A Bayesian-Inspired Item Response Theory–Based Framework to Produce Very Short Versions of MacArthur–Bates Communicative Development Inventories

Jun Ho Chai[b,*], Chang Huan Lo[b,*] and Julien Mayor[a]

[a]University of Oslo, [b]University of Nottingham Malaysia

Author Note

Correspondence to Julien Mayor: julien.mayor@psykologi.uio.no

*Jun Ho Chai and Chang Huan Lo share first authorship.

Purpose:

This study introduces a framework to produce very short versions of the MacArthur–Bates Communicative Development Inventories (CDIs) by combining the Bayesian-inspired approach introduced by Mayor and Mani (2019) with an item response theory–based computerized adaptive testing that adapts to the ability of each child, in line with Makransky et al. (2016).

Method:

We evaluated the performance of our approach— dynamically selecting maximally informative words from the CDI and combining parental response with prior vocabulary data—by conducting real-data simulations using four CDI versions having varying sample sizes on Wordbank—the online repository of digitalized CDIs: American English (a very large data set), Danish (a large data set), Beijing Mandarin (a medium-sized data set), and Italian (a small data set).

Results:

Real-data simulations revealed that correlations exceeding .95 with full CDI administrations were reached with as few as 15 test items, with high levels of reliability, even when languages (e.g., Italian) possessed few digitalized administrations on Wordbank.

Conclusions:

The current approach establishes a generic framework that produces very short (less than 20 items) adaptive early vocabulary assessments—hence considerably reducing their administration time. This approach appears to be robust even when CDIs have smaller samples in online repositories, for example, with around 50 samples per month-age.

A Bayesian-Inspired Item Response Theory–Based Framework to Produce Very Short Versions of MacArthur–Bates Communicative Development Inventories

The MacArthur–Bates Communicative Development Inventories (CDIs) are one of the most widely used sets of parent report instruments for assessing young children's early language development (Fenson et al., 2007). Originally developed in American English (Fenson et al., 1993), CDIs have since been adapted into nearly 100 languages (e.g., Danish, Mandarin, and Italian), including language variations (e.g., British English, Australian English, and Singaporean English). Adaptations have also been developed in a number of sign languages, including American Sign Language and British Sign Language (see CDI Advisory Board, 2015, for a list of available adaptations).

CDIs typically consist of three forms: The CDI: Words and Gestures (CDI:WG)—targeting children approximately8–18 months of age—assesses both comprehension and production of early vocabulary, as well as production of communicative gestures; the CDI: Words and Sentences (CDI:WS)—targeting children approximately 16–30 months of age—assesses productive vocabulary and morphosyntactic skills; and the CDI-III—a short form targeting children approximately 30–37 months of age—assesses productive vocabulary, syntactic maturity, and language use (Dale et al., 1998; Fenson et al., 2007).

These assessment tools rely on parents' knowledge about their children's language and allow a representative picture of children's early language development (Fenson et al., 2000). Beyond being cost-effective, CDIs are also reliable and valid, not only with children who are developing typically (Fenson et al., 1993, 2007; Law & Roy, 2008; Pan et al., 2004; Rescorla et al., 2005) but also with children with developmental disabilities (Galeote et al., 2016; Luyster et al., 2007; Mayne et al., 1999, 1998; Thal et al., 2007).

Through the application of CDIs in various languages, similarities have been observed in lexical development trajectories among children speaking different languages (Bleses et al., 2008; Braginsky et al., 2019; Frank et al., in press). The evidence suggests that most children produce their first words between 12 and 20 months of age (Bleses et al., 2008; Devescovi et al., 2005; Fernald et al., 1998), that their vocabulary acquisition rate increases rapidly after 18 months of age (Bates & Goodman, 1997; Fernald et al., 2006, 1998), and that there is a strong relationship between lexical and grammatical development (e.g., Bates & Goodman, 1997; Caselli et al., 1999; Conboy & Thal, 2006; Devescovi et al., 2005; Marjanovič-Umek et al., 2013; Stolt et al., 2009). CDIs have also been used as additional criteria for identifying late language emergence; for example, starting from 24 months of age, a child is typically considered to be a late talker or a late language learner if they have an expressive vocabulary at or below the 10th percentile on the CDI (Dale et al., 2003; Desmarais et al., 2008; Ellis Weismer, 2007; Rescorla & Dale, 2013).

Despite the many advantages and widespread applications of CDIs, completion of the forms requires a significant amount of time and that the parent should be literate. The American English CDI:WS, for example, includes a vocabulary checklist of 680 words, organized into 22 semantic categories (e.g., vehicles, toys, people, action words, descriptive words, and question words). Under circumstances when a rapid assessment is desirable (whether in a battery of tests or in multilingual environments) or when parents have low literacy skills, the applicability of CDIs becomes limited. To address these drawbacks, Fenson et al. (2000) developed the first short-form versions (CDI:SF) of the CDI:WG and CDI:WS with items drawn from the original full forms. The former consists of an 89-item checklist, whereas the latter consists of two 100-item checklists to allow for repeated administrations. As with the full CDIs, these short forms have demonstrated high validity and reliability, and are at the same time highly correlated with the full forms, thus making them a useful alternative when time or parental literacy is limited (Fenson et al., 2000). Nevertheless, due to their brevity, these short forms may not be as precise as the full forms and may fail to capture individual differences. The short-form CDI:WS suffers from a ceiling effect after 27– 28 months and even more so when children have a large vocabulary. Furthermore, it takes much time and effort to develop such forms for each language in order to maintain a good balance of items from different semantic categories, as well as items with different levels of difficulty.

With the objective to develop a short-form version of CDI:WS that maintains the accuracy and precision of the full form and is tailored to each child, Makransky et al. (2016) applied computerized adaptive testing (CAT; van der Linden & Glas, 2010), whose principle is based on item response theory (IRT; Embretson & Reise, 2000). In their approach (hereafter referred to as "CDI:CAT"), items in the American English CDI:WS norming sample[1] are fitted to an IRT model. During CAT, 10 items with maximal item information are initially sampled at random from the full CDI. The algorithm then selects subsequent items that reflect the ability parameter of the child that is estimated at each point (i.e.,

[1] The American English CDI-WS norming sample consists of 1;461 children between 16 and 30 months of age.

item) in the test. Based on the results obtained from CDI:CAT simulations with 5, 10, 25, 50, 100, 200, 400, and 680 (the full form) items, it was found that, at 50 items and above, CDI:CAT performed well, with correlations above .95 with the full CDI, average SE below .20, and reliability coefficients above .96 (above what Makransky et al. described as a minimal threshold for test acceptability). Although this may be a viable solution to reducing the lengths of the full CDI, the performance of CDI:CAT with novel empirical data, as pointed out by Makransky et al., may be lower due to a systematic or random error. It is also possible that the respondents would respond differently as items in CDI:CAT are presented in a semantically unstructured order as opposed to the semantic grouping adopted in CDIs.

Recently, Mayor and Mani (2019) presented a languagegeneral approach that takes advantage of the richness of Wordbank (Frank et al., 2017), an open repository for crosslinguistic CDI data from over 75,000 children across 29 languages. Their approach combines a subset of items drawn randomly from the full forms with (prior) CDI data sampled from language-, gender-, and age-matching children on Wordbank. Real-data simulations conducted using CDI:WS data of American English (Fenson et al., 2007), German (Szagun et al., 2014), and Norwegian (Simonsen et al., 2014) revealed that, at 50 items, correlations reached .97, with average SEs of .05, and reliability coefficients of .99, suggesting that their approach, which takes into account children's age and gender, outperforms CDI:CAT. Empirical validation with 25- and 50-item checklists administered to the parents of German-speaking children further demonstrated good performance, with correlations of .96, average SEs of .14, and a reliability of .98, above Makransky et al.'s (2016) recommended thresholds, even when parents showed inconsistencies (about 10%–15% of responses) in responding in the full and short forms. However, to capture the full extent of the large variations in vocabulary acquisition (e.g., withinand between-age variations, gender differences; Fenson et al., 2007), Mayor and Mani's approach requires a considerably large sample size on Wordbank—for example, the German CDI:WS data set, the smallest data set used in the study of Mayor and Mani, has over 70 children in each age group. Thus, it is unclear how their approach would perform with smaller sample sizes, for example, for languages having fewer computerized forms on Wordbank.

Another obvious limitation to this approach is that items are randomly selected during the test. Consequently, the sampled items may be minimally informative, that is, items that are either too easy (e.g., "cat" is produced by over 95% of 30-month-olds in American English) or too difficult (e.g., "snowman" is produced by just 1% of 16-month-olds), and hence may inform little about a child's language ability.

To address these issues, our approach aims to produce language-general, short-form versions of CDIs in which items are selected to be maximally informative. Building on Mayor and Mani's (2019) work, we applied aprincipled selection of items in place of random selection. More specifically, we applied IRT-based CAT in real-data simulations, as in Makransky et al. (2016). In IRT, items may differ in discrimination value, which determines how well each item discriminates the level of knowledge across individuals (Fraley et al., 2000). For instance, items with high difficulty level may discriminate two children having a high degree of knowledge but may not for weaker children. Thus, by applying IRT, the risk of sampling minimally informative items can be circumvented. To validate our approach, we selected four CDI:WS versions for which their sample sizes on Wordbank vary: American English (a very large data set; Fenson et al., 2000), Danish (a large data set; Bleses et al., 2008), Beijing Mandarin (a mediumsized data set; Tardif et al., 2008), and Italian (a small data set; Rinaldi et al., 2019). This, in turn, helped to evaluate the possibility of applying a CAT approach to languages for which only small samples are represented on Wordbank and to establish short language assessments that are reliable. An evaluation of performance was conducted across different age groups and genders.

The next section details the two main components of our approach, that is, the IRT-based selection of test items coupled with a full CDI score estimation based on Mayor and Mani's (2019) model. We then present the result and discuss the implications of our findings for researchers and practitioners intending to use short forms for quick and cost-effective assessments of young children's vocabulary.

Method

IRT-Based Item Selection

Prior to an assessment, items on the CDI are fitted to a two-parameter IRT model. Item parameters are estimated using the mirt function from Chalmers's (2012) mirt package in R. Each item is assigned two parameters: a discrimination parameter and a difficulty parameter. These item parameters are computed with marginal maximum likelihood estimation using the expectation–maximization algorithm (Bock & Aitkin, 1981). During CAT assessments, items with maximum information are prioritized and tested. The estimation of the ability parameter of each child is conducted using the weighted likelihood estimation method (Warm, 1989) and is subsequently used to select items that can maximally inform about the knowledge level of the child. The test items are selected using Chalmers's (2016) mirtCAT package in R. This principled selection of test items on the basis of both the child's estimated ability and the properties of the test items is expected to improve the relevance of the items sampled in the algorithm. Based on children's responses on the items administered in CATs, the next step computes estimates of their full CDI scores.

CDI Score Estimation

The method of fitting each of the item-based histograms of full CDI scores to a normal distribution and the estimation of CDI scores closely resembled those described in Mayor and Mani (2019). First, language-, gender-, and age-matched children are retrieved from Wordbank (Frank et al., 2017) using Braginsky's (2018) wordbankr package in R. Then, for each test item i that is reported by the parent as either known or not known by the child, a histogram of full CDI scores of all other children with the same response is extracted from Wordbank. The resulting histogram is fitted with a normal distribution using maximum likelihood estimation. A polynomial is fitted with the parameters (mean and standard deviation) extracted from the fitted histogram to smoothen out random fluctuations. Unlike Mayor and Mani whose polynomial fitting was conducted with a degree of three, the present model took a more flexible approach, in which the degree of the polynomial adjusts to the breadth of the distribution of the vocabulary counts.[2] Once normalized, this histogram can be seen as the probability distribution of full CDI scores given the child's response for item i. The histograms of all test items are subsequently log-summed, and we retrieve the mode of the resulting histogram. Finally, a linear transformation of the mode, ensuring that the full range of CDI values associated with language-, age-, and gendermatching children can be reached, produces the estimate of the child's full CDI score.

[2] That is, to this end, the median absolute deviation (MAD) of productive vocabulary is computed for each age, in months. When MAD < 100, a linear function is fitted to improve generalization. When MAD > 100, a cubic polynomial is fitted to obtain a finer model.

Real-Data Simulations

The data used in this study were from the American English (Fenson et al., 2000), Danish (Bleses et al., 2008), Beijing Mandarin (Tardif et al., 2008), and Italian (Rinaldi et al., 2019) CDI:WS, retrieved from Wordbank (Frank et al., 2017). The American English data set was categorized as very large sized, with more than 200 samples for each age (in months); the Danish data set was categorized as large sized, with between 100 and 200 samples for each age; the Beijing Mandarin data set was categorized as medium sized, with between 50 and 100 samples for each age; whereas the Italian data set was categorized as small sized, with fewer than 50 samples for each age. These versions of the CDI:WS were selected for having relatively homogeneous sample sizes across all ages.

Using real-data simulations, we compared the performance of the present model (the IRT version) with the original model by Mayor and Mani (2019) in estimating full CDI scores. We adopt the following standard for test acceptability, as introduced by Makransky et al. (2016): correlations above .95 with the full CDI, average SE below .20, and reliability coefficients (1–SE^2) above .96. In addition, the sum of words that are reported as known on a random selection of items from the CDI, scaled up to the full CDI size, was used as a baseline measure and compared with the IRT version. Results for the original model and the baseline measure were averaged over 10 simulations, whereas those for the IRT version were based on single simulations, since the item selection process establishes individual parameters for each child, consequently constraining the selection of words for that child. In line with previous work using real-data simulations (i.e., Makransky et al., 2016; Mayor & Mani, 2019), correlations between the estimates (based on 5, 10, 25, 50, 100, 200, 400, and all items on each CDI) and the full CDI scores, average standard errors, and reliability (1–SE^2) were reported across different age groups and genders.

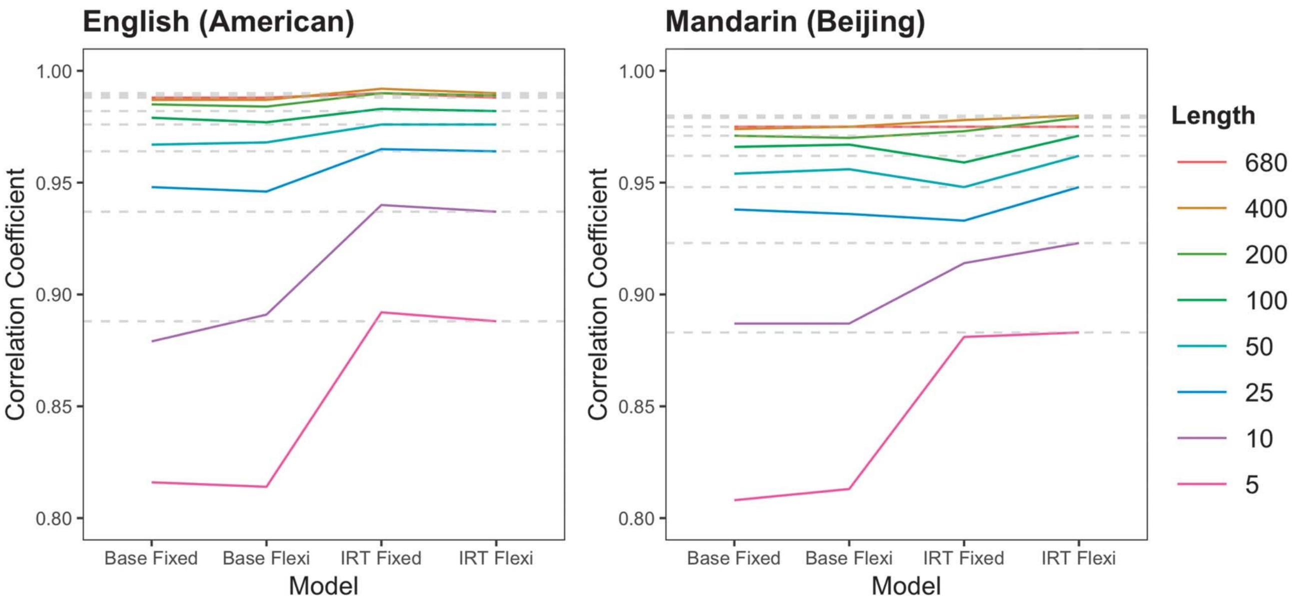
Further evaluation of the performance of the IRT version was conducted using established CDI:SFs. CDI:SF scores were obtained by summing raw vocabulary count based on responses on CDI:SF items and scaling these scores up to the instrument size to fit the range of full CDI scores. Likewise, correlation coefficients of these scores with the full CDI scores, reliability, and average standard errors were computed and compared to those obtained from the IRT version.

Results

Model Selection

Two changes were made to the original model (Mayor & Mani, 2019): the application of IRT in item selection and the flexible approach to polynomial fitting. Preliminary comparisons between correlation coefficients of the original model, the original model equipped with flexible polynomial fitting, the original model with IRT (but without flexible polynomials), and the IRT version (with both flexible polynomial fitting and IRT) are shown in Figure 1. Correlations were compared using two CDIs having different sample sizes: the very large-sized American English CDI data set and the medium-sized Beijing Mandarin CDI data set, in order to select the final model. When applied to the medium-sized data set, the combination of flexible polynomial fitting and IRT led to the largest improvements. For the very large-sized data set, the mere application of IRT improved the model the most, although performance was comparable to the level of performance attained with the maximal model (IRT and flexible polynomials). With the merit that improvements were observed with the smaller data set when both flexible polynomial fitting and IRT were applied, we selected the IRT version as the final model.

Figure 1. A comparison between correlation coefficients of the original model (Base Fixed), the original model with flexible polynomial fitting (Base Flexi), the original model with item response theory (IRT; IRT Fixed), and the IRT version with flexible polynomial fitting (IRT Flexi).

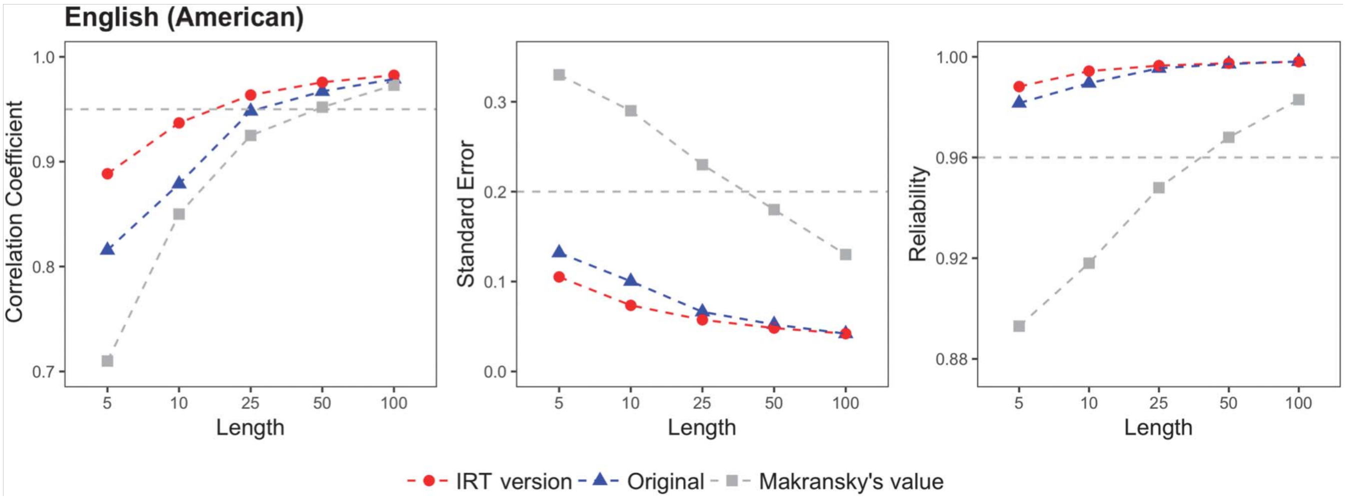


American English CDI:WS

The original model and the IRT version were used in real-data simulations on the very large-sized American English CDI:WS data set for children between 16 and 30 months of age, for each gender, using a different number of test items (5, 10, 25, 50, 100, 200, 400, and 680, the full CDI size). Figure 2 shows the correlations between the estimated scores and the full CDI scores using 5, 10, 25, 50, and 100 items, along with the average standard errors, and the reliability of both the IRT version and the original model, as well as Makransky et al.'s (2016) values for comparison.[3]

[3] For the full list of values at all test lengths, across gender, see Appendix Table A1.

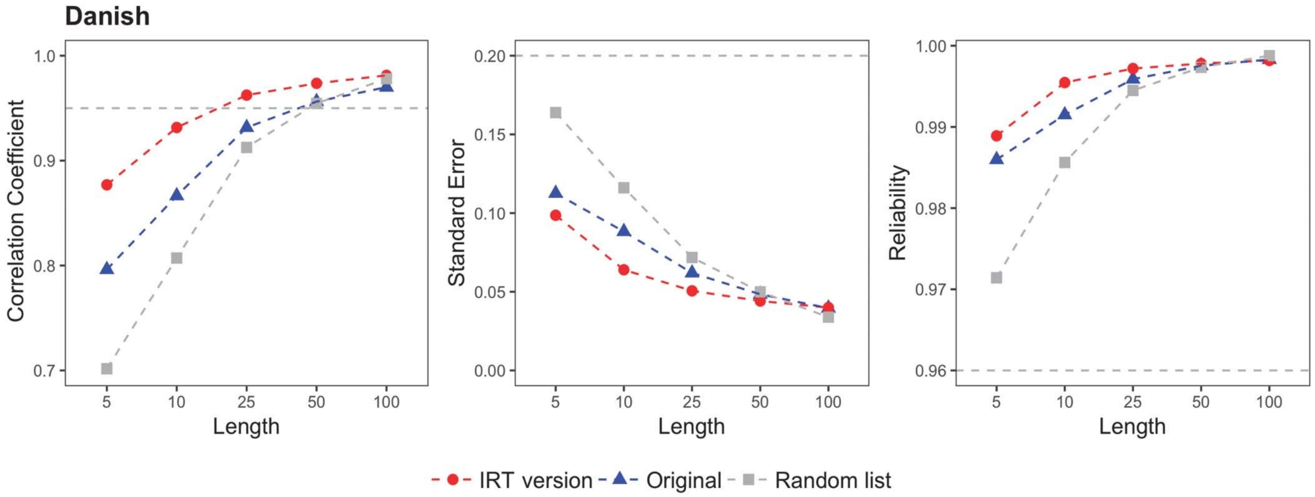
Figure 2. Comparison of the item response theory (IRT) version and the original model with different test lengths (5, 10, 25, 50, 100) on the American English MacArthur–Bates Communicative Development Inventories: Words and Sentences, with Makransky et al.'s (2016) values for reference. The gray dashed lines at .95 on correlation, .20 on standard error, and .96 on reliability represent the cutoff points suggested by Makransky et al. The x-axes are not linear.



In terms of correlations, the IRT version performed better than the original model, with correlations above .9, provided that the test has more than just five items. In terms of both average standard errors and reliability, the IRT version had values similar to the original model at 100 items but outperformed the latter at 50 items and below. At 25 items, correlations greater than the .95 cutoff point, as suggested by Makransky et al. (2016), were achieved. Additional real-data simulations revealed that a correlation of .95 was already achieved at 14 items, with an average SE of .07 and a reliability of .995.

To further examine the effectiveness of the IRT version across ages, an analysis was conducted per age group (see Appendix Table A2). Improvements in correlations were observed for all age groups when compared to the original model. It is noteworthy that, at 25 items, correlations across all age groups were already higher than .95, whereas in the original model, the youngest age group (16–18 months) required at least 50 items to achieve correlations of .95 and above. In line with Makransky et al. (2016) and Mayor and Mani (2019), a marked reduction in performance was observed for both the youngest and oldest age groups, when the test featured less than 10 items.

Figure 3. Comparison of the item response theory (IRT) version and the original model with different test lengths (5, 10, 25, 50, 100) on the Danish MacArthur–Bates Communicative Development Inventories: Words and Sentences, with random list as the baseline measure. The gray dashed lines at .95 on correlation, .20 on standard error, and .96 on reliability represent the cutoff points suggested by Makransky et al. (2016). The x-axes are not linear.



Danish CDI:WS

Real-data simulations were conducted using the large-sized Danish CDI:WS (Bleses et al., 2008) data set. Figure 3 depicts the performance of the IRT version and the original model for children between 16 and 30 months of age when having five to 100 items, with random lists as the baseline measure for comparisons.[4] Similar to the American English data, consistent improvements in correlations, average standard errors, and reliability were observed for the IRT version. With the Danish CDI:WS data set, the IRT version was again able to achieve correlations of above .95 at 25 items, as it did with the American English CDI:WS data set, whereas the original model required at least 50 items to achieve a similar performance. Furthermore, the IRT version had better correlations, average standard errors, and reliability than the baseline measure at 50 items and below. Additional real-data simulations revealed that a correlation of .95 was already achieved at 17 items, with an average SE of .06 and a reliability of .997.

Beijing Mandarin CDI:WS

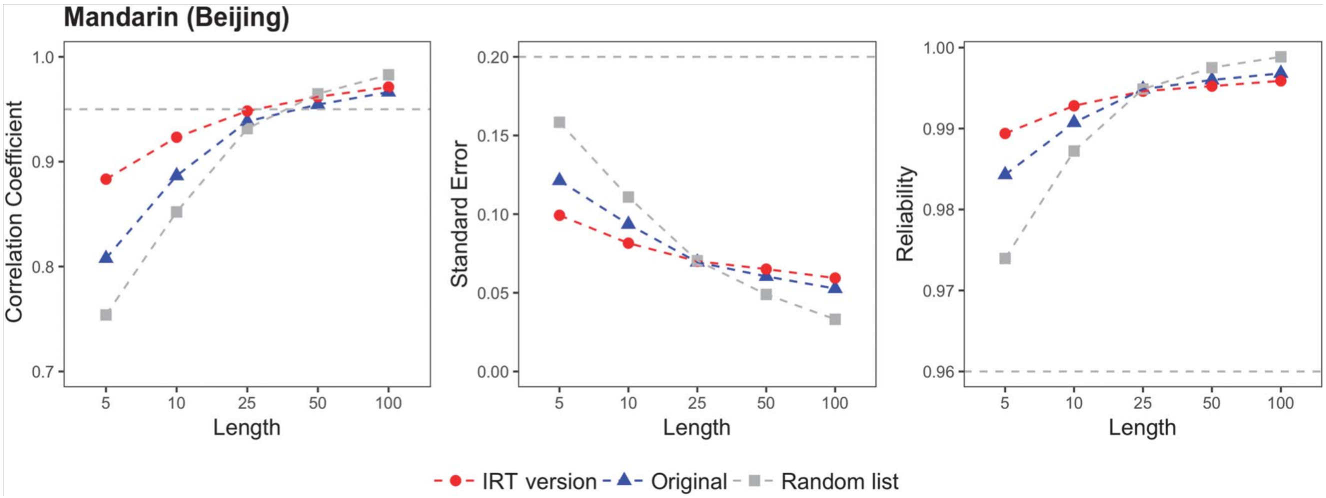
Real-data simulations were run on the medium-sized Beijing Mandarin CDI:WS (Tardif et al., 2009) data set for children between 16 and 30 months of age, and the results[5] are illustrated in Figure 4 for tests with 100 items and below. In terms of correlations, the IRT version performed better than the original model, with similar or better average standard errors and reliability. With a relatively smaller sample size, the IRT version achieved correlations of .95 at 25 items for male and at 50 items for female. When compared to the baseline measure, the IRT version had better correlations at 25 items and below, whereas the average standard errors and reliability were similar at 25 items and only better at 10 items and below. Additional real-data simulations revealed that a correlation of .95 was

[4] For the full list of values at all test lengths, across gender, see Appendix Table A3.

[5] For the full list of values at all test lengths, across gender, see Appendix Table A4.

achieved at 23 items for male, with an average SE of .09 and a reliability of .992, and at 36 items for female, with an average SE of .08 and a reliability of .993.

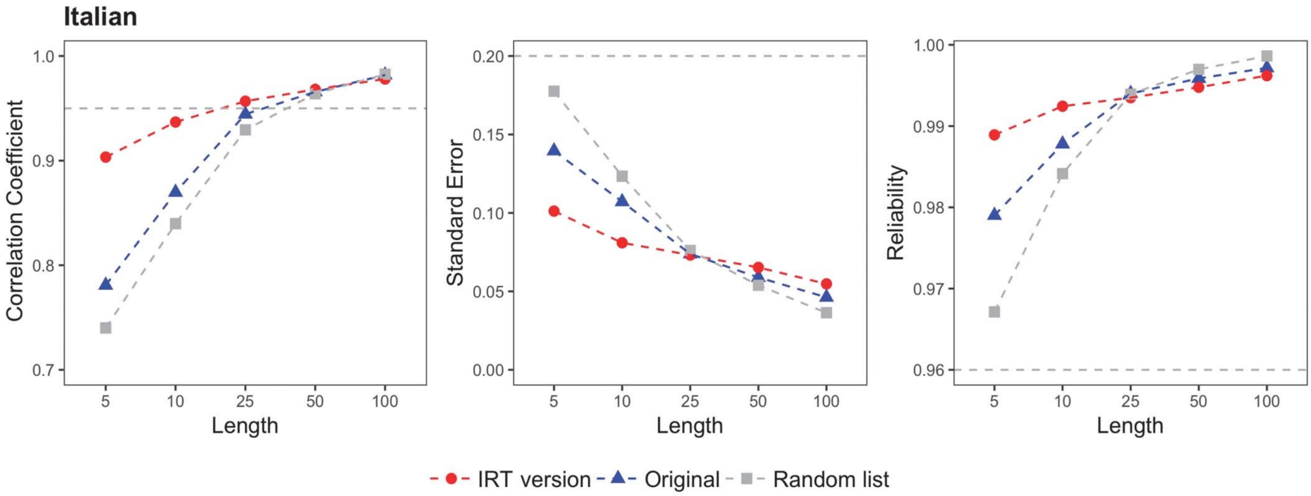
Figure 4. Comparison of the item response theory (IRT) version and the original model with different test lengths (5, 10, 25, 50, 100) on the Beijing Mandarin MacArthur–Bates Communicative Development Inventories: Words and Sentences, with random list as the baseline measure. The gray dashed lines at .95 on correlation, .20 on standard error, and .96 on reliability represent the cutoff points suggested by Makransky et al. (2016). The x-axes are not linear.



Italian CDI:WS

Real-data simulations were conducted on the smallsized Italian CDI:WS (Caselli et al., 1995) data set for children between 18 and 30 months of age. For this particular data set, the original model (with fixed degree of polynomial fit of 3) was unable to reliably estimate scores. Thus, we could only compare the performance of the IRT version with the original model with flexible polynomial fitting. As shown in Figure 5, the IRT version outperformed the original version in terms of correlations at 25 items and below, with better or similar average standard errors and reliability.[6] Correlations of above .95 were already achieved with the IRT version starting at 25 items, whereas the original model achieved the same for females, but not for males (starting at 50 items). Correlations of the IRT version were better than the baseline measure at 50 items and below, whereas average standard errors were better at 25 items and below and reliability at 10 items and below. Additional real-data simulations revealed that a correlation of .95 was already achieved at 15 items, with an average SE of .08 and a reliability of .993.

Figure 5. Comparison of the item response theory (IRT) version and the original model with flexible polynomial fitting with different test lengths (5, 10, 25, 50, 100) on the Italian MacArthur–Bates Communicative Development Inventories: Words and Sentences, with random list as the baseline measure. The gray dashed lines at .95 on correlation, .20 on standard error, and .96 on reliability represent the cutoff points suggested by Makransky et al. (2016). The x-axes are not linear.



[6] For the full list of values at all test lengths, across gender, see Appendix Table A5.

Comparisons With Established Short-Form Versions of CDIs

Using 100-item tests (110-item tests for Beijing Mandarin), the performance of the IRT version was compared with the American English (Fenson et al., 2000), Danish (Bleses et al., 2010), Beijing Mandarin (Tardif et al., 2008), and Italian (Rinaldi et al., 2019) CDI:SFs, as well as random lists as the baseline measure. For a more detailed evaluation, comparisons were made across different age groups. Overall, all three approaches met the criterion for test acceptability suggested in Makransky et al. (2016) across all age groups and CDIs.

For the very large-sized American English CDI data set, comparisons were made using Form A of the American English CDI:SF (Fenson et al., 2000). Table 1 reports correlations, average standard errors, and reliability scores for the IRT version, CDI:SF, and the baseline measure, across five age groups. The IRT version performed better than CDI:SF in terms of correlations between 16 and 24 months, whereas CDI:SF performed better between 25 and 30 months, though they both had similar average standard errors and reliability. The baseline measure outperformed the IRT version between 22 and 30 months as well as CDI:SF across all age groups, with better correlations, average standard errors, and reliability.

Table 1. Comparisons between the IRT version of the model, Fenson et al. (2000)'s American English CDI-SF and the baseline measure using 100 test items, by age group.

Age	IRT version			CDI-SF			Baseline		
	r with full CDI	Avg.SE	Rel.	r with full CDI	Avg.SE	Rel.	r with full CDI	Avg.SE	Rel.
16 - 18	.982	.03	.999	.954	.04	.998	.975	.03	.999
19 - 21	.990	.03	.999	.973	.05	.997	.985	.04	.999
22 - 24	.985	.04	.998	.984	.05	.997	.988	.04	.998
25 - 27	.978	.05	.997	.986	.06	.997	.988	.04	.999
28 - 30	.978	.05	.997	.985	.04	.998	.987	.04	.999

Comparisons made using the large-sized Danish CDI data set, across five age groups, among the IRT version, the Danish CDI:SF (Bleses et al., 2010), and the baseline measure are reported in Table 2. The IRT version had better correlations than CDI:SF and the baseline measure between 16 and 24 months. After 24 months, CDI:SF performed best in terms of correlations. Overall, the average standard errors and reliability of the IRT version were similar, if not slightly poorer, when compared to both CDI:SF and the baseline measure.

Table 2. Comparisons between the IRT version of the model, Bleses et al. (2010)'s Danish CDI-SF and the baseline measure using 100 test items, by age group.

Age	IRT version			CDI-SF			Baseline		
	r with full CDI	Avg.SE	Rel.	r with full CDI	Avg.SE	Rel.	r with full CDI	Avg.SE	Rel.

16 - 18	.986	.02	.999	.968	.02	1.000	.972	.02	1.000
19 - 21	.978	.05	.997	.969	.03	.999	.973	.03	.999
22 - 24	.990	.04	.999	.983	.04	.998	.982	.04	.998
25 - 27	.981	.05	.997	.984	.05	.997	.983	.04	.998
28 - 30	.971	.06	.997	.985	.05	.997	.98	.04	.998

For the medium-sized Beijing Mandarin CDI data set, 110 items were administered in accordance with the number of items in the Beijing Mandarin CDI:SF (Tardif et al., 2008). The results reported across five age groups in Table 3 indicated poorer performance of the IRT version in terms of the correlations, except for the youngest age group, that is, 16–18 months. Average standard errors and reliability scores were also poorer than both CDI:SF and the baseline measure.

Table 3. Comparisons between the IRT version of the model, Tardif and Fletcher (2008)'s Beijing Mandarin CDI-SF and the baseline measure using 110 test items, by age group.

Age	IRT version			CDI-SF			Baseline		
	r with full CDI	Avg.SE	Rel.	r with full CDI	Avg.SE	Rel.	r with full CDI	Avg.SE	Rel.
16 - 18	.986	.06	.995	.980	.04	.999	.979	.04	.999
19 - 21	.984	.05	.998	.990	.05	.998	.990	.05	.998
22 - 24	.963	.07	.995	.981	.04	.998	.986	.04	.998
25 - 27	.961	.06	.997	.979	.04	.998	.983	.04	.999

28 - 30	.970	.06	.996	.981	.03	.999	.976	.03	.999

The final comparisons were made using the small-sized Italian CDI data set, among the IRT version, the Italian CDI:SF (Rinaldi et al., 2019), and the baseline measure, across four age groups. As reported in Table 4, the IRT version had better correlations than CDI:SF and the baseline measure between 18 and 24 months. Between 25 and 30 months, CDI:SF had the highest correlations. Average standard errors and reliability were similar across all three approaches.

Table 4. Comparisons between the IRT version of the model, Rinaldi et al. (2019)'s Italian CDI-SF and the baseline measure using 100 test items, by age group.

Age	IRT version			CDI-SF			Baseline		
	r with full CDI	Avg.SE	Rel.	r with full CDI	Avg.SE	Rel.	r with full CDI	Avg.SE	Rel.
18 - 21	.981	.04	.998	.972	.03	.999	.975	.03	.999
22 - 24	.990	.03	.999	.983	.04	.998	.982	.04	.998
25 - 27	.981	.04	.998	.984	.05	.997	.983	.05	.998
28 - 30	.971	.05	.998	.985	.05	.997	.980	.05	.998

Discussion

CDIs are a cost-effective, reliable, and valid set of parent report instruments for assessing children's early language development from 8 up to 37 months of age. However, due to their size, the administration of CDIs is time-consuming and require that parents be literate, thus restricting the applicability of CDIs when rapid assessments are desirable or when parents have low literacy skills. To deal with these drawbacks, researchers have sought to reduce the lengths of CDIs using different approaches, including developing short forms in different languages (e.g., Bleses et al., 2010; Fenson et al., 2000); administering CDIs as CAT (Makransky et al., 2016); and, more recently, estimating full CDI scores based on CDI data from language-, gender-, and age-matching children on Wordbank (Mayor & Mani, 2019).

The present approach, that is, the IRT version, combined Mayor and Mani's (2019) Bayesian-inspired approach with an IRT-based CAT that dynamically selects test items that are maximally informative based on both the child's ability and the properties of the test items (as in Makransky et al., 2016). To evaluate the IRT version, real-data simulations were conducted using four CDI:WS versions with varying sample sizes on Wordbank: American English (a very large data set; Fenson et al., 2000), Danish (a large data set; Bleses et al., 2008), Beijing Mandarin (a medium-sized data set; Tardif et al., 2008), and Italian (a small data set; Rinaldi et al., 2019). Results obtained were subsequently compared with three other approaches: Mayor and Mani's model (in a novel implementation, in R), a baseline measure (i.e., the sum of responses obtained directly from a set of items sampled randomly from the full forms), and CDI:SF. For the American English CDI:WS, Makransky et al.'s (2016) results were also included in the comparisons.

Overall, the IRT version achieved correlations with the full CDI above .95, average SE below .20, and reliability above .96 (a criterion for test acceptability suggested in Makransky et al., 2016) with fewer than 17 items for American English, Danish, and Italian. For the Mandarin data set, this criterion was only met from 23 items for males and 36 items for females.

To explain the uneven performance between both genders in the Mandarin data set, we further inspected the data set and found much lower variation (quantified by MAD) in the female data than in the male data. More specifically, starting from 23 months, the female data were more left-skewed than the male data; that is, a majority of females had high CDI scores, whereas males' scores continued to vary until about 27 months, when a majority, like females, began to have high CDI scores. The implication is twofold: First, it may be that, for girls, a larger sample size is needed for a better representation of the population; second, many items in the Mandarin CDI appear to be too easy, in particular, for girls older than 23 months, hence reaching a ceiling earlier than boys. Despite this exception, our results suggest that a 25-item test can reliably estimate a child's CDI scores in most cases, regardless of gender and language. Analyses conducted per age group on the American English data set extend this finding, further suggesting that a 25-item checklist is suitable for use with children across all age groups (16–30 months).

Comparisons with Mayor and Mani (2019) revealed that the IRT version had similar or better performance in terms of correlations, average standard errors, and reliability, across all four CDIs and both genders, regardless of the number of test items. In other words, the scores established by the IRT version matches more closely the full CDI scores. When compared against the baseline measure, the IRT version performed better in terms of correlations, average standard errors, and reliability for all short tests, that is, having 50 items and below. It is noteworthy, however, that the baseline measure—

summing a random selection of words—performed well across all four CDIs, already achieving correlations of above .95 with good average standard errors and reliability, starting at just 50 items. At 100 items, the baseline measure's performance was also comparable to CDI:SFs. Such impressive performance should be attributed to the high internal consistency of CDIs (e.g., Bleses et al., 2008; Fenson et al., 1994; Tardif et al., 2009).

The final comparisons were made with CDI:SFs and the baseline measure across age groups, with 100 test items (or 110 for the Beijing Mandarin CDI:SF). While the IRT version typically outperformed CDI:SFs in the younger age groups, that is, between 16 and 24 months (with the exception of the Mandarin CDI:SF), both the baseline measure and CDI:SFs performed better in the older age groups, that is, between 25 and 30 months. Nevertheless, the performance of the IRT version was still comparable to the baseline measure and CDI:SFs, with all three approaches meeting Makransky et al.'s (2016) suggested criterion for test acceptability across all age groups. An important point to note here is that the development of CDI:SF for even just one language is labor intensive, whereas the IRT version has the advantage of being cost-effective in that it is generalizable; that is, it can be directly applied to CDIs of any languages, as long as sufficient CDI data are available online. Crucially, our objective is to develop a brief test that allows for rapid assessments—a 100-item checklist may still be considered too long in cases requiring multiple forms to be completed (e.g., in a multilingual environment, a clinical setting) or intimidating when parents have low literacy. The IRT version, on the other hand, is able to provide reliable estimates with just 14–25 items, gaining a factor of 4–7 compared to CDI:SFs.

The results reported here are based on real-data simulations. A full assessment of the psychometric properties of the IRT version should be conducted with new participants, in particular, to establish its test–retest reliability and its validity using an array of validity tests. With new participants, we also expect reduced level of performance as a result of parents responding differently to the same item in the full and short forms. This was demonstrated in Mayor and Mani (2019), with parents responding more positively in both the 25- and 50-item checklists than in the full CDI. In addition, as opposed to the more structured full forms that organize items into different semantic categories, our approach presents items in a semantically unstructured order, which may in turn affect parents' response behavior. Therefore, the essential next steps include investigating the differences in parents' response behavior and validating the model on new participants.

Finally, the reliability of our generic approach relies upon the availability of CDI data from children with matching key demographics (e.g., language, age, and gender). Based on our findings, even with a small data set having less than 50 samples available online for each age group (in months), our approach is able to reliably estimate children's full CDI scores with just 25 items, effectively reducing administration time to a mere couple of minutes. Thus, it is vital that data collected on children's vocabulary be shared publicly to enable access to and reuse of these data that will allow for the establishment of computerized adaptive tests that are tailored to each child.

Acknowledgments

This work was supported by the following Malaysian grants awarded to J. M.: MOSTI-06-02-12-SF0351 and FRGS-NFHY0004.

References

Bates, E., & Goodman, J. C. (1997). On the inseparability of grammar and the lexicon: Evidence from acquisition, aphasia and real-time processing. Language and Cognitive Processes, 12(5–6), 507–584. https://doi.org/10.1080/016909697386628

Bleses, D., Vach, W., Jørgensen, R. N., & Worm, T. (2010). The internal validity and acceptability of the Danish SI-3: A language-screening instrument for 3-year-olds. Journal of Speech, Language, and Hearing Research, 53(2), 490–507. https://doi. org/10.1044/1092-4388(2009/08-0132)

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T. O., & Basbøll, H. (2008). The Danish communicative developmental inventories: Validity and main developmental trends. Journal of Child Language, 35(3), 651–669. https://doi.org/ 10.1017/S0305000907008574

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46(4), 443–459. https://doi.org/10.1007/ BF02293801

Braginsky, M. (2018). wordbankr: Accessing the Wordbank database (R package Version 0.3.0) [Computer software]. https:// CRAN.R-project.org/package=wordbankr

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and variability in children's word learning across languages. Open Mind: Discoveries in Cognitive Science, 3, 52–67. https://doi.org/10.1162/opmi_a_00026

Caselli, M. C., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L., & Weir, J. (1995). A cross-linguistic study of early lexical development. Cognitive Development, 10(2), 159–199. https://doi.org/10.1016/0885-2014(95)90008-X

Caselli, M. C., Casadio, P., & Bates, E. (1999). A comparison of the transition from first words to grammar in English and Italian. Journal of Child Language, 26(1), 69–111. https://doi.org/10.1017/S0305000998003687

CDI Advisory Board. (2015). Adaptations in other languages. http://mb-cdi.stanford.edu/adaptations.html

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. Journal of Statistical Software, 48(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory Journal of Speech, Language, and Hearing Research • 1–13 applications. Journal of Statistical Software, 71(5), 1–39. https:// doi.org/10.18637/jss.v071.i05

Conboy, B. T., & Thal, D. J. (2006). Ties between the lexicon and grammar: Cross-sectional and longitudinal studies of bilingual toddlers. Child Development, 77(3), 712–735. https://doi.org/10.1111/j.1467-8624.2006.00899.x

Dale, P. S., Price, T. S., Bishop, D. V., & Plomin, R. (2003). Outcomes of early language delay: I. Predicting persistent and transient delay at 3 and 4 years. Journal of Speech, Language, and Hearing Research, 46(3), 544–560. https://doi.org/10.1044/1092-4388(2003/044)

Dale, P. S., Reznick, J. S., & Thal, D. J. (1998). A parent report measure of language development for three-year-olds. Infant Behavior & Development, 21, 370. https://doi.org/10.1016/S0163-6383(98)91583-1

Desmarais, C., Sylvestre, A., Meyer, F., Bairati, I., & Rouleau, N. (2008). Systematic review of the literature on characteristics of late-talking toddlers. International Journal of Language &

Communication Disorders, 43(4), 361–389. https://doi.org/10.1080/13682820701546854

Devescovi, A., Caselli, M. C., Marchione, D., Pasqualetti, P., Reilly, J., & Bates, E. (2005). A crosslinguistic study of the relationship between grammar and lexical development. Journal of Child Language, 32(4), 759–786. https://doi.org/10.1017/S0305000905007105

Ellis Weismer, S. (2007). Typical talkers, late talkers, and children with specific language impairment: A language endowment spectrum? In R. Paul (Ed.), Language disorders from a developmental perspective: Essays in honour of Robin S. Chapman (pp. 83–101). Erlbaum. https://doi.org/10.4324/9781315092041-3.

Embretson, S. E., & Reise, S. P. (2000). Multivariate applications books series. Item response theory for psychologists. Erlbaum.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., & Pethick, S. J. (1994). Variability in early communicative development. Monographs of the Society for Research in Child Development, 59(5), 1–185. https://doi.org/10.2307/1166093

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

Fenson, L., Marchman, V., Thal, D., Dale, P., Reznick, J. S., & Bates, E. (2007). MacArthur–Bates Communicative Development Inventories: User’s guide and technical manual (2nd ed.). Brookes. https://doi.org/10.1037/t11538-000.

Fenson, L., Pethick, S., Renda, C., Cox, J. L., Dale, P. S., & Reznick, J. S. (2000). Short-form versions of the MacArthur Communicative Development Inventories. Applied Psycholinguistics, 21(1), 95–116. https://doi.org/10.1017/S0142716400001053

Fernald, A., Perfors, A., & Marchman, V. A. (2006). Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. Developmental Psychology, 42(1), 98–116. https://doi.org/10.1037/0012-1649.42.1.98

Fernald, A., Pinto, J., Swingley, D., Weinberg, A., & McRoberts, G. W. (1998). Rapid gains in speed of verbal processing by infants in the 2nd year. Psychological Science, 9(3), 228–231. https://doi.org/10.1111/1467-9280.00044

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. Journal of Personality and Social Psychology, 78(2), 350–365. https://doi.org/10.1037/0022-3514.78.2.350

Frank, M. C., Braginsky, M., Marchman, V. A., & Yurovsky, D. (in press). Variability and consistency in early language learning: The Wordbank project. MIT Press. https://langcog.github.io/wordbank-book/

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. https://doi.org/10.1017/S0305000916000209

Galeote, M., Checa, E., Sánchez-Palacios, C., Sebastian, E., & Soto, P. (2016). Adaptation of the MacArthur–Bates Communicative Development Inventories for Spanish children with Down syndrome: Validity and reliability data for vocabulary. American Journal of Speech-Language Pathology, 25(3), 371–380. https://doi.org/10.1044/2015_AJSLP-15-0007 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. https://doi.org/10.1111/j.1475-

3588.2008.00503.x

Luyster, R., Lopez, K., & Lord, C. (2007). Characterizing communicative development in children referred for autism spectrum disorders using the MacArthur–Bates Communicative Development Inventory (CDI). Journal of Child Language, 34(3), 623–654. https://doi.org/10.1017/S0305000907008094

Makransky, G., Dale, P. S., Havmose, P., & Bleses, D. (2016). An item response theory–based, computerized adaptive testing version of the MacArthur–Bates Communicative Development Inventory: Words & Sentences (CDI:WS). Journal of Speech, Language, and Hearing Research, 59(2), 281–289. https://doi. org/10.1044/2015_JSLHR-L-15-0202

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. https://doi. org/10.1017/S0305000912000244

Mayne, A. M., Yoshinaga-Itano, C., & Sedey, A. L. (1999). Receptive vocabulary development of infants and toddlers who are deaf or hard of hearing. The Volta Review, 100(5), 29–52. Mayne, A. M., Yoshinaga-Itano, C., Sedey, A. L., & Carey, A. (1998). Expressive vocabulary development of infants and toddlers who are deaf or hard of hearing. The Volta Review, 100(5), 1–28.

Mayor, J., & Mani, N. (2019). A short version of the MacArthur– Bates Communicative Development Inventories with high validity. Behavior Research Methods, 51(5), 2248–2255. https:// doi.org/10.3758/s13428-018-1146-0

Pan, B. A., Rowe, M. L., Spier, E., & Tamis-Lemonda, C. (2004). Measuring productive vocabulary of toddlers in low-income families: Concurrent and predictive validity of three sources of data. Journal of Child Language, 31(3), 587–608. https://doi. org/10.1017/S0305000904006270

Rescorla, L., & Dale, P. S. (2013). Late talkers: Language development, interventions, and outcomes. Brookes.

Rescorla, L., Ratner, N. B., Jusczyk, P., & Jusczyk, A. M. (2005). Concurrent validity of the Language Development Survey: Associations with the MacArthur–Bates Communicative Development Inventories: Words and Sentences. American Journal of Speech-Language Pathology, 14(2), 156–163. https://doi.org/ 10.1044/1058-0360(2005/016)

Rinaldi, P., Pasqualetti, P., Stefanini, S., Bello, A., & Caselli, M. C. (2019). The Italian Words and Sentences MB-CDI: Normative data and concordance between complete and short forms. Journal of Child Language, 46(3), 546–566. https://doi.org/ 10.1017/S0305000919000011

Simonsen, H. G., Kristoffersen, K. E., Bleses, D., Wehberg, S., & Jørgensen, R. N. (2014). The Norwegian Communicative Development Inventories: Reliability, main developmental trends and gender differences. First Language, 34(1), 3–23. https://doi. org/10.1177/0142723713510997

Stolt, S., Haataja, L., Lapinleimu, H., & Lehtonen, L. (2009). Associations between lexicon and grammar at the end of the second year in Finnish children. Journal of Child Language, 36(4), 779–806. https://doi.org/10.1017/S0305000908009161

Szagun, G., Stumper, B., & Schramm, S. A. (2014). Fragebogen zur fruhkindlichen Sprachentwicklung (FRAKIS) und FRAKIS-K (Kurzform) [Questionnaire on Early Language Development (FRAKIS) and FRAKIS–K (Short Form Version)]. Pearson Assessment.

Tardif, T., Fletcher, P., Liang, W.-L., & Kaciroti, N. (2009). Early vocabulary development in Mandarin (Putonghua) and Cantonese. Journal of Child Language, 36(5), 1115–1144. https:// doi.org/10.1017/S0305000908009185

Tardif, T., Fletcher, P., Zhang, Z.-X., & Liang, W.-L. (2008). The Chinese Communicative Development Inventory (Putonghua and Cantonese versions): Manual, forms, and norms. Peking University Medical Press.

Thal, D., DesJardin, J. L., & Eisenberg, L. S. (2007). Validity of the MacArthur–Bates Communicative Development Inventories for measuring language abilities in children with cochlear implants. American Journal of Speech-Language Pathology, 16(1), 54–64. https://doi.org/10.1044/1058-0360 (2007/007) van der Linden, W. J., & Glas, C. A. (2010). Elements of adaptive testing. Springer.

https://doi.org/10.1007/978-0-38785461-8.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. Psychometrika, 54(3), 427–450. https:// doi.org/10.1007/BF02294627

Appendix A

Comparisons of the IRT version and the original model (in parentheses) with different test item sizes on the CDI-WSs and the baseline.

Table A1

Comparison of the IRT version and the original model (in parentheses) with different test item sizes on the American English CDI-WS and the baseline (random list), by gender.

Length	Females			Males			Baseline		
	r with full CDI	Avg.SE	Rel.	r with full CDI	Avg.SE	Rel.	r with full CDI	Avg.SE	Rel.
680	.988 (.988)	.03 (.03)	.999 (.999)	.989 (.989)	.03 (.03)	.999 (.999)	1.000	.00	1.00
400	.990 (.987)	.03 (.03)	.999 (.999)	.990 (.987)	.03 (.03)	.999 (.999)	.998	.01	1.00
200	.988 (.985)	.03 (.04)	.999 (.999)	.989 (.985)	.04 (.04)	.999 (.999)	.993	.02	.999
100	.982 (.979)	.04 (.04)	.998 (.998)	.982 (.978)	.04 (.04)	.998 (.998)	.985	.04	.999
50	.976 (.968)	.05 (.05)	.997 (.997)	.976 (.966)	.05 (.05)	.997 (.997)	.967	.05	.997
25	.963 (.950)	.06 (.07)	.996 (.995)	.964 (.946)	.06 (.07)	.997 (.995)	.936	.07	.994
10	.937 (.884)	.07 (.10)	.994 (.990)	.937 (.873)	.07 (.10)	.994 (.989)	.856	.12	.985
5	.891 (.820)	.11 (.13)	.988 (.982)	.886 (.812)	.10 (.13)	.989 (.982)	.765	.17	.97

Table A2

Correlations from the IRT version and the original model (in parentheses) on the American English CDI-WS, by age group.

Length	16 - 18	19 - 21	22 - 24	25 - 27	28 - 30
680	.97 (.97)	.99 (.99)	1.00 (1.00)	1.00 (1.00)	.98 (.98)
400	.98 (.96)	.99 (.99)	1.00 (1.00)	.99 (1.00)	.98 (.98)
200	.99 (.96)	.99 (.99)	.99 (.99)	.99 (.99)	.98 (.98)
100	.98 (.95)	.99 (.98)	.99 (.99)	.98 (.99)	.98 (.98)
50	.98 (.94)	.99 (.97)	.98 (.98)	.97 (.98)	.97 (.96)
25	.96 (.92)	.98 (.95)	.97 (.96)	.96 (.96)	.95 (.94)
10	.92 (.81)	.95 (.87)	.95 (.90)	.94 (.90)	.92 (.89)
5	.87 (.74)	.92 (.82)	.92 (.84)	.89 (.85)	.84 (.82)

Table A3

Comparison of the IRT version and the original model (in parentheses) with different test item sizes on the Danish CDI-WS and the baseline (random list), by gender.

Length	Females			Males			Baseline		
	r with full CDI	Avg.SE	Rel.	r with full CDI	Avg.SE	Rel.	r with full CDI	Avg.SE	Rel.
725	.982 (.982)	.03 (.04)	.999 (.998)	.983 (.983)	.03 (.04)	.999 (.998)	1.000	.00	1.000
400	.985 (.980)	.03 (.04)	.999 (.998)	.987 (.981)	.03 (.04)	.999 (.998)	.997	.01	1.000
200	.985 (.977)	.03 (.04)	.999 (.998)	.985 (.979)	.04 (.04)	.998 (.998)	.990	.02	.999
100	.981 (.969)	.04 (.05)	.998 (.998)	.981 (.971)	.04 (.05)	.998 (.998)	.978	.03	.999
50	.974 (.957)	.04 (.06)	.998 (.997)	.974 (.956)	.05 (.05)	.998 (.997)	.955	.05	.997
25	.964 (.931)	.05 (.07)	.997 (.995)	.961 (.932)	.05 (.07)	.997 (.995)	.913	.07	.995
10	.924 (.863)	.06 (.09)	.996 (.991)	.939 (.870)	.06 (.09)	.995 (.991)	.807	.12	.986
5	.866 (.792)	.10 (.12)	.989 (.985)	.888 (.801)	.10 (.11)	.989 (.986)	.702	.16	.971

Table A4. Comparison of the IRT version and the original model (in parentheses) with different test item sizes on the Beijing Mandarin CDI-WS and the baseline (random list), by gender.

Length	Females			Males			Baseline		
	r with full CDI	Avg.SE	Rel.	r with full CDI	Avg.SE	Rel.	r with full CDI	Avg.SE	Rel.
799	.976 (.976)	.05 (.06)	.997 (.994)	.974 (.974)	.04 (.05)	.997 (.997)	1.000	.00	1.000
400	.981 (.975)	.05 (.06)	.997 (.994)	.979 (.973)	.05 (.05)	.997 (.997)	.997	.01	1.000
200	.980 (.971)	.05 (.07)	.997 (.994)	.978 (.970)	.06 (.06)	.996 (.996)	.993	.02	1.000
100	.969 (.964)	.06 (.07)	.996 (.994)	.974 (.968)	.06 (.06)	.996 (.996)	.983	.03	.999
50	.957 (.950)	.06 (.07)	.995 (.993)	.967 (.959)	.07 (.07)	.995 (.995)	.965	.05	.998
25	.942 (.930)	.07 (.08)	.995 (.991)	.955 (.947)	.07 (.07)	.994 (.994)	.932	.07	.995
10	.916 (.871)	.08 (.11)	.994 (.987)	.930 (.902)	.09 (.09)	.992 (.991)	.852	.11	.987
5	.873 (.790)	.10 (.13)	.990 (.979)	.893 (.826)	.10 (.13)	.989 (.983)	.754	.16	.974

Table A5. Comparison of the IRT version and the original model with flexible approach in fitting of polynomial (in parentheses) with different test item sizes on the Italian CDI-WS and the baseline (random list), by gender.

Length	Females			Males			Baseline		
	r with full CDI	Avg.SE	Rel.	r with full CDI	Avg.SE	Rel.	r with full CDI	Avg.SE	Rel.
670	.993 (.993)	.02 (.03)	.999 (.998)	.996 (.996)	.03 (.03)	.997 (.999)	1.000	.00	1.000
400	.992 (.992)	.03 (.04)	.999 (.998)	.995 (.994)	.04 (.03)	.997 (.999)	.998	.01	1.000
200	.987 (.989)	.04 (.04)	.998 (.998)	.990 (.990)	.05 (.04)	.996 (.998)	.992	.02	.999
100	.976 (.983)	.05 (.05)	.997 (.997)	.981 (.981)	.06 (.05)	.996 (.997)	.983	.04	.999
50	.965 (.970)	.06 (.06)	.995 (.996)	.971 (.962)	.07 (.06)	.994 (.996)	.964	.05	.997
25	.954 (.950)	.07 (.08)	.994 (.994)	.960 (.939)	.08 (.08)	.993 (.993)	.929	.08	.994
10	.943 (.877)	.08 (.11)	.993 (.987)	.931 (.862)	.08 (.11)	.992 (.986)	.840	.12	.984
5	.912 (.797)	.10 (.15)	.990 (.976)	.895 (.765)	.10 (.16)	.988 (.973)	.740	.18	.967