Early vocabulary inventory for Mandarin Chinese

MEILING HAO
Beijing Language and Culture University, Beijing, China

HUA SHU AND AILING XING
Beijing Normal University, Beijing, China

AND

PING LI
Pennsylvania State University, University Park, Pennsylvania
and National Science Foundation, Arlington, Virginia

Early vocabulary development is a reliable predictor of children's later language skills. The MacArthur–Bates Communicative Development Inventory (CDI) has provided a powerful tool to assess early vocabulary development in English and other languages. However, there have been no published CDI norms for Mandarin Chinese. Given the importance of large-scale comparative data sets for understanding the early childhood lexicon, we have developed an early vocabulary inventory for Mandarin. In this article, we report our efforts in developing this instrument, and discuss the data collected from 884 Chinese families in Beijing over a period of 12–30 months, based on our instrument. Chinese children's receptive and expressive lexicons as assessed by our inventory match well with those reported for English on the basis of CDI. In particular, our data indicate comprehension–production differences, individual differences in early comprehension and in later production, and different lexical development profiles among infants versus toddlers. We also make the checklists and norms of our inventory available to the research community via the Internet; they may be accessed from the Psychonomic Society's Archive of Norms, Stimuli, and Data, at www.psychonomic.org/archive.

Early vocabulary development is not only an important milestone in children's language development; it is also a reliable predictor of children's later language skills. The study of this subject has significant implications, therefore, not only for normal language development, but also for language delays and other abnormally developing patterns in the clinical context. In response to the significance of the early childhood lexicon, researchers have developed vocabulary norms in a variety of languages. Among the existing vocabulary norms, the parental report has been the most widely used instrument, because it is less time consuming, easy to conduct, and easy to obtain in large quantities in a relatively short amount of time. As its name implies, parental report requires parents to judge whether their children can understand or produce the words in a given wordlist. Fenson et al. (1994) and Dale and Fenson (1996) pioneered this line of work. They named their parental report norm the Communicative Development Inventory (CDI),[1] with which they investigated American children's comprehension and production vocabulary between the ages of 0;8 (8 months) and 2;4 (2 years 4 months). They have also made available a Web-based interface that contains the English and Spanish lexical norms at www.sci.sdsu.edu/cdi/.

Building on the original CDI, a number of researchers have developed different versions of the CDI in other languages and linguistic variants, such as British English (Hamilton, Plunkett, & Schafer, 2000), Spanish (Jackson-Maldonado, Thal, Marchman, Bates, & Gutierrez-Clellen, 1993), and New Zealand English (Reese & Read, 2000). However, there are no published CDI norms yet for Mandarin Chinese. Given the language-specific properties of Chinese, and the differences between Chinese and Indo-European languages in acquisition (Li, Tan, Bates, & Tzeng, 2006), it is important to develop similar measures for Mandarin Chinese.

A few studies have used parental reports to evaluate Chinese children's early vocabulary development. Wu (1997) used a version of CDI in her dissertation research of Chinese children's lexical development. However, the extent to which her results truly capture early Chinese lexicon is questionable, because her child participants all resided in the U.S.A. (30 children of Chinese immigrants in Boulder, Colorado) and their parents had mixed linguistic and cultural backgrounds (from Mainland China, Taiwan, and Singapore). Many were bilingual Chinese–English speakers or spoke English as a second language. In addition, some items in the questionnaire list were unknown to Mandarin-speaking children in Main-

P. Li, p-li@nsf.gov

land China—for example, 脚踏车 (*jiaotache*, bicycle); 窗户帘 (*chuanghulian*, curtain); 邮差 (*youchai*, post-man); 警员 (*jingyuan*, policeman).[2] Some were also low-frequency words in the spoken language, such as 打击 (*daji*, hit); 关闭 (*guanbi*, close); 摇动 (*yaodong*, move); and 攀登 (*pandeng*, climb). Part of the problem might be that Wu simply translated the original English word list into Chinese, so it cannot be used as a reliable tool to faithfully assess Mandarin-speaking children's early vocabulary development. In light of this problem, Tardif and her colleagues (Liang et al., 2001; Tardif, Gelman, & Xu, 1999) developed their own versions of Chinese CDI to examine children learning Mandarin and Cantonese. Although Tardif et al.'s Chinese CDI forms represented an improvement over Wu's, two problems still limited the use of their CDIs. One is the same as that in Wu's CDI; namely, the problem that some words were simply translations of the original English words and were therefore not colloquial to Mandarin-speaking children. The other is that the actual lexical norms are not as readily available to researchers as are the English lexical norms. Thus, it is difficult to either use or validate their norms in language acquisition in a variety of research contexts.

In this article, we report data collected from 884 Chinese families in Beijing (children's age range, 12–30 months) and describe the ways in which we constructed the early vocabulary inventory for Mandarin Chinese. We attempt to correct the problems seen in the Wu (1997) and Tardif et al. (1999) versions of the Chinese CDI in order for our word lists to reflect the language characteristics and faithfully capture early vocabulary development in Mandarin Chinese. It is important to consider validity issues in constructing early vocabulary inventories, and our study indicates that a significant portion of the words are not shared between English and Chinese for the early vocabulary (see the Method section and the archived materials). At the same time, we attempt to make the actual lexical norm readily available to the general research community, via the Internet.

METHOD

Construction of the Chinese Vocabulary Checklist

To achieve high validity of our instrument, we followed the following steps closely to derive the word checklists for our early vocabulary inventory.

First, one student trained in linguistics wrote down all the possible Chinese equivalents of the lexical items in the English CDI according to Fenson et al. (1994). Items rarely used in Chinese daily life were excluded—for example, *church*, *farm*, *lawnmower*. Items frequently used in Chinese daily life, such as 熊猫 (*xiongmao*, panda), 稀饭 (*xifan*, porridge), and 排骨 (*paigu*, short ribs),[3] were added according to Dong's (1998) classified dictionary.

Second, two mothers trained in developmental psychology revised the word list. One mother had a child 22 months old and another had a 3-year-old. In the course of this revision, they either excluded or reformulated the words that Chinese children and their caregivers seldom

use, even though such words may be frequent in Chinese daily life. For example, the official Chinese word for *crib* is 婴儿床 (*yingerchuang*), but Chinese mothers often say simply 床 (*chuang*, bed).

Third, seven mothers whose children, between 17 and 30 months of age, went through the remaining words to identify whether their children could understand or produce these words. They were also asked to write down additional words that their children could say but that were not included in the checklist.

Our Mandarin inventory includes two parts, following the example of the original English CDI: Part 1 is the infant checklist, used for infants 12 to 16 months of age. It is divided into the following 14 categories: action words; animals; body parts; clothing; descriptive words; food and drink; furniture and rooms; games and routines; outside things and places to go; people; pronouns; small household items; toys; and vehicles. We did not include the following 4 categories: sound effects and animal sounds; quantifiers; question words; and words about time. These categories were not included, either because these words were difficult to match between languages (e.g., sound effects) or because they would not be acquired by infants before the age of 17 months (e.g., quantifiers).

Part 2 is the toddler checklist, used for children 17 to 30 months of age. It includes the following 20 categories: action words; animals; body parts; clothing; connecting words; descriptive words; food and drink; furniture and rooms; games and routines; helping verbs; outside things; people; places to go; pronouns; quantifiers and articles; question words; small household items; toys; vehicles; and words about time.

The infant checklist includes 232 words, among which 160 items (69%) are translation equivalents of the items in the English CDI. Next to each word on the list parents are asked to indicate whether their child (1) understands the word and (2) produces the word. The toddler checklist includes a total of 710 words, among which only 407 (57%) are translation equivalents of the English CDI items. For the toddler list, parents are only asked to indicate whether their children have ever said the word, as is done for the English CDI. The number of words in the toddler list is comparable to that in the English CDI (680 words), although the infant checklist contains fewer words than the English CDI (396 words).

The complete infant checklist and toddler checklist, along with the English glosses, are shown in the archived Appendix. For the infant list, there are often alternative forms for the same name in parent–child interactions, and therefore all the alternative forms are presented in the questionnaire for parental report.

Data Collection

All the forms were completed by parents in Beijing 301 Hospital, Beijing Haidian Hospital, and Tieying Hospital, where their children received regular physical checkups. A trained nurse helped to collect the questionnaires. Data were excluded if the child had any or all of the following conditions: premature birth; exposure to a second lan-

guage; or insufficient birth information. Parents provided consent to the data collection procedure, understanding that they were free to withdraw from the process at any time during the completion of the inventory forms.

Below, we report data collected from 884 families whose parents completed the inventory forms, with 231 forms for the infant checklist and 653 forms for the toddler list. For each age (month) group, there were around 47 parents who completed the forms. The sample sizes for each age (month) group are shown in Tables 1 (infants) and 2 (toddlers). The number of boys and girls was roughly the same for each age group.

RESULTS

Infant Checklist

Table 1 provides the number of words that infants can understand or produce on average for each month from 12 to 16 months, and the maximum and the minimum of the number of words that they can understand or produce during the period. These results show, not surprisingly, that children's receptive vocabulary exceeds their expressive vocabulary (i.e., comprehension generally proceeds production). Children make more progress in receptive vocabulary during this period (from a mean of 81 words to one of 115), but less progress on expressive vocabulary (from a mean of 5.07 words to one of 16.82). Individual differences in receptive vocabulary are larger than those in expressive vocabulary, as reflected in the standard deviation (*SD*) and the range.

For the infant checklist, each child had a comprehension score and a production score. The median, 10th, 25th, 75th, and 90th percentiles of comprehension and production vocabulary found for each age group are shown in Figures 1 and 2.

Figure 1 shows the development of receptive vocabulary from 12 to 16 months, for children at each of the percentiles. As is also indicated in Table 1, children's vo-

Table 1
Mean Number of Words (With Standard Deviations) in the Receptive and Expressive Vocabularies of Chinese Infants

Age (months)	Receptive (Comprehension)					Expressive (Production)				
	M	*SD*	*N*	Max.	Min.	*M*	*SD*	*N*	Max.	Min.
12	81	37	45	162	28	5	6	45	23	0
13	90	49	50	193	11	7	9	50	32	0
14	88	50	48	194	20	8	11	48	47	0
15	101	60	47	215	30	15	15	47	46	0
16	116	50	41	224	31	17	17	41	70	0

Table 2
Mean Number of Words (With Standard Deviations) in the Expressive Vocabularies of Chinese Toddlers

Age (months)	*M*	*SD*	*N*	Max.	Min.
17	101	104	40	383	3
18	113	98	50	337	4
19	168	114	50	420	3
20	181	147	50	516	4
21	224	165	49	609	6
22	222	143	46	606	23
23	367	206	50	705	32
24	334	193	50	700	37
25	360	203	43	699	29
26	376	189	47	700	93
27	444	181	44	686	99
28	450	187	44	709	87
29	497	167	49	710	141
30	536	160	41	708	165

cabulary development during this age period shows large individual differences, consistent with findings based on the original English CDI (Bates et al., 1994). For example, at 12 months infants at the 10th percentile can understand only 17% of the vocabulary in our inventory (39 words), but infants at the 90th percentile can understand 46% of the vocabulary in the inventory (132 words). In contrast to the individual differences, however, growth rate is lim-
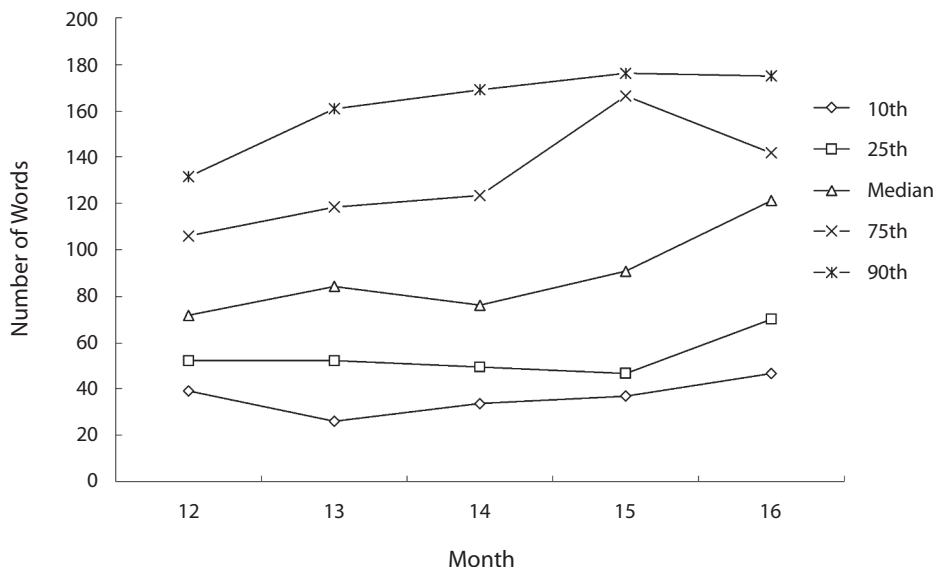

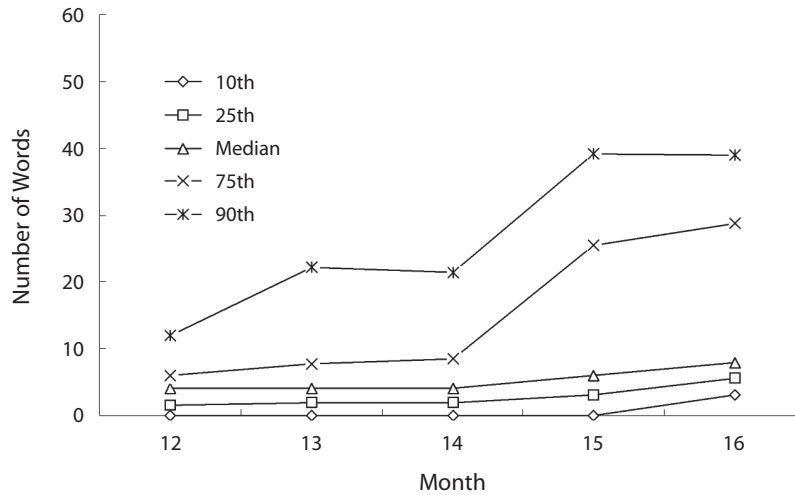Figure 1. Receptive vocabulary of Chinese infants (12–16 months).

Figure 2. Expressive vocabulary of Chinese infants (12–16 months).

ited for this period. Children at the median can understand 31% of the vocabulary (72 words) at 12 months and 52% (121 words) at 16 months.

Figure 2 shows the development of expressive vocabulary from 12 to 16 months. Unlike the comprehension profiles shown in Figure 1, the individual variation for production is relatively small for children below the 50th percentile. On the other hand, children above the 50th percentile show both clear growths in productive vocabulary and individual differences. For example, at 12 months, even children at the 90th percentile can produce only about 5% of the inventory words (12 words), whereas at 16 months these children can produce up to 17% of the vocabulary (39 words).

Toddler Checklist

Table 2 provides the number of words that toddlers can produce on average for each month from 17 to 30 months,

and the maximum and minimum of the number of words that they can produce during the period. These results show that children experience a substantial vocabulary growth from 17 to 30 months of age. For example, children at 17 months can produce only about 100 words on average, but can produce more than 500 words at 30 months, a fivefold increase in a little bit over a year.

Figure 3 shows the median, 10th, 25th, 75th, and 90th percentiles of expressive vocabulary for each age group from 17 to 30 months. These results show that there is a considerable amount of individual variation in production for this growth period. Children at the 10th percentile can produce only 2% of the vocabulary in the inventory (15 words) at 17 months, whereas at the 90th percentile they can produce 36% of the vocabulary (252 words). The individual variation remains large, even when children grow older. At 30 months, children at the 10th percentile
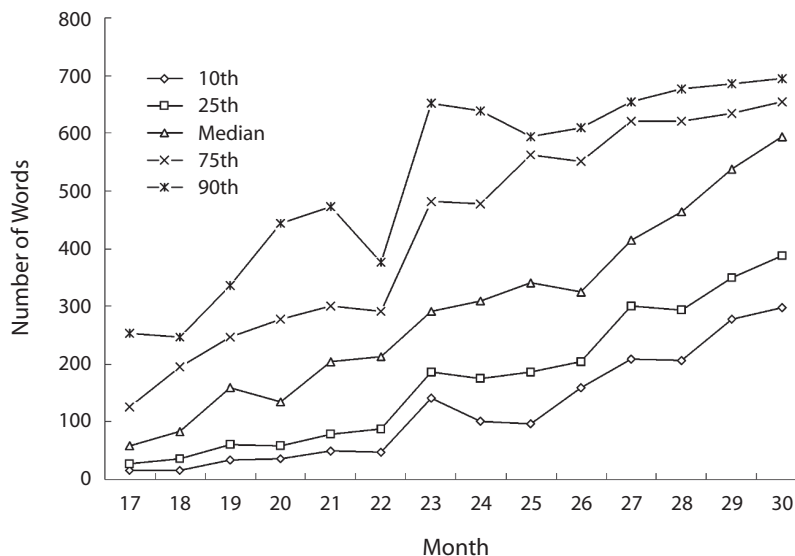

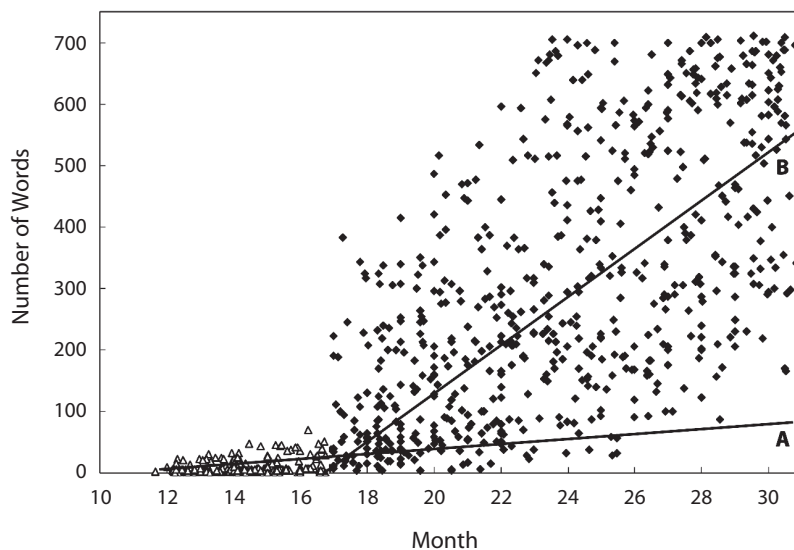**Figure 3. Expressive vocabulary of Chinese toddlers (17–30 months).**

Figure 4. Expressive vocabulary of Chinese children and a regression analysis of the infant (12–16 months, line A) and the toddler (17–30 months, line B) data. Triangles denote the infant data; dots denote the toddler data.

can produce 42% of the vocabulary in the inventory (298 words), whereas at the 90th percentile they can produce 98% of the vocabulary (694 words).

Productive Patterns in Infants Versus Toddlers

Consistent with Table 2, Figure 3 also indicates a rapid and substantial growth in Chinese children's vocabulary acquisition between 17 and 30 months of age. For example, children at the median can produce 8% of the vocabulary (59 words) at 17 months, but they can produce 84% (594 words) at 30 months. This rapid increase in vocabulary is consistent with patterns found with English-speaking children in the original CDI (Bates et al., 1994; Fenson et al., 1994), and also consistent in general with the "vocabulary spurt" phenomenon reported in the literature (Hamilton et al., 2000; see a recent analysis in Li, Zhao, & MacWhinney, 2007). On the other hand, the rapid increase during this period is in sharp contrast to the slow growth in the 12- to 16-month period shown by the infant data. To see this contrast more clearly, in Figure 4 we plot the expressive vocabulary data from both the infant checklist and the toddler checklist. Each data point represents the number of words (y-axis) that a given child can produce at a given month (x-axis), from a total of 884 children assessed by our inventory. A linear regression was run separately on the data from the 12- to 16-month period, and on the data from the 17- to 30-month period, as shown in line A and line B. This analysis shows distinct growth profiles for the two age periods, with slow growth followed by rapid increase in the overall size of vocabulary, and small variation followed by large individual differences.

CONCLUSION

Early vocabulary development is a reliable predictor of children's later language skills. Given the significance

of the early childhood lexicon, and the tools available to assess it, it is important that we develop large-scale comparative data sets for the understanding of lexical development in a variety of languages. CDI has provided a powerful tool in this regard, and it has been applied to the study of several languages, and for a variety of questions (see Goodman, Dale, & Li, 2008, for a recent analysis of the role of frequency with CDI).

In this article, we report our efforts in developing a similar inventory for Mandarin Chinese, and discuss the data collected, using this new instrument, from 884 Chinese families in Beijing with children from 12 to 30 months of age. We have attempted to correct problems seen in previous efforts in using the CDI to study early vocabulary acquisition in Chinese-speaking children. Chinese children's receptive and expressive lexicons as assessed by our instrument match well with those reported in the English CDI analyses. In particular, our data indicate comprehension–production differences, individual differences in early comprehension and in later production, and different lexical development profiles among infants (12–16 months—relatively slow growth) versus toddlers (17–30 months—much more rapid increase).

In developing the Mandarin early vocabulary inventory, we also have in mind the aim of making the actual lexical norms available to the general research community over the Internet. Readers of this article should consult our Web site (cogsci.richmond.edu) and the Psychonomic Society Archive of Norms, Stimuli, and Data (www.psychonomic.org/ archive) for updated versions of the lexical norms discussed in this article.

REFERENCES

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

Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers*, **28**, 125-127.

Dong, D. N. (1998). *Classified Dictionary of Contemporary Chinese*. Chinese Dictionary Press, Shanghai.

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**, v-179.

Goodman, J., Dale, P., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, **35**, 1-17.

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

Jackson-Maldonado, D., Thal, D., Marchman, V., Bates, E., & Gutierrez-Clellen, V. (1993). Early lexical development in Spanish-speaking infants and toddlers. *Journal of Child Language*, **20**, 523-549.

Li, P., Tan, L., Bates, E., & Tzeng, O. (2006). New frontiers in Chinese psycholinguistics: An introduction. In P. Li, E. Bates, L.-H. Tan, & O. Tzeng (Eds.), *Handbook of East Asian psycholinguistics* (Vol. 1, pp. 1-8). Cambridge: Cambridge University Press.

Li, P., Zhao, X., & MacWhinney, B. (2007). Dynamic self-organization and early lexical development in children. *Cognitive Science*, **31**, 581-612.

Liang, W., Hao, B., Wang, S., Yang, Y., Zhang, Z., Zuo, Q., et al. (2001). The study of Chinese language and communicative development-word part. *Chinese Journal of Child Health Care*, **9**, 295-297.

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

Tardif, T., Gelman, S. A., & Xu, F. (1999). Putting the "noun bias" in context: A comparison of English and Mandarin. *Child Development*, **70**, 620-635.

Wu, J. (1997). *Language, play and general development for Chinese infant-toddlers: Using adapted assessments*. Unpublished doctoral dissertation, University of Colorado, Boulder.

NOTES

1. It is also called MCDI, since the work was sponsored by the MacArthur Foundation. It was later named MacArthur–Bates Communicative Development Inventory, in memory of Elizabeth Bates's contribution to language development research.

2. Looking at the English translations of the Chinese words might be misleading, since neither the register nor the frequencies associated with the specific Chinese words is reflected in the English translations.

3. The classified dictionary is organized by semantic categories. Words belonging to the same semantic category are listed together, which helped us to find all the possible items within a given semantic category.

ARCHIVED MATERIALS

The following materials may be accessed through the Psychonomic Society's Norms, Stimuli, and Data archive, www.psychonomic.org/archive.

To access these files, search the archive for this article using the journal name (*Behavior Research Methods*), the first author's name (Hao), and the publication year (2008).

File: Li-BRM-2008.zip

Description: The compressed archive file contains three files:

infant_data.xls, containing the percentages of children reported understanding and producing words between the ages of 12 and 16 months;

toddler_data.xls, containing the percentages of children reported producing words between the ages of 17 and 30 months;

appendix.pdf, containing the infant checklist and toddler checklist in our early vocabulary inventory for Mandarin Chinese.

Corresponding author's e-mail address: p-li@nsf.gov.