3-2017

Research to Establish the Validity, Reliability, and Clinical Utility of a Comprehensive Language Assessment of Mandarin

Xueman Lucy Liu
University of Texas at Dallas

Jill de Villiers
Smith College, jdevilli@smith.edu

Chunyan Ning
Tianjin Normal University

Eric Rolfhus
Bethel Hearing and Speaking Training Center

Teresa Hutchings
Bethel Hearing and Speaking Training Center

See next page for additional authors

Recommended Citation

Liu, Xueman Lucy; de Villiers, Jill; Ning, Chunyan; Rolfhus, Eric; Hutchings, Teresa; Lee, Wendy; Jiang, Fan; and Zhang, Yi Wen, "Research to Establish the Validity, Reliability, and Clinical Utility of a Comprehensive Language Assessment of Mandarin" (2017). Philosophy: Faculty Publications, Smith College, Northampton, MA.
https://scholarworks.smith.edu/phi_facpubs/42

Running Head: Assessment of Mandarin

“Research to Establish the Validity, Reliability and Clinical Utility of a Comprehensive Language Assessment of Mandarin”

Xueman Lucy Liu
University of Texas at Dallas, Communication Sciences and Disorders
Bethel Hearing and Speaking Training Center, Research,
Dallas, Texas, United States

Jill de Villiers
Smith College, Psychology and Philosophy
Northampton, Massachusetts, United States

Bethel Hearing and Speaking Training Center, Research
Dallas, Texas, United States

Chunyan Ning
Tianjin Normal University, Institute of Linguistics
Tianjin, China

Bethel Hearing and Speaking Training Center, Research
Dallas, Texas, United States

Eric Rolfhus
Bethel Hearing and Speaking Training Center, Research
Dallas, Texas, United States

Teresa Hutchings
Bethel Hearing and Speaking Training Center, Research
Dallas, Texas, United States

Wendy Lee
University of Texas at Dallas, Communication Sciences and Disorders
Bethel Hearing and Speaking Training Center, Research
Dallas, Texas, United States

Fan Jiang
Shanghai Children's Medical Center Affiliated to Medical School of Shanghai Jiao Tong University,
Department of Developmental and Behavioral Pediatrics
Shanghai, China

Yi Wen Zhang
Shanghai Children's Medical Center Affiliated to Medical School of Shanghai Jiao tong University, Department of
Developmental and Behavioral Pediatrics
Shanghai, China

Abstract

Purpose

With no existing *gold standard* for comparison, challenges arise for establishing validity of a new standardized Mandarin language assessment normed in mainland China.

Methods

A new assessment, Diagnostic Receptive and Expressive Assessment of Mandarin (DREAM)[1], was normed with a stratified sample of 969 children (ages 2;6–7;11) in multiple urban and non-urban regions in northern and southern China. In this study of 230 children the sensitivity and specificity of DREAM were examined against an a priori judgment of disorders. External validity was assessed using two indices of language production for different age groups.

Results

External validity was assessed against spontaneous language indices (correlations range r=0.6 to 0.7, all p<.01) and narrative indices (overall r=0.45, p<.01). Sensitivity (0.73) and specificity (0.82) of DREAM are moderate to good using a priori judgment as the standard. The values improved to 0.95 and 0.82 when spontaneous language and narratives were added to a priori judgment to define typicality. Divergent validity was moderate with non-linguistic indices.

Conclusions

DREAM holds promise as a diagnostic test of Mandarin language impairment for children aged 2;6–7;11.

Introduction

The Background Problem

Children with specific language impairment (SLI) are clinically defined as having language skills one standard deviation or more below the age-group mean, with a non-verbal IQ being 85 or above, and without medical or neurological diagnoses, such as hearing impairment (Rice, 2013; Tomblin et al., 1997). It is a matter of some controversy whether there is a genuine category under the term SLI, or whether more attention needs to be paid instead to the problem of language impairment in a broader sense (Reilly et al., 2014). Regardless of the answer to this question, it is true that children with hearing loss, or with genetic disorders such as Down Syndrome, are easily recognized as needing language services by the medical profession, whereas children who only have a language problem are likely to be missed.

The prevalence of SLI for kindergartners in the upper Midwestern region of the United States was found to be 7.4% overall, 6% for females, and 8% for males (Tomblin et al., 1997)[2]. Assuming that the impairment is a worldwide phenomenon with some significant genetic component (Rice, 2013), applying the United States prevalence rate to the 2010 Chinese census (China Data Center, 2012) would suggest there are in China approximately 230,000 children with SLI just between five and six years of age who are currently in need of identification and rehabilitation services. The profession of speech-language pathology is just beginning in mainland China, so there are few trained clinical professionals with the linguistic knowledge to assess a child with language impairments, as well as a lack of language assessments that meet validity and reliability standards. There have been some attempts by pediatricians to develop language screeners and assessments for early identification of a possible language disorder among mainland Chinese children. For example, there is a checklist—the Infant and Child

Language Development Screener (Zhang, Jin, & Shen, 2003)—for early functional language communication from birth to 36 months that has items such as “makes noise while smiling” and “is able to speak a simple sentence”. This screener has been tested on over 8,000 children in Shanghai, China and is currently being used by multiple pediatric hospitals in China. The Mandarin MacArthur-Bates Communicative Development Inventories (Tardif, Fletcher, Zhang, Liang, & Zuo, 2008) was adapted from the MacArthur-Bates Communicative Development Inventories (Fenson et al., 2007) and normed only for the Beijing region. It is a vocabulary checklist completed by parents of children between birth to 30 months. However, researchers have reported that late talkers often catch up to their peers in language skills by 3 to 5 years of age (Leonard, 2014), while children with a language disorder do not. As a result, it is important that diagnostic tests be available for children older than three years. The problem in China is that there are until now no formal standardized and comprehensive language assessment tools normed in mainland China that meet psychometric standards (Friberg, 2010) to diagnose whether children have a language impairment when an overt medical diagnosis, such as hearing impairment, is not present.

Designing a New Test for Mandarin

To fill the need in China, the DREAM test (Ning, Liu, & de Villiers, 2014) was developed as a standardized, norm-referenced language assessment representative of Mandarin speaking children aged 2;6–7;11 years old in mainland China. We will first discuss the specific challenges associated with designing a new test in a language, and how items were developed for DREAM.

First, it is difficult to begin the process of developing a new test when there are no existing systematic assessments. In English, language acquisition has been studied in great depth for about the past 50 years, leading to a very good empirical knowledge base of what children at different ages know and can answer in natural circumstances of language interaction (e.g., the extensive databases in CHILDES: MacWhinney & Snow, 1985). In parallel with naturalistic studies, experimental work has been carried out testing particular aspects of knowledge, such as relations among lexical concepts, understanding of tense, or production of relative clauses (Crain & Thornton, 1998; de Villiers & Roeper, 2011; Guasti, 2004; Hoff, 2014; McDaniel, McKee, & Cairns, 1996). Furthermore, large numbers of assessments have been developed and normed, including tests of vocabulary for age 2 to 90 years (Dunn & Dunn, 2007), and batteries of syntax, morphology, semantics, and pragmatics (Hresko, Reid, & Hammill, 1999; Semel, Wiig, & Secord, 1996; Seymour, Roeper, & de Villiers, 2005; Zimmerman, Steiner, & Pond, 2002). Hence, in developing a new assessment in English, a wealth of gold standard data already exists against which to examine the validity of the new items. Nevertheless, even in the United States, issues arise about the appropriate standard for children who are dual language learners (Iglesias, 2015) and for children who speak non-mainstream dialects of English (de Villiers & de Villiers, 2010).

Not only are there no gold standard assessments in Mandarin, but the knowledge base about normal language acquisition is considerably less well established. Yet it is a mistaken approach to translate assessment tests from English into another language, because languages differ in significant ways (Peña, 2007). Semantic constructs may be expressed straightforwardly in one language but in a highly complex way, or not at all, in another language. Grammar development can take quite different paths in different languages (Slobin, 1986).

To begin to solve this problem a large variety of test items was designed and piloted in the Beijing/Tianjin area by a team comprised of linguistic experts in Mandarin from mainland China, western-trained bilingual speech-language pathologists, and experts in assessment development. The items were selected to represent the variety of linguistic forms and structures that are acquired by typically developing children from age 2 through 8 who speak Mandarin (Cheng, 1988; Lee, 1982, 1986, 1992; Lee & Naigles, 2008; Li, Huang, & Hsiao, 2010; Liu, 2009; Liu & Ning, 2009; Zhou, 2002, 2004; Zhou & Crain, 2009, 2011). Items were chosen not only by close attention to the empirical and theoretical literature on language development, but also by considering evidence on the nature of language deficits in childhood. The research team took care to reflect properties of Mandarin that could create challenges for a child with language impairment (Cheung, 2009; Fung, 2009).

Mandarin is radically different from English in several ways. Inflectional and derivational morphology—standbys for English language tests—are virtually nonexistent in Mandarin, where in contrast, compounding of morphemes is common. Unlike English, the time markers in Mandarin focus on Aspect, not Tense. Discourse allows the dropping of both subject (S) and object (O) noun phrases, and even the verb (V). Wh-phrases do not move, they remain *in situ.* Nouns often require classifiers, but there are no determiners. The passive takes several forms, but controversy rages about its underlying structure. About the only thing Mandarin and English grammars have in common is a preference for S-V-O word order.

The content of the items went through a rigorous selection and testing procedure. The first focus was on the unique aspects of Mandarin (e.g. classifiers) for which small-scale experimental studies have been conducted, allowing estimates of the age at which the constructions might be mastered, and also clues about the likely errors a child might make in production or comprehension. (Chien, Lust & Chiang, 2003; Li et al., 2010).

Secondly, the DREAM includes items recognized as creating difficulties for children with language impairment, such as fast mapping of novel words (de Villiers & Johnson, 2007; Rice, Buhr, & Nemeth*,* 1990), wh-questions (de Villiers, Roeper, Bland-Stewart, & Pearson, 2008), tense/aspect markers (Duchesne, 2015; Rice & Wexler, 1996), and embedded clauses (Friedmann & Novogrodsky, 2004).

Thirdly, a specific area of focus was to consider measures of process. The problem with measures of what has been achieved is that they can fail to distinguish between children who have genuine difficulties in learning the language, despite adequate exposure, from those who are delayed because their exposure to sufficiently varied and complex language has been inadequate (Hirsh-Pasek, Kochanoff, Newcombe, & de Villiers, 2005; Rice, Buhr, & Nemeth, 1990).

Fourthly, we considered the variation in the language—even the Mandarin spoken in different regions. According to Sun (2006), China is home to at least seven mutually unintelligible dialects and numerous additional regional languages mostly associated with minority peoples. However, Mandarin seems to be the place to start because it is China's educational policy in urban and suburban regions for teachers to use Mandarin exclusively in schools for children age 3 and above. Parents filled out a questionnaire regarding the child's exposure to and use of other dialects and languages. Only children who were reported to speak Mandarin could be included, though it was common for members of the household to speak other local dialects. Much care was taken to avoid lexical items or expressions that might be biased against or in favor of certain dialects and not others. As a check, the assessment was administered across different urban and suburban regions, as well as northern and southern dialect regions, and analyses were conducted to finalize the items on the test by ensuring that they were not biased by the dialect the child spoke or region of China from which the child came. A differential item functioning (DIF) analysis was used to determine if an item on an assessment may be biased against a subgroup of respondents due to characteristics of the item that are not related to the construct being assessed (Zumbo, 2007). Items were flagged for possible DIF using two approaches, Mantel-Haenszel (Holland & Thayer, 1988) and item response theory (Rasch DIF in Winsteps, Linacre, 2014). Approximately 3% of the originally developed item pool was flagged and modified based on this process. Revised items were subsequently included in standardization and tested again for evidence of DIF to ensure item modifications were effective. Nevertheless, we fully recognize that this is only the first of a series of potential tests that will be needed to maximize fair testing across China.

A final consideration in item design was the performance of the items throughout the age range. Within any given type, it was necessary to design easy and difficult items that might discriminate at different points in development.

Standardization

Detailed information about the piloting and standardization of the DREAM test is included in the assessment manual, which is available upon request from the first author. After extensive piloting and tryout in different regions within mainland China, a final set of items was selected by subjecting the data to Rasch analyses (a variety of item response theory; Embretson & Reise, 2000). In a Rasch analysis, “ability” is defined by the number of items a child gets correct, and “item difficulty” by the number of children who get the item incorrect. As in classical test theory, the overall index of whether the test is satisfactory is measured by the degree to which the items cohere in providing the same estimate of a child’s relative ranking. Rasch psychometrics have the advantage of being less affected by the properties of the sample.

Well-behaved test items are selected that spread across the range of abilities in terms of their discriminative potential, and are not misfits; for example, they do not create undesirable u-shaped developmental curves.

The linguistic components were compiled into four composites: Expressive score, Comprehension score, Syntax score (adding across modes) and Semantics score (adding across modes). The table in Appendix A shows a breakdown of the major parts of the test and the number of items finalized for each of the receptive and expressive subcomponents. It also lists some examples from the subtests.

The nationally representative standardization sample consisted of 969[3] Mandarin-speaking children between the ages of 2;6 and 7;11, with equal numbers of boys and girls. Between 2;6 and 5;11 years, half-year age groups were distinguished, with year-long age groups for 6 and 7 year olds. Sampling included multiple cities and suburbs in both the northern and southern regions of China, and was stratified by multiple variables such as age, gender, urban versus suburban, region, and highest primary caregiver education level, according to the most recent census data (China Data Center, 2012).

Reliability and Validity Studies and the Establishment of Clinical Utility

Data from standardization provided estimates of the internal consistency reliability of the DREAM Total scale (Cronbach's alpha=0.94; N=969) and test-retest reliability over a 2–4 week period (r=0.85; N=60).

In the normative sample, females demonstrate a 4.2 point advantage on DREAM total scaled score, on average. These results are in keeping with the general finding world-wide that boys are more likely to have language delay than girls (Snowling, Duff, Nash, & Hulme, 2015; Zambrana, Pons, Eadie, & Ystrom, 2014). In DREAM data this advantage is less pronounced at younger ages (<4.0 years, 1.2 points), and more so at older ages (>=4.0 years, 5.6 points), in keeping with other findings that boys may have more persistent language disorder.

The item development process and dialect differential analysis indicated that DREAM has good content validity. The standardization phase revealed that DREAM has appropriate psychometric properties, including internal reliability and test-retest reliability, with a large sample of children aged 2;6–7;11 years.

The Present Study

The quality of a test is also judged by its ability both to identify children in need of language intervention services (sensitivity), and not to select those who do not need such services (specificity). The study we report next will evaluate the sensitivity and specificity of DREAM in assessing children with potential language impairment. The question that needs to be addressed is how to establish the groups of typical and atypical children in order to judge the specificity and sensitivity. We assess two approaches. One uses the judgment of pediatricians, who usually make such decisions in China, solely based on a thorough parent interview. The second adds to this by using language samples to further refine the pediatricians' decisions. Language sampling and linguistic analyses are often recommended to assess language impairment when there is no established instrument, but the process is time consuming and requires expertise. Nevertheless, results from language sampling can be used for examining convergent validity of a new instrument, especially when there are no other suitable measures (Bedore, Pena, Gillam, & Ho, 2010; Pearson, Jackson & Wu, 2014). The samples also provided a way to refine the discrimination of the pediatricians' decisions.

Method

Participants

Three hundred children aged 2;6–7;11 were recruited for this study at a major urban pediatric hospital, Shanghai Children's Medical Center in China. This hospital serves as a central agency for evaluating children with developmental problems, including language development. All 300 children received a regular physical examination from the pediatricians at the Medical Center. If a concern about the child's communication was expressed by the parents or teachers, pediatricians then evaluated the child through a thorough informal interview addressed to the parents. Ninety-four children were classified as possibly having atypical language development according to pediatricians' judgment based on the parent interview without any assistance from comprehensive language assessments. These children were termed the a priori atypical group. A parallel sample of a priori typical children (*N*=136) with approximately the same demographic characteristics were reported to have no concerns for language development by their parents and/or teachers. Seventy children were excluded because they were reported to have autism, a neurological diagnosis, genetic disorders, intellectual disability (<60), severe Cerebral Palsy, a hearing loss, or blindness. Table 1 summarizes demographics and other characteristics of this sample.

Table 1 here

Procedures

All children received the DREAM test, administered via a tablet, with a standardized narration provided by a female Mandarin speaker who works in a professional capacity on a children's radio program. Each child heard pre-recorded questions while viewing pictures, and responded by touching the screen in the comprehension part and by giving a verbal response in the expressive part. When the child gave a verbal answer, the test administrator recorded responses by touching the corresponding word, picture, or phrase, choice buttons on the tablet screen. Test administration time averaged about 45 minutes to complete and took place in the child's school or preschool in a quiet place. Five examiners received a full-day training from a bilingual speech-language pathologist certified by the American Speech-Language Hearing Association, and a two-day practicum, until each examiner demonstrated competency in administering all tests used in this study. All children were tested by the trained examiners under the supervision of two speech-language pathologists certified by the American Speech-Language Hearing Association.

In the absence of an existing language test in China, elicited language samples were chosen as the accepted standard for convergent validity. A reasonable body of literature provided guidance as to what to expect at different ages in Mandarin language development, and the assessment research team made use of these resources (Cheng, 1988; Lee, 1982, 1986, 1992; Li et al., 2010; Lin, 1986; Liu, 2009; Miao, 1986; Zhou, 2002, 2004).

Convergent validity was investigated using language samples collected in the same test session. Given the wide age band, it was necessary to use different means of language sample collection for two broad age groups. For younger children aged 2;6 to 4;5, a spontaneous language sample was collected from a play session designed to elicit varied kinds of talk with the examiner (See Appendix B). For children aged 4;6 to 7;11, children received three wordless pictured narratives to describe (the Mandarin Expressive Narrative Test[4)], targeting aspects of grammar and semantics likely to be discriminating at this later age, to gain a broader perspective on language skills (see Appendix C for details). Very little relevant work has been published on Mandarin narratives in mainland China (Zhou & Zhang, 2010), though there are some small-scale studies in Taiwan (Chang, 2004). Once these narratives were recorded, a group of linguistically trained researchers listened to the samples and coded each utterance along a series of dimensions. In each case, specific criteria were used to score the child's language along a 0–3 point scale. There were 16 overall indices formed by the five scales for each of the three stories plus a composite measure of adequacy of answers to the questions following the stories. Cronbach's alpha was 0.82, suggesting that together these form a good scale. An overall narrative score was then derived by summing these together.

Several other measures were collected on the same children with the purpose of providing more information about their nonverbal intellectual capacities such as spatial reasoning (PTONI), executive function (Day-Night Stroop), and short-term auditory memory (Forward digit span). The details of these are provided in Appendix D.

Results

Sensitivity and Specificity

In order to estimate sensitivity and specificity, one needs a *true* or *gold-standard* classification status of *typical* or *atypical* for each child. Given the early state of speech-language diagnostics in China, no such true classification was available in the current study. Instead, a priori judgment status was used. Sensitivity represents the probability that a child who is judged to be atypically developing will receive a score below the DREAM test's at-risk cut-score. Specificity represents the probability that a child who is judged to be typically developing will receive a DREAM score above the at-risk cut score. Sensitivity and specificity values will vary depending on the cut-score selected.

To improve the usability of DREAM to identify at-risk students, a simple decision rule was sought. Instead of implementing a different cut-score for each DREAM scale, a rule was identified where a single cut-score was applied to all the scale scores simultaneously. For a single cut-score, an optimal balance of sensitivity and specificity was determined to occur when a child received a scale score of <80 on any one of the five DREAM test scales. This represents approximately 1.33 SDs below the mean, which is consistent with other at-risk cut-points in the literature[5].

The approach was intentionally designed not to rely on the Total Scale score alone, as this is not available if the child does not complete the entire test administration. Furthermore, unusual difficulty on one of the scales alone could warrant a classification of at-risk. Table 2 shows the findings for various cut-scores applied to the set of DREAM scale scores available to the practitioner.

Table 2 here

A second analysis tightened the criteria for atypical language development to improve the a priori classification. In addition to being referred for likely language problems by the pediatrician based solely on parent interview, the child also had to exhibit poor performance (*z* score <-1.25 SDs below the mean) on the language sample measures, whether that was the play session or the narrative. To count as typically developing, the child had to score above that level and also not have an a priori judgment of disorder.

One-hundred and eight children satisfied these dual criteria as typical or atypical, and a further statistical analysis was conducted to look at the sensitivity and specificity of their DREAM scores. As might be expected, the extra refinement of a priori classification improved sensitivity dramatically, to 95%. The DREAM test missed very few children who were classified as potentially language impaired and had poor language sample measures. Specificity did not improve, staying at 82%.[6]

In this study, 70% of the a priori atypical children were males and 30% females. However, there would be no reason to expect gender differences in the DREAM scores for children within the a priori atypical group. In fact, there were no statistically significant differences in DREAM total scaled scores attributable to gender within the a priori atypical group.

Evidence for External Validity

The study incorporated several additional measures discussed in Appendix D, to explore the construct validity of the DREAM scales. Correlations of DREAM scales with external measures are reported based on the different measures administered for two age ranges.

For ages 2;6– 4;5, correlations among the DREAM, PTONI, Digit Span, Executive Function, and Spontaneous Language are provided in Table 3. The following Spontaneous Language measures were computed: A Grammar score based on complexity of sentences, a Vocabulary score based on variety and types of words, and a Morpheme score based on the range of likely grammatical morphemes observed. These were combined into a Total Spontaneous Language score by averaging the *z*-scores of the three measures.

Table 3 here

For ages 4;6–7;11, correlations among the DREAM, PTONI, Digit Span, Executive Function, and Narrative are provided in Table 4. In this particular sample, the correlations among the DREAM scales are very high. The PTONI scores are in the moderate range, indicating discriminant validity between DREAM (a general language measure) and PTONI (a cognitive measure specific to visual-spatial reasoning).

Table 4 here

The intercorrelation matrix among the narrative indices was examined, this time taking the totals across the three stories for each index a–e (see Table 5). Results were mixed, with the most highly intercorrelated item being reference specification, and least effective index being descriptions of character's desires. However, different indices may contribute useful information at different points over this broad age span as found in other work on narrative. The individual scores were converted to *z*-scores by age band. Then a total narrative *z*-score was composed of the average of these component *z*-scores, to give them equal weight.

Table 5 here

Discriminant Relations with Spontaneous Language and Narratives: The DREAM Versus A Priori Judgment Status

This paper has discussed two approaches for classifying children as at-risk: a priori judgment of disorder without comprehensive language assessments, and any scale score <80. In the absence of a standard for atypical status, this section explores how well each approach relates to other skills such as Narrative production and Spontaneous language. Table 6 provides evidence that atypical classification defined by the DREAM cut-scores is more highly related to Narrative production and Spontaneous language measures than a classification based on a priori judgment. All of the DREAM atypical status correlations are significantly higher statistically (Lee & Preacher, 2013) than those for a priori atypical status (at p<.05 or lower). Note that being judged a priori typical or atypical is not a predictor of children's Narrative performance as the correlation is not statistically significant (p>.05).

Table 6 here

Discussion

The predictive validity of the new DREAM test was assessed in several ways. First, its sensitivity and specificity were evaluated against an a priori judgment status. In the absence of other comprehensive language assessments, it is unlikely that the children judged as atypically developing all meet the definition of pure language impairment. Despite that qualification, a respectable level of sensitivity and specificity was achieved if any DREAM component standard score was set at 80, or approximately the 9th percentile. This score is within the range of expected level of language impairments estimated to exist worldwide (Leonard, 2014; Tomblin et al., 1997). Sensitivity was much higher (0.95) if an extra criterion was added to a priori status, namely, whether the children also fell into the normal range or below -1.25 SDs on the indices of spontaneous language or narratives.

Specificity was still only moderate, however. This would be expected if the new test were to measure properties of language about which non-linguist professionals would be unaware, such as quantifier scope or verb complement structures. Neither would these properties necessarily be picked up in spontaneous language or even narratives, which are biased towards lexical items and structures that a child can use with confidence. As a result of both factors, a well-designed and demanding linguistic test is likely to pick out more children with subtle difficulties, categorizing fewer children as "typical". Therefore, the specificity against the a priori judgment is moderate. In a case such as this with no alternative gold standard, relatively lower specificity does not necessarily mean that the test is not doing its job.

Secondly, the validity was assessed by comparing the standardized test results to language samples, argued to be the best alternative in the absence of another gold standard language test. For the younger cohort, this was elicited in various ways in a play session that encouraged a variety of language forms and uses. The results showed excellent correlations with the DREAM subscores and total score. In addition, the a priori status was more weakly associated with the spontaneous language measures than the DREAM scores, reflecting the fact that a priori judgment is not yet as refined as instruments that have been carefully designed to measure linguistic content and avoid test bias. For the older cohort, spontaneous language was considered unlikely to reveal subtle language properties. For that reason, narratives were elicited using wordless picture stories that the children were encouraged to tell. DREAM scores correlated with narrative scores in a way that a priori judgment failed to do. The implication is that some children may have subtler problems that are manifest under careful testing but are concealed in ordinary communication with family and in school. Other children may be mistaken as having language problems in this age range. Though the narrative measure (MENT) added useful information, the index still has more variance than is desirable and needs refinement. It served its purpose here as providing further validity for DREAM in this older age group where there is too much unconstrained variability in ordinary spontaneous language.

Other measures that tap general cognitive abilities, such as visual pattern making (PTONI), executive function (Day-Night Stroop), and short term auditory memory (Digit Span) proved to be modestly related to the DREAM standard scores. For the younger group, these indices were less well correlated with DREAM than the spontaneous language measures were. However, for the older group, the narrative and cognitive measures both correlated with DREAM to about the same degree. The difference needs further exploration, as there were possibly floor effects for the younger children on the cognitive measures. We could interpret the pattern to suggest that language development is increasingly intertwined with other cognitive skills as children master the fundamentals and begin to use language for wider purposes. Though language has a considerable link to and dependence on general intelligence, it is not fully reducible to a general cognitive skill. For example, sentence repetition was not only related to Digit Span, but also to syntax comprehension. Repeating a sentence correctly requires grammatical knowledge, not just short-term memory for sounds. However, a child with weak auditory memory may show language difficulties as a result. Other work has shown that children with language impairment often have difficulties in executive function skills (Henry, Messer, & Nash, 2012; Sabbagh, Xu, Carlson, Moses, & Lee, 2006), but the direction of effect is not clear. As children acquire language they begin to use it as a tool for control of memory, rehearsal, and planning, but it is undoubtedly true that language learning itself requires auditory memory and controlled attention. The general implication is that each of these tasks provides useful information about a child's functioning, but the language test gives information of a specific sort relevant for language-based therapy, as it reveals the child’s state of linguistic competency.

Limitations

Developing a standardized, norm-referenced assessment for mainland China does not end with the demonstration that the test meets appropriate standards, and at the present time the sample size is still small relative to the population of China. It will be necessary to expand the range of children who take this and other assessments to investigate whether such tests can play a satisfactory role in all of the diverse circumstances that affect children in need of language intervention.

In addition, the issue of bilingualism and bi-dialectalism needs to be directly addressed in future work. The current norming took careful account of regional and dialectal influences in the areas studied, and found relatively little to adjust. Nevertheless, the concern is that, especially below age 3, there may be children who are just beginning to be exposed to Mandarin. It is therefore unwise to compare their skills to those of children who have been native Mandarin speakers from the first year. Even in this age range the spontaneous language was remarkably confirmatory of the child's level of attainment, but both may underestimate the language skills of the youngest bilingual children. Interesting work is underway in the United States (Iglesias, 2015; Peña, Gutierrez-Clellen, Iglesias, Goldstein & Bedore, 2014) and Europe (Armon-Lotem, 2012) to derive the best practice for evaluating bilingual children for speech and language disorders, and more work is needed in China on this front.

On a final note, the assessment of children for speech and language disorders cannot happen in a vacuum: the educational and pediatric care must be prepared for the consequences of such identification by training therapists, designing and testing appropriate interventions and re-evaluations (Rogers et al., 2012). China's progress has been rapid in this regard, and it will be vitally important to match the preparation of the therapists to the sophistication of the instruments, particularly with respect to knowledge about language acquisition and linguistics.

References

Armon-Lotem, S. (2012). Introduction: Bilingual children with SLI–the nature of the problem. *Bilingualism: Language and Cognition*, *15*(01), 1–4. http://dx.doi.org/10.1017/S1366728911000599

Bedore, L. M., Peña, E. D., Gillam, R. B., & Ho, T. (2010). Language sample measures and language ability in Spanish-English bilingual kindergarteners. *Journal of Communication Disorders*, *43*(6), 498–510. http://dx.doi.org/10.1016/j.jcomdis.2010.05.002

Chang, C. J. (2004). Telling stories of experiences: Narrative development of young Chinese children. *Applied Psycholinguistics*, *25*(01), 83–104. http://dx.doi.org/10.1017/S0142716404001055

Cheng, S.-W. (1988). Beginning negative sentences among Mandarin-speaking toddlers. *Chinese Journal of Psychology, 30*, 47–63.

Cheung, H. (2009). Grammatical characteristics of Mandarin-speaking children with specific language impairment. *Language Disorders in Speakers of Chinese*, 33–52.

Chien, Y. C., Lust, B., & Chiang, C. P. (2003). Chinese children's comprehension of count-classifiers and mass-classifiers. *Journal of East Asian Linguistics*, *12*(2), 91–120. http://dx.doi.org/10.1023/A:1022401006521

China Data Center (2012). *China 2010 County Population Census Data*. Ann Arbor, MI: University of Michigan.

Crain, S., & Thornton, R. (1998). Investigations in universal grammar: a guide to research on the acquisition of syntax and semantics. Cambridge, MA: MIT Press.

de Villiers, J., & Johnson, V. E. (2007). Implications of new vocabulary assessments for minority children. *Vocabulary acquisition: Implications for reading comprehension*, 157-181.

de Villiers, J., & Roeper, T. (Eds.). (2011). *Handbook of generative approaches to language acquisition* (Vol. 41). New York, NY: Springer Science & Business Media. http://dx.doi.org/10.1007/978-94-007-1688-9

de Villiers, J., Roeper, T., Bland-Stewart, L., & Pearson, B. (2008). Answering hard questions: Wh-movement across dialects and disorder. *Applied Psycholinguistics*, *29*(01), 67–103. http://dx.doi.org/10.1017/S0142716408080041

de Villiers, P. A., & de Villiers, J. G. (2010). Assessment of language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(2), 230–44. http://dx.doi.org/10.1002/wcs.30

Duchesne, L. (2015). Grammatical competence after early 8 cochlear implantation. *The Oxford Handbook of Deaf Studies in Language* (pp. 113–31). New York, NY: Oxford University Press.

Dunn, D. M., & Dunn, L. M. (2007). *Peabody picture vocabulary test: Manual*. San Antonio, TX: Pearson.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Newark, NJ: Erlbaum.

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur-Bates Communicative Development Inventories: User's guide and technical manual* (2nd ed.). Baltimore, MD: Brookes.

Friberg, J. C. (2010). Considerations for test selection: How do validity and reliability impact diagnostic decisions? *Child Language Teaching and Therapy*, *26*(1), 77–92. http://dx.doi.org/10.1177/0265659009349972

Friedmann, N., & Novogrodsky, R. (2004). The acquisition of relative clause comprehension in Hebrew: A study of SLI and normal development. *Journal of Child Language*, *31*(03), 661–81. http://dx.doi.org/10.1017/S0305000904006269

Fung, R. S. Y. (2009). Characteristics of Chinese in relation to language disorders. *Language Disorders in Speakers of Chinese* (pp. 1–18). Bristol, UK: Multilingual Matters.

Guasti, M. T. (2004). *Language acquisition: The growth of grammar.* Cambridge, MA: MIT Press.

Henry, L. A., Messer, D. J., & Nash, G. (2012). Executive functioning in children with specific language impairment. *Journal of Child Psychology and Psychiatry*, *53*(1), 37–45. http://dx.doi.org/10.1111/j.1469-7610.2011.02430.x

Hirsh-Pasek, K., Kochanoff, A., Newcombe, N., & de Villiers, J. G. (2005). Using scientific knowledge to inform preschool assessment: Making the case for empirical validity. *Social Policy Report (SRCD), 19*(1), 3–19.

Hoff, E. (2014). *Language development*. San Francisco, CA: Cengage Learning.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. *Test Validity* (pp. 129–45). Abingdon, UK: Routledge.

Hresko, W. P., Reid, D. K., & Hammill, D. D. (1999). *TELD-3: Test of Early Language Development: Examiner's Manual*. Austin, TX: Pro-Ed.

Iglesias, A. (2015). Language impairment in bilingual children: From theory to practice. *Seminars in Speech and Language*, *36*(2), 87. http://dx.doi.org/10.1055/s-0035-1549103

Lee, J. N., & Naigles, L. R. (2008). Mandarin learners use syntactic bootstrapping in verb acquisition. *Cognition*, *106*(2), 1028–37. http://dx.doi.org/10.1016/j.cognition.2007.04.004

Lee, I. A., & Preacher, K. J. (2013, September). Calculation for the test of the difference between two dependent correlations with one variable in common [Computer software]. Available from http://quantpsy.org.

Lee, T. H. T. (1982). The development of negation in a Mandarin-speaking child. *Language Learning and Communication, 1*(3), 269–81.

Lee, T. H. T. (1986). Acquisition of quantificational scope in Mandarin Chinese. Paper presented at the Annual Stanford Child Language Forum (18th), Stanford, CA, April 4–6, 1986.

Lee, T. H. T. (1992). The inadequacy of processing heuristics: Evidence from relative clause acquisition in Mandarin Chinese. *Research in Chinese Linguistics in Hong Kong* (pp. 47–85. Hong Kong, CN: Linguistic Society of Hong Kong.

Leonard, L. B. (2014). *Children with specific language impairment*. Cambridge, MA: MIT Press.

Li, P., Huang, B., & Hsiao, Y. (2010). Learning that classifiers count: Mandarin-speaking children's acquisition of sortal and mensural classifiers. *Journal of East Asian Linguistics*, *19*(3), 207–30. http://dx.doi.org/10.1007/s10831-010-9060-1

Lin, O. H. J. (1986). *A developmental study of the acquisition of aspect markers in Chinese children.* (Unpublished master's thesis). Fu Jen Catholic University, Taipei, TW.

Linacre, J. M. (2014). *Reliability and separation measures: Winsteps Help*. Retrieved from: http://www.winsteps.com/winman/reliability.htm

Liu, H. (2009). The acquisition of Mandarin aspects and modals: Evidence from the acquisition of negation. *Language and Linguistics*, *10*(1), 133–60.

Liu, H. & Ning, C. (2009). *Phase impenetrability condition and the acquisition of unaccusatives, object-raising ba-constructions and passives in Mandarin-speaking children.* Proceedings of the Third Conference on Generative Approaches to Language Acquisition North America (GALANA 2008). Somerville, MA: Cascadilla Press.

MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of Child Language*, *12*(2), 271–95. http://dx.doi.org/10.1017/S0305000900006449

McDaniel, D., McKee, C., & Cairns, H. S. (1996). *Methods for assessing children's syntax*. Cambridge, MA: MIT Press.

Miao, X. (1986). Young children's understanding of interrogatives: The developmental peculiarities of answering wh-questions in young children. *Psychological Science*, *3, 1–5.*

Ning, C.Y., Liu, X.L., & de Villiers, J.G. (2014). *The Diagnostic Receptive and Expressive Assessment of Mandarin.* Dallas, TX: Bethel Hearing and Speaking Training Center.

Norbury, C. F., Gooch, D., Wray, C., Baird, G., Charman, T., Simonoff, E., Vamvakas, G. and Pickles, A. (2016). The impact of nonverbal ability on prevalence and clinical presentation of language disorder: evidence from a population study. Journal of Child Psychology and Psychiatry. doi:10.1111/jcpp.12573

Pearson, B. Z., Jackson, J. E., & Wu, H. (2014). Seeking a valid gold standard for an innovative, dialect-neutral language test. *Journal of Speech and Hearing Research*, 39, 1239–57. http://dx.doi.org/10.1044/2013_jslhr-l-12-0126

Peña, E. D. (2007). Lost in translation: Methodological considerations in cross-cultural research. *Child Development*, *78*(4), 1255–64. http://dx.doi.org/10.1111/j.1467-8624.2007.01064.x

Peña, E. D., Gutierrez-Clellen, V., Iglesias, A., Goldstein, B., & Bedore, L. M. (2014). *BESA: Bilingual English-Spanish Assessment Manual*. San Diego, CA: AR-Clinical Publications.

Reilly, S., Tomblin, B., Law, J., McKean, C., Mensah, F. K., Morgan, A., ... & Wake, M. (2014). Specific language impairment: A convenient label for whom? *International Journal of Language & Communication Disorders*, *49*(4), 416–51. http://dx.doi.org/10.1111/1460-6984.12102

Rice, M. L. (2013). Language growth and genetics of specific language impairment. *International Journal of Speech-Language Pathology*, *15*(3), 223–33. http://dx.doi.org/10.3109/17549507.2013.783113

Rice, M. L., Buhr, J. C., & Nemeth, M. (1990). Fast mapping word-learning abilities of language-delayed preschoolers. *Journal of Speech and Hearing Disorders*, *55*(1), 33–42. http://dx.doi.org/10.1044/jshd.5501.33

Rice, M. L., & Wexler, K. (1996). Toward tense as a clinical marker of specific language impairment in English-speaking children. *Journal of Speech, Language, and Hearing Research*, *39*(6), 1239–57. http://dx.doi.org/10.1044/jshr.3906.1239

Rogers, M., Bentler, R., Liu, B., Liu, X., Xu, L., Louko, L., & Hallowell, B. (2012). *Strategic building for the CSD professions in China*. CAPCSD 2012 Annual Conference: Sustaining Global Alliances. Lecture conducted from Council of Academic Programs in Communication Sciences and Disorders, Newport Beach, CA.

Sabbagh, M. A., Xu, F., Carlson, S. M., Moses, L. J., & Lee, K. (2006). The development of executive functioning and theory of mind a comparison of Chinese and U.S.

preschoolers. *Psychological Science*, *17*(1), 74–81. http://dx.doi.org/10.1111/j.1467-9280.2005.01667.x

Semel, E. M., Wiig, E. H., & Secord, W. (1996). *CELF 3, Clinical Evaluation of Language Fundamentals: Observational rating scales.* San Antonio, TX: The Psychological Corporation.

Seymour, H., Roeper, T., & de Villiers, J. (2005) *The DELV-NR. (norm-referenced version): The Diagnostic Evaluation of Language Variation*. San Antonio, TX: The Psychological Corporation.

Slobin, D. I. (Ed.). (1986). *The crosslinguistic study of language acquisition: Theoretical issues*. New York, NY: Psychology Press.

Snowling, M. J., Duff, F. J., Nash, H. M., & Hulme, C. (2015). Language profiles and literacy outcomes of children with resolving, emerging, or persisting language impairments. *Journal of Child Psychology and Psychiatry*. http://dx.doi.org/10.1111/jcpp.12497

Sun, C. (2006). *Chinese: A linguistic introduction*. Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511755019

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

Tomblin, J. B., Records, N. L., Buckwalter, P., Zhang, X., Smith, E., & O’Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research*, *40*(6), 1245–60. http://dx.doi.org/10.1044/jslhr.4006.1245

Zambrana, I. M., Pons, F., Eadie, P., & Ystrom, E. (2013). Trajectories of language delay from age 3 to 5: Persistence, recovery and late onset. *International Journal of Language & Communication Disorders, 49*(3), 304–-16. http://dx.doi.org/10.1111/1460-6984.12073

Zhang, Y., Jin, X., & Shen, X. (2003). The establishment of screening standard for the language development of 2-3 year old children in Mainland China. *Zhong Guo Er Tong Bao Jian Za Zhi*, *11*(5), 308–11.

Zhou, F. & Zhang, Y. (2010). Research on narrative ability in microstructure and macrostructure of preschool children. *Chinese Journal of Child Health Care*, *18*(1), 18–21.

Zhou, G. (2002). An investigation of the children's use of the negative 不 and the structure of negation [J]. *Applied Linguistics*, *4*, 6.

Zhou, G. (2004). An investigation into children's acquisition of Chinese time system. *Applied Linguistics*, *4*, 33–40.

Zhou, P. & Crain, S. (2009). Scope assignment in child language: Evidence from the acquisition of Chinese. *Lingua*, *119*(7), 973–88. http://dx.doi.org/10.1016/j.lingua.2009.01.001

Zhou, P., & Crain, S. (2011). Children's knowledge of the quantifier *dou* in Mandarin Chinese. *Journal of Psycholinguistic Research*, *40*(3), 155–76. http://dx.doi.org/10.1007/s10936-010-9161-z

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2002). *Preschool Language Scale, (PLS-4).* San Antonio, TX: The Psychological Corporation.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language assessment quarterly*, *4*(2), 223–33. http://dx.doi.org/10.1080/15434300701375832

Author Note

Authors acknowledge the financial support of Bethel Hearing and Speech Training Center for the standardization and the current study. We would like to thank the Linguistics department graduate students and faculty, including the Dean, Dr. Gang Gu, at Tianjin Normal University and the families and schools who allowed us to assess their children throughout China. Authors Jiang and Zhang were supported with funding (Chinese Ministry of Health Professional Research Fund #201002006) to assist the validity work in the study through Shanghai Children's Medical Center.

Footnotes

[1] DREAM is a proprietary assessment tool of Bethel Hearing and Speech Training Center Inc.

[2] Under DSM-5, relaxing the criteria for non-verbal IQ results in new estimates of around 11% (Norbury et al, 2016).

[3] Of 1,183 children who participated not all were included in the norming process due to incomplete administrations or because students did not fit the census stratification.

[4] Mandarin Expressive Narrative Test (MENT) © 2014 by Bethel Hearing and Speaking Training Center Inc.

[5] DSM-5 criteria: Language scores −1.5*SD* or more below normative mean on 2/5 language composite scores. Tomblin et al. (1997): Language scores −1.25*SD* or more below normative mean on 2/5 language composite scores.

[6] Using the language sample measures alone resulted in unacceptable classification accuracy, as no decision rule resulted in classification accuracy over 50%.

[7] The Scoring Sheet for Spontaneous Language ©2014 by Bethel Hearing and Speaking Training Center Inc.

Table 1. Demographic characteristics of the validation sample.

Group	*A priori atypical*		*A priori typical*	
N-Count	94		136	
Region				
North (Region I & Region II Combined)	0		0	
South (Region I & Region II Combined)	100		100	
Gender (%M)	77.1		65.1	
Age-groups				
2;6–2;11	14		21	
3;0–3;5	16		18	
3;6–3;11	12		17	
4;0–4;5	8		18	
4;6–4;11	8		16	
5;0–5;5	10		11	
5;6–5;11	6		9	
6;0–6;11	10		15	
7;0–7;11	10		11	
Parent education level (%)				
Low	21		13	
Medium	30		21	
High	49		66	
DREAM scales	(n=69)		(n=102)	
Total (Mean/SD)	79.7 (16.0)		97.3 (16.0)	
Receptive (Mean/SD)	79.7 (14.6)		97.8 (16.5)	
Expressive (Mean/SD)	78.7 (12.7)		95.2 (16.5)	
Syntax (Mean/SD)	78.0 (10.1)		91.5 (13.5)	
Semantic (Mean/SD)	81.1 (18.6)		103.2 (19.8)	
P-TONI (ages 2;6 – 7;11)	95.1 (22.8)	[n=69]	111.4 (19.4)	[n=43]
Digit Span (ages 2;6 – 7;11)	5.0 (2.0)	[n=69]	5.8 (1.6)	[n=120]
Stroop (ages 2;6-7;11)	12.4 (4.3)	[n=59]	13.6 (3.9)	[n=114]
Spontaneous Speech Total[2] (ages 2;6-4;5)	-0.44 (0.9)	[n=39]	0.28 (0.9)	[n=53]
Ment Total[2] (ages 4;5-7;11)	-0.17 (0.7)	[n=25]	-0.01 (0.6)	[n=47]

[1] A greater proportion of males was sampled in the non-referred group to match the preponderance of males in the clinically referred group.

[2] The *Spontaneous Speech* and *Ment* measures are reported as z-scores.
Note. Primary caregiver level *Low* represents less than a high-school diploma, *Medium* includes high-school diploma and post-secondary Associates degrees or equivalent, *High* represents a Bachelor’s Degree or above.

Table 2. Comparison of cut-score criteria evaluated for atypical determination.

Criteria	*Criterion*	*Sensitivity*	*Specificity*	*Correctly Classified*	*LR+*	*LR-*
Any DREAM < 70	A priori Status	41.30%	85.71%	67.56%	2.8913	0.6848
Any DREAM < 75	A priori Status	55.43%	84.21%	72.44%	3.5109	0.5292
Any DREAM < 80	A priori Status	73.31%	81.55%	79.02%	3.6235	0.3247
Any DREAM < 85	A priori Status	82.61%	68.42%	74.22%	2.6159	0.2542
Any DREAM < 90	A priori Status	86.96%	57.14%	69.33%	2.029	0.2283

Note: The criterion in the middle row was selected as exhibiting the highest classification accuracy and balanced sensitivity and specificity.

Table 3. Correlations among DREAM scales and validity measures administered to ages 2;6–4;5.

		DREAM								*Spont. Language*		
		Total	Recept -ive	Express -ive	Syntax	Semantic	PTONI[a]	Digit Span	Executive	Grammar	Vocabulary	Morpheme
DREAM	Total	1										
	Receptive	.995**	1									
	Expressive	.858**	.806**	1								
	Syntax	.948**	.929**	.899**	1							
	Semantic	.981**	.986**	.788**	.868**	1						
	PTONI[a]	.567**	.557**	.577**	.625**	.514**	1					
	Digit Span	.463**	.454**	.481**	.469**	.450**	.457**	1				
	Executive	.434**	.432**	.425**	.437**	.423**	.177	.436**	1			
Spont.Lang.	Grammar	.635**	.626**	.523**	.551**	.644**	.196	.322**	.200	1		
	Vocabulary	.615**	.617**	.515**	.542**	.634**	.220	.503**	.264*	.830**	1	
	Morpheme	.712**	.705**	.621**	.667**	.702**	.337**	.405**	.261*	.846**	.799**	1

**Correlation is significant at the 0.01 level (2-tailed).

*Correlation is significant at the 0.05 level (2-tailed).

[a] PTONI was administered only to ages 3;0 and above.

Table 4. Correlations among DREAM scales and validity measures administered to ages 4;6–7;11.

DREAM

		Total	Receptive	Expressive	Syntax	Semantic	PTONI	Digit Span	Executive	Narrative
DREAM	Total	1								
	Receptive	.991**	1							
	Expressive	.846**	.771**	1						
	Syntax	.935**	.907**	.877**	1					
	Semantic	.969**	.975**	.761**	.820**	1				
	PTONI	.516**	.494**	.499**	.499**	.489**	1			
	Digit Span	.567**	.519**	.638**	.582**	.509**	.307**	1		
	Executive	.485**	.470**	.486**	.488**	.454**	.253*	.378**	1	
	Narrative	.489**	.489**	.399**	.485**	.457**	.284*	.188	.169	1

**Correlation is significant at the 0.01 level (2-tailed).

*Correlation is significant at the 0.05 level (2-tailed).

[a] PTONI was administered only to ages 3;0 and above.

Table 5. Correlations among Narrative indices.

	Desire	*Emotion*	*Mental*	*Reference Specification*	*Time Specification*	*Questions after stories*
Desire	1					
Emotion	0.16	1				
Mental	0.15	0.95**	1			
Reference Specification	0.42**	0.23*	0.33**	1		
Time Specification	0.19	0.16	0.26*	0.47**	1	
Questions after stories	0.27*	0.28*	0.37**	0.52**	0.28*	1

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 6. Differential correlations between a priori atypical status, the DREAM atypical status, and Narrative production and Spontaneous language

Ages (N)	*Construct*	*Validity Score*	*A priori atypical Status*	*DREAM atypical Status*	*Statistical significance of the difference between a priori and Dream correlations—two-tailed.*[a]
2:6–4:5 (*N*=92)	Spontaneous Language	Total	.369**	.571**	*p* =.004
		Grammar	.306**	.516**	*p* =.004
		Vocabulary	.384**	.545**	*p* = .011
		Morphemes	.350**	.548**	*p* = .005
4:6–7:11 (*N*=70)	Narrative	Total	.148	.421**	*p* = .029

[a] Estimated using Lee & Preacher (2013).
** Correlation is significantly different from 0 at the 0.01 level (2-tailed).

Appendix A: Structure of DREAM

DREAM-C is organized into seven subscales that are used to estimate five composite scores. The mapping of subscales to composites is provided in the table below.

The vocabulary used in the grammar test items was chosen to be relatively easy, and a major sample of that is assessed in the vocabulary portion of the instrument. However, the child's ability to learn new words is also tapped via fast mapping, to reflect the process of learning. Closed class items specific to Mandarin are tested, including those that mark aspect and passive voice in the verb phrase, and no test would be complete without taking note of the special classifier system of Mandarin. The logical aspects of language, such as quantifiers and connectives, are examined. The sentence structures common to most languages but taking particular form in Mandarin were selected for the test, both in production and comprehension. Sentence repetition (which is sensitive to grammar as well as memory), and the production of sentences and phrases in the description of events and scenes were chosen for expressive portions of the test. By choosing items that ranged in difficulty in these major areas of knowledge, the test contains material appropriate for children from early grammar to language more relevant at age eight.

Appendix A Table. Relationship of the DREAM subscales to the DREAM Indices.

Category	**Receptive (N=165)**					**Expressive (N=67)**	
Sub-scale	**Vocabulary**	**Fast-Mapping**	**Closed Class**	**Logical expressions**	**Sentence contrasts**	**Event Description**	**Sentence Repetition**
Indices							
Receptive	X	X	X	X	X		
Expressive			X			X	X
Syntax					X	X	X
Semantic	X	X	X	X			
Total Language	X	X	X	X	X	X	X
Example	Point to where the cat is.	Which one is *dafu*"?	Which picture shows “an apple is in the bowl”?	Which picture has a lot of apples?	*Whose dog is eating the corn?*	Elicit description of pictured event.	Repeat a sentence.

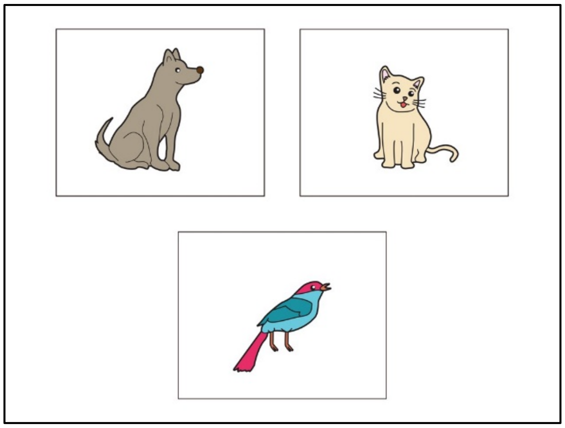
W

Examples from Appendix A Table:

A1. 指一指小猫在哪里。

zhi3 yi4 zhi3 xiao3 mao1 zai4 na3 li3

Point to where the cat is.



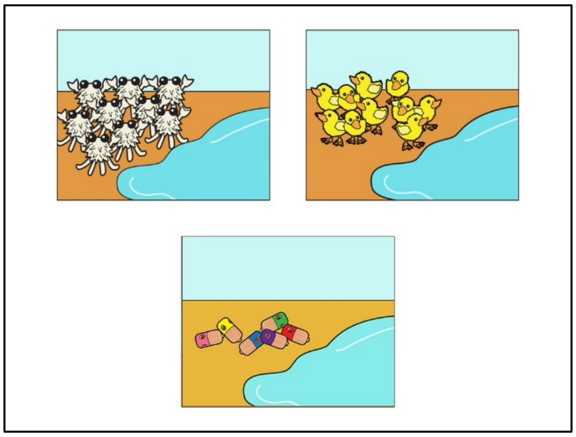
An example of a vocabulary item (from Ning, Liu, & de Villiers, 2014). Diagnostic Receptive and Expressive Assessment of Mandarin (DREAM). Copyright © 2014 Bethel Hearing and Speaking Training Center Inc. Reproduced with permission. All rights reserved.

A.2. The figure below provides an illustration of an item (1) in comprehension. The item involves a process-type problem, namely fast mapping a novel word via cues from the sentence context, in this case, from the classifier that precedes it.

(1) 看，河边有一群***dafu***。**哪张图里的是*****dafu*****？**

kan4, he2 bian1 you3 yi4 qun2 da1fu1. na3 zhang1 tu2 li3 de shi4 da1fu1?

"There is a flock of *dafu* by the river side. Which picture shows *dafu*?"



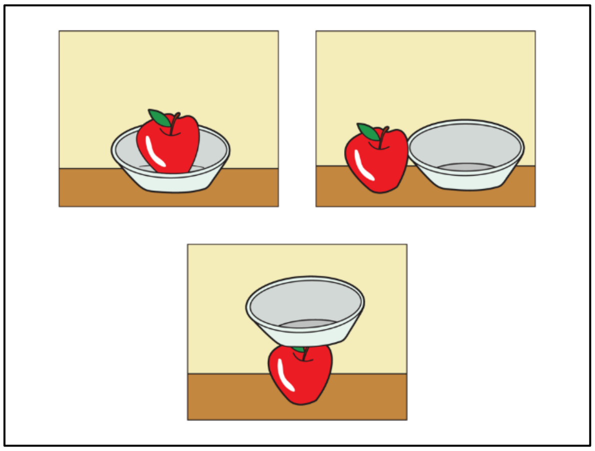
An example of a fast mapping item (from Ning, Liu, & de Villiers, 2014). Diagnostic Receptive and Expressive Assessment of Mandarin (DREAM). Copyright © 2014 Bethel Hearing and Speaking Training Center Inc. Reproduced with permission. All rights reserved.

The classifier qun2 (群）in Mandarin is used with animates; the child likely already has a word for the other animates, which therefore cannot be *dafu*. The unknown *inanimate* could not be dafu because then the classifier would be inappropriate. Hence the only option is to choose the small crab-like entities as the reference for the novel word *dafu*. Classifiers in Mandarin play a much more important role grammatically than English words such as "flock", which is used as the nearest equivalent to illustrate the point.

A3. 哪张图画的是苹果在碗里面？

na3 zhang1 tu2 hua4 de shi4 ping2 guo3 zai4 wan3 li3 mian4?

Which picture shows "an apple is in the bowl"?

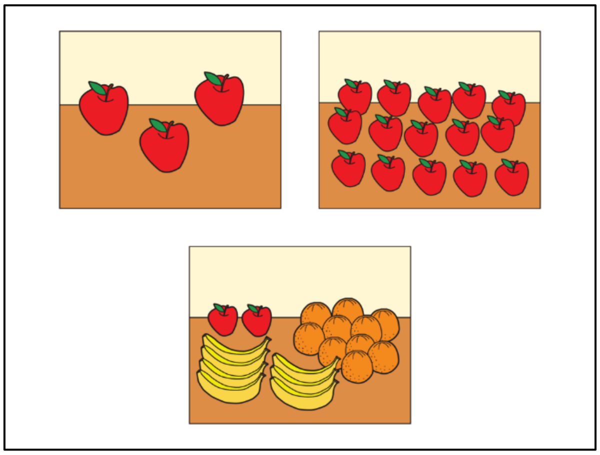


An example of a closed class item (from Ning, Liu, & de Villiers, 2014). Diagnostic Receptive and Expressive Assessment of Mandarin (DREAM). Copyright © 2014 Bethel Hearing and Speaking Training Center Inc. Reproduced with permission. All rights reserved.

A4. 哪张图里有很多苹果？

na3 zhang1 tu2 li3 you3 hen3 duo1 ping2 guo3?

Which picture has a lot of apples?



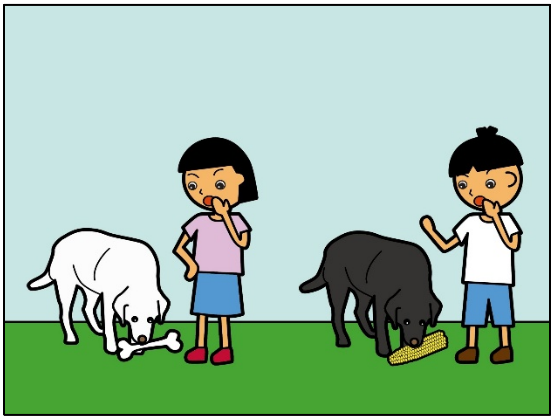
An example of a logical expressions item (from Ning, Liu, & de Villiers, 2014). Diagnostic Receptive and Expressive Assessment of Mandarin (DREAM). Copyright © 2014 Bethel Hearing and Speaking Training Center Inc. Reproduced with permission. All rights reserved.

A5. 看，男孩有一只小狗。他的小狗在吃玉米。女孩也有一只小狗，她的小狗在吃骨头。谁的小狗在吃玉米?

Kan4, nan2 hai2 you3 yi4 zhi1 xiao3 gou3. Ta1 de xiao3 gou3 zai4 chi1 yu4 mi3. Nǚ3hai2 ye3 you3 yi1 zhi1 xiao3 gou3, ta1 de xiao3 gou3 zai4 chi1 gu3 tou. Shui2 de xiao3 gou3 zai4 chi1 yu4 mi3?

Elicitation: “Look, this boy has a dog. His dog is eating corn. This girl has a dog. Her dog is eating a bone.”

Question: “Whose dog is eating the corn?”



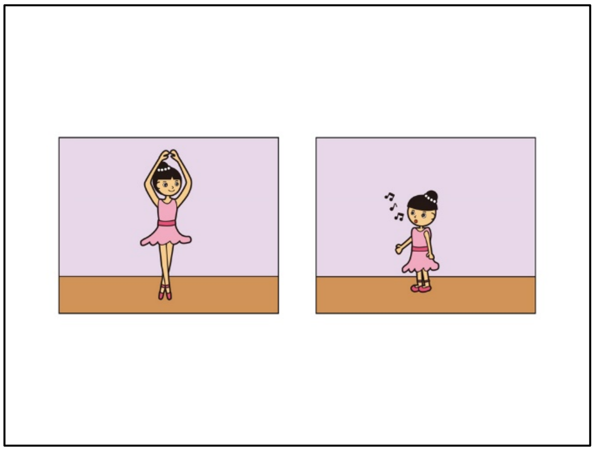
An example of a sentence contrasts item (from Ning, Liu, & de Villiers, 2014). Diagnostic Receptive and Expressive Assessment of Mandarin (DREAM). Copyright © 2014 Bethel Hearing and Speaking Training Center Inc. Reproduced with permission. All rights reserved.

A6. 小朋友请看图，这个女孩高，她在跳舞；那个女孩矮，她在唱歌。哪个女高?

xiao3 peng2 you3 qing3 kan4 tu2, zhe4 ge4 nü3 hai2 gao1, ta1 zai4 tiao4 wu3; na4 ge4 nü3 hai2 ai3, ta1 zai4 chang4 ge1. na3 ge4 nü3 hai2 gao1?

Elicitation: “Look at the picture: There is a girl. She is tall and she is dancing. There is another girl. She is short and she is singing.”

Question: “Which girl is tall?”



An example of an event description item (from Ning, Liu, & de Villiers, 2014). Diagnostic Receptive and Expressive Assessment of Mandarin (DREAM). Copyright © 2014 Bethel Hearing and Speaking Training Center Inc. Reproduced with permission. All rights reserved.

The above figure is from a subtest in the expressive part of the test, where a child is encouraged to use a certain type of sentence structure by hearing it modeled in a related context. The relative clause in Mandarin is marked by the particle DE (Cheng, 1995; Wang, 2009) and this is a task that requires the child to use a relative clause to distinguish between two referents.

A7. 妈妈歇会吧

Ma1 ma1 xie1 hui4 er ba1.

Mother, have a rest.

An example of a sentence repetition item (from Ning, Liu, & de Villiers, 2014). Diagnostic Receptive and Expressive Assessment of Mandarin (DREAM). Copyright © 2014 Bethel Hearing and Speaking Training Center Inc. Reproduced with permission. All rights reserved.

References

Cheng, S. Y.-Y. (1995) *The Acquisition of Relative Clauses in Chinese*. Unpublished MA thesis, National Taiwan Normal University.

Ning, C.Y., Liu, X.L., & de Villiers, J.G. (2014). *The Diagnostic Receptive and Expressive Assessment of Mandarin.* Dallas, TX: Bethel Hearing and Speaking Training Center.

Wang, Zhe. (2009). *Investigation and research on acquisition of relative clauses of Mandarin speaking children. (*Unpublished master's thesis). Changsha, CN: University of Hunan.

Appendix B: Language Samples in Spontaneous Play

For the spontaneous language sample, the researchers used a variety of toys and pictures with the child to elicit language, including descriptions, and not just naming. For instance, the child was shown pictures of illogical or unusual situations, such as a boy riding a tricycle with a square-shaped front wheel and was asked whether the boy could pedal forward and why or why not. The session was also arranged so that certain things went wrong, for example, the activity of coloring a picture would be thwarted by lack of the crayon that the child was instructed to use. This provided appropriate functional opportunities for requests, questions, or negations. This strategy has worked well in other tests for young children (Pearson, Jackson, & Wu, 2014; Peña, Gutierrez-Clellen, Iglesias, Goldstein, & Bedore, 2014).

Once the 15 minute language samples were recorded, they were played back to a group of linguistically skilled researchers who listened for certain specific properties in the language sample, covering word use, grammatical complexity, and morphology. The diversity of vocabulary was assessed on a five-point scale derived in part from previous work on developmental milestones (Hao, Shu, Xing, & Li, 2008; Hoff, 2014). The grammatical complexity was also assessed on a five-point scale, derived by consideration of the complexity of the clause types used, much as in Scarborough's (1990) work on the IPSYN in English. Based on previous language acquisition studies of the emergence of grammatical morphemes of aspect and classifiers (Zhou, 2004; Zong, 2011), the morphemes heard in the transcript were checked off and the variety of morphemes used was then totaled. Each checklist was designed to represent simpler or earlier forms and then increasingly complex forms, based on empirically based knowledge about Mandarin use by young children. An overall spontaneous language score[7] was derived by adding together the points from these different aspects.

References

Hao, M., Shu, H., Xing, A., & Li, P. (2008). Early vocabulary inventory for Mandarin Chinese. *Behavior Research Methods*, *40*(3), 728–33. http://dx.doi.org/10.3758/BRM.40.3.728

Hoff, E. (2014). *Language development*. San Francisco, CA: Cengage Learning.

Pearson, B. Z., Jackson, J. E., & Wu, H. (2014). Seeking a valid gold standard for an innovative, dialect-neutral language test. *Journal of Speech and Hearing Research*, 39, 1239–57. http://dx.doi.org/10.1044/2013_jslhr-l-12-0126

Pena, E. D., Gutierrez-Clellen, V., Iglesias, A., Goldstein, B., & Bedore, L. M. (2014). *BESA: Bilingual English-Spanish Assessment Manual*. San Diego, CA: AR-Clinical Publications.

Scarborough, H. S. (1990). Index of productive syntax. *Applied Psycholinguistics*, *11*(1), 1–22.http://dx.doi.org/10.1017/S0142716400008262

Zhou, G. (2004). An investigation into children’s acquisition of Chinese time system. *Applied Linguistics*, *4*, 33–40.

Zong, L. (2011). A Standardized Test of Syntactic Comprehension Ability for Mandarin-Speaking Preschool Children. Masters thesis in Linguistics, Tianjin University.

Appendix C: Narratives

Narratives were recorded and analyzed for the presence of properties typically developed between the ages of 4 to 8 years, based on cross-linguistic work. The measures focus on markers of temporal cohesion, reference specificity to distinguish characters for the listener, and the landscape of consciousness or mental state references about the characters. As in the Dialectal Evaluation of Language Variation test (Seymour et al., 2005) children saw three short picture stories, seeing one picture at a time, then they were asked to start at the beginning of the sequence and describe what happened. One sequence represented a classic theory-of-mind scenario (Leslie, 1987; Wellman, Cross, & Watson, 2001), in which a character sees something placed in one location, leaves the scene, and then another character moves the object to a new location, out of sight. The first character returns, with a thought balloon to indicate that the character now wants the object. The second and third stories added more and different elements, including unintended mistakes and small dramas of deception. These were also designed to allow children to describe events either at a purely action level, or at the level of character's desires, motives, beliefs, and emotions. The transition from one form of storytelling to the other is a major development in this age range 4;6–7;11 (Seymour et al, 2005; Burns, de Villiers, Pearson, & Champion, 2012). After each story, the researcher asked the child some questions designed to promote such causal explanations, such as, "Why is he looking there?" or "Why couldn't the boy climb down the tree?" to further elicit complex elements of desires and emotions etc.

The dimensions coded included higher levels of grammatical complexity appropriate for this age range, to include mental verbs plus complements, or complex use of sentence connectives such as "while", "before", etc. A level of overall quality was also marked, having to do with the adequacy of a single picture description. In addition, indices of how well the child specified referents and whether they had sophisticated time reference were scored (Burns et. al, 2012). The sophistication of references to emotion, desire, and mental states was also coded, as an index of their ability to employ Theory of Mind to describe the characters. Finally, one or two questions requiring a causal explanation were asked after each story.

References

Burns, F., de Villiers, P. A., Pearson, B., & Champion, T. (2012). Dialect neutral indices of narrative cohesion and evaluation. *Language, Speech and Hearing Services in Schools,43,* 132–52. http://dx.doi.org/10.1044/0161-1461(2011/10-0101)

Leslie, A. M. (1987). Pretense and representation: The origins of "theory of mind." *Psychological Review*, *94*(4), 412. http://dx.doi.org/10.1037/0033-295X.94.4.412 Scarborough, H. S. (1990). Index of productive syntax. *Applied Psycholinguistics*, *11*(1), 1–22. http://dx.doi.org/10.1017/S0142716400008262

Seymour, H., Roeper, T., & de Villiers, J. (2005) *The DELV-NR. (norm-referenced version): The Diagnostic Evaluation of Language Variation*. San Antonio, TX: The Psychological Corporation.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, 73 (2) 655–684. http://dx.doi.org/10.1111/1467-8624.00304

Appendix D: Supplementary Tests

The tests needed to be culturally neutral and to provide a snapshot of the child's intellectual abilities of a different kind than linguistic testing reveals. The tests chosen included the Day-Night Stroop executive function test, a version of the color Stroop test (Golden, 1978) adapted for young children. The Day-Night Stroop (Gerstadt, Hong, & Diamond, 1994) tested children's inhibition of the prepotent response of matching the word they said to the object shown. The children had to say 'night' for each card showing the sun and 'day' for each card showing the night sky. Following two warm-up items, there were 16 test trials in random order with no corrective feedback. Children have difficulty suppressing the more usual associations of sun/day and moon/night, and their ability to maintain the new rule is measured by the total correct out of 16. No norms have been established for Chinese samples, though the test should be fairly culturally neutral.

A visual spatial intelligence test, the Primary Test of Nonverbal Intelligence (PTONI) (Ehrler & McGhee, 2008), was used as an index of nonverbal intelligence. The PTONI is especially appropriate for testing children who have underdeveloped verbal and/or motor skills. The publishers claim a certain cultural neutrality, in that PTONI directions are provided in eight alternative languages (including Mandarin), making it an appropriate assessment of intelligence for children from diverse language backgrounds, according to Ehrler and McGee (2008). It was normed on a culturally and ethnically diverse demographic sample of 1,010 children from 38 states in the United States. The test format requires a child to look at a series of pictures and to point to the one picture that does not belong with the others. Items are arranged in order of difficulty. The PTONI provides standard scores, percentile ranks, and age equivalents. Though not normed in mainland China, PTONI is a non-verbal test and is purported to be largely free of cultural bias.

Finally, short-term auditory memory was measured using a forward digit span task (backward digit span has some advantages in tapping working memory but would have been unsuitable for children younger than five). This task was designed so that there were four items of each length, for example, four 4-digit strings, four 5-digit strings and so on. Increasingly long digit strings were presented until the child failed to get 75% of the items at that length correct. The highest level at which they achieved 3 or 4 items correct was taken as their digit span.

This procedure closely mimics the procedure used in standardized tests such as WISC-V (Wechsler, 2014) and the Differential Ability Scales (Elliott, 2006). It is also parallel to earlier research that examined digit span in China. That research found scores to be slightly higher than Western samples (Chen & Stevenson, 1988). Properties of the Mandarin digits were considered in creating a digit span test for Mandarin. All Mandarin digits are monosyllabic, but some of them rhyme, so we chose strings in which rhymes were not adjacent, and in which there were no difficult phonetic sequences. As in other digit span tests, we checked carefully to ensure that no number sequences were frequent idioms (as with 911 in English).

References

Chen, C., & Stevenson, H. W. (1988). Cross-linguistic differences in Digit Span of preschool children. *Journal of Experimental Child Psychology*, 46, 150–58.

Ehrler, D. J., & McGhee, R. L. (2008). *Primary Test of Nonverbal Intelligence (PTONI). Manual.* Austin, TX: pro-ed.

Elliott, C.D. (2006). *Differential Ability Scales—Second Edition (DAS-II)*. Ontario, Canada: Pearson.

Gerstadt, C., Hong, Y., & Diamond, A. (1994). The relationship between cognition and action: Performance of children 3–7 years old on a Stroop-like day-night test. *Cognition*, *53*, 129–53. http://dx.doi.org/10.1016/0010-0277(94)90068-X

Golden, C. J. (1978). *Stroop color and word test: A manual for clinical and experimental uses*. Wood Dale, IL: Stoelting Company.

Wechsler, D. (2014). Wechsler intelligence scale for children-fifth edition. Bloomington, MN: Pearson.