https://doi.org/10.1016/j.bandl.2023.105268

Developmental Language Disorder in Chinese Children: A Systematic Review of Research from 1997 to 2022

Li Sheng[a], Jiayu Yu[a], Pumpki Lei Su[b], Danyang Wang[c], Tzu-Hung Lu[d], Lue Shen[e], Ying Hao[f], & Boji Pak Wing Lam[g]

[a]The Hong Kong Polytechnic University, [b]University of Texas-Dallas, [c]University of California-Urvine, [d]University of Delaware, [e]Boston University, [f]Nanjing Normal University, [g]University of North Texas

Address correspondence to Li Sheng, Research Centre for Language, Cognition, and Neuroscience & Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong SAR, China. Email: lsheng@polyu.edu.hk. Tel: 852-27667446, Fax: 852-23340185.

Acknowledgments

The authors wish to thank Dr. Anita Wong for her comments on a previous version of the paper, and Dr. Julie (Hsin-Jen) Hsu for providing information on language assessment tools in Taiwan.

Abstract

Developmental language disorder (DLD) is one of the most common neurodevelopmental disorders. The influences of DLD on language development have been delineated in detail in English. The same is not true for Chinese, a group of Sinitic languages with distinct typological features that may modify the profile of DLD crosslinguistically. We conducted a systematic search of English and Chinese journal databases and reviewed 59 studies on the manifestations of DLD in Chinese. Methodological quality appraisal of the literature revealed several areas of improvement to enhance transparency and replicability. A bibliometric analysis indicated a steep growth trajectory of this literature. Examination of the participant selection and diagnostic criteria revealed limitations and calls for the development of assessment tools and increased knowledge of evidenced-based diagnostic practice. Areas of deficits demonstrated by Chinese children with DLD were synthesized qualitatively and discussed in light of the literature on clinical markers of DLD in English.

Keywords: developmental language disorder, Chinese, Mandarin, Cantonese, clinical marker, diagnostic accuracy

1. Introduction

1.1. The need for a systematic review on Chinese DLD

Developmental language disorder (DLD, also known as Specific Language Impairment, or SLI) is a neurodevelopmental disorder that negatively affects a person's ability to understand and speak their native and subsequently learned language(s) (Bishop et al., 2017). Since the 1980s, this population has received intense research attention (Leonard, 2020). Deficits in finite verb morphology, nonword repetition, and sentence repetition are the most established phenotypic markers of DLD in young children who speak English, the most studied and served population (Archibald & Joanisse, 2009). Much of the theorizing on DLD is also centered on findings from English and other Indo-European languages, leading to theoretical accounts that revolve around characteristics of these well-studied languages that may not be universally applicable (Kidd & Garcia, 2022). For instance, the extended optional infinitive account (Rice & Wexler, 1995) and the surface account (Leonard et al., 1992) both aim to explain the extraordinary difficulties with tense/agreement marking morphemes exhibited by English-speaking children with DLD. These theories bear less directly on languages that do not mark tense and agreement. In this review, we focus on Chinese, a group of Sinitic languages. Chinese is tonal, has relatively simple syllable structures, a sparse morphology, a heavy reliance on word order, and a rich inventory of aspect markers and classifiers. Nouns and verbs are not marked for number, gender, case, or tense. The use of aspect markers is grammatically optional but interfaces with event semantics and pragmatic conditions (Li & Thompson, 1989; Matthews & Yip, 1994). These distinct linguistic features in Chinese place different demands on its learner and likely induce language-specific vulnerabilities. Thus, the study of Chinese DLD may provide novel evidence for building generalizable accounts of this disorder.

Practically speaking, there is a pressing need to raise awareness of and enhance clinical services for DLD in all countries. Though the linguistic and cognitive phenotypes of DLD in English speakers are well delineated (Leonard, 2014), this disorder remains a hidden disability in English-speaking countries (McGregor et al., 2020). With a prevalence rate of around 7-11% (Norbury et al., 2016; Tomblin et al., 1997), DLD is more common than other widely known developmental disorders, such as autism and attention deficit and hyperactivity disorder. However, only 18% of children who suffer from DLD are identified and receiving professional help in the United States (Zhang & Tomblin, 2000). Under-identification is likely to be far more frequent in speakers of Chinese, given the lack of awareness of language disorders and the lack of referral systems and assessment instruments in mainland China (He, 2020; Zhang et al., 2021), where the largest number of Chinese speakers reside. While still in its infancy, research on Chinese DLD is gaining traction among scholars from various disciplines. The goal of the current study is to provide an up-to-date review of the literature on Chinese DLD and summarize the evidence from studies that assessed the deficits demonstrated by children with DLD relative to typically-developing controls (for reviews of grammatical deficits in Cantonese DLD, see Fletcher et al., 2009; and Wong, 2023).

1.2. Chinese language background

Chinese languages belong to the Sinitic language family and are commonly divided into 10 main dialect/language[1] groups (Ministry of Education of the People's Republic of China, 2021). Standard Mandarin is the largest Chinese dialect and the official language of mainland China, Taiwan, and Singapore (see Table 1). Though the 10 main Chinese dialects are not

[1] In official documents, the different varieties of Chinese are labeled as dialects, due to the fact that the different spoken forms share the same written form. From a linguistic perspective, the differences between these dialects are analogous to those of mutually unintelligible languages of the same family (e.g., French and Spanish) (Tardif et al., 2009). Here, we use "dialect" and "language" interchangeably to recognize both of these realities.

mutually intelligible to each other, the majority of the population can communicate via the common tongue of Mandarin, as 80.72% of the population in mainland China can speak Mandarin (Ministry of Education of the People's Republic of China, 2021) and 96.8% of the population in Taiwan uses standard Mandarin as either a primary (66.4%) or a secondary (30.4%) language (Directorate General of Budget, Accounting and Statistics, Executive Yuan, 2020). The only other Chinese language that enjoys the official status is Cantonese, which is the official language in Hong Kong and Macau Special Administrative Regions (hereafter referred to as Hong Kong and Macau). Mandarin is the most spoken language in the world, with 918 million first-language speakers and 199 million second-language speakers; Cantonese is also a widely spoken language with more than 73 million first-language speakers worldwide (Kan et al., 2020).

Cantonese and Mandarin differ the most in phonology. For instance, one frequently cited difference is that there are six lexical tones in Cantonese but only four in Mandarin (Fung, 2009). Cantonese and Mandarin share many similarities in the domain of morphosyntax: both rely on word order to express grammatical relationships and both have subject-verb-object as the basic word order. Differences also exist. For example, Cantonese ditransitive sentences place the direct object before the indirect object, whereas Mandarin ditransitive sentences use the opposite word order. Also, the omission of the agent is permissible in Mandarin passive sentences but not in Cantonese passives. The Mandarin vernaculars spoken in mainland China and Taiwan show some differences in vocabulary and pronunciation, due to the influences of Japanese and Southern Min dialect on Taiwanese Mandarin, differences in translations of foreign words, and historical and political reasons (Cheng, 1985; Zhou & Zhou, 2019).

In terms of written language, the Chinese orthography may adopt the simplified script, which is used in mainland China and Singapore; or the traditional script, which is used in Hong

Kong, Macau, and Taiwan. Many traditional characters are visually more complex than simplified characters. For instance, the word "deaf" is written as "聋" in simplified Chinese, and "聾" in traditional Chinese. However, readers who are taught to read one script are usually capable of reading the other script. Moreover, despite phonological, semantic, and syntactic differences in the oral form, Mandarin, Cantonese, and other Chinese languages share the same writing system, giving rise to "a situation of multiple languages in the oral format but not in written form" (Reetzke et al., 2015, p.817).

To capture the growth trajectory of Chinese DLD research, we will describe the bibliometric characteristics of the literature focusing on the geographic locations of the participants and the specific Chinese dialect/language spoken by the participants. To present a more comprehensive picture of the status of this literature, we will include in our search both English language journals and Chinese language journals.

1.3. Evidence-based diagnostic criteria

To investigate the area and degree of deficits exhibited by a clinical population, one must apply a pre-determined set of criteria to select the clinical group and a relevant comparison group. An integral component of DLD diagnosis involves administering standardized norm-referenced language tests and separating individuals into typical and disordered groups using a cutoff score (Bishop et al., 2017; Reilly et al., 2014). DLD diagnosis is not a cut and dried process. Much variability exists in the criteria applied in both research studies and real-life decisions in English-speaking children, the most studied population. Spaulding et al. (2006) noted that researchers used scores between 1SD and 1.5SD below the normative mean on one or more language tests to select participants with DLD, whereas school systems in the United States required between 1.5SD and 2SD below the mean on one or more language tests to qualify

students for service. A well-known epidemiological study in the US (Tomblin et al., 1997) operationalized DLD as scores of 1.25SD below the mean on 2/5 language composites. A more recent epidemiological study in the UK (Norbury et al., 2016) operationalized DLD as scores of 1.5SD below the mean on 2/5 language composites. Finally, the ICD-10 criteria for language disorders (World Health Organization, 1992) require scores of -2SD or below on 2/5 language composites (Norbury et al., 2016). In short, stakeholders have different definitions of the demarcating line dividing typical from impaired language performance.

Evidence-based diagnostic criteria are advocated to overcome the arbitrariness in diagnostic decision making (Nitido & Plante, 2020; Spaulding et al., 2006). Specifically, tests used for identifying language impairment should have at least acceptable diagnostic accuracy (i.e., greater than 80% sensitivity and specificity, Plante & Vance, 1994). Moreover, instead of using a single arbitrary cut score across tests, the cut score should be test- and age-specific and should show empirical evidence of acceptable sensitivity and specificity values (Plante & Vance, 1994). Whereas only one out of 21 English tests of child language achieved acceptable diagnostic accuracy in the mid-1990s (Plante & Vance, 1994), five out of 43 commercially available English tests of child language reported acceptable diagnostic accuracy in the mid-2000s (Spaulding et al., 2006). Though the number of tests that meet diagnostic standards is on the rise, not all researchers, arguably the most informed segment of stakeholders, routinely use evidence-based methods to identify their participants. Nitido and Plante (2020) examined the use of diagnostic criteria in studies of DLD published between 2015 and 2019. They limited the evidence to studies conducted with native English speakers in the United States because validated diagnostic measures are not available in many countries and languages. The review showed that the majority of the 90 studies conducted in the US with English speakers did not use

evidence-based diagnostic criteria. To be specific, only 38 articles (42%) reported that they strictly used validated diagnostic methods in participant selection, despite that 13 out of the 30 tests used by these studies reported more than acceptable sensitivity and specificity. Failures to use evidence-based diagnostic practice could threaten the validity of research findings: using a test with low sensitivity could lead to under-identification and a biased DLD sample consisting of children with more severe impairment, and inflated probability of finding significant differences between the DLD and control groups on researcher-designed tasks. Conversely, using a test with low specificity could lead to over-identification and a sample consisting of children not affected by DLD, making the comparison with the control group invalid.

The variability in and the evolving nature of the diagnostic standards observed in the English DLD literature highlight the importance to examine the participant selection criteria in the Chinese DLD literature. Cognizant of the shortage of assessment instruments in Chinese, our intention is not to criticize studies for not using evidence-based diagnostic criteria, but to document the field's current practice in order to inform future directions.

1.4. Aims of the review

This review aims to give a thorough portrayal of Chinese DLD research, understand the status quo of the literature, identify gaps, and make recommendations for future studies. The specific questions we aim to address are:

1. How is the methodological quality of this literature? We deemed this an important question to guide the future development of this line of research and to determine whether or not a meta-analysis should be conducted in the current review. We used a critical appraisal checklist to evaluate the quality of identified articles and summarize the results.

2. What are the bibliometric characteristics of the Chinese DLD literature? To address this question, we reported the level of research output by year, journal type (English language, Chinese language), and participants' geographic region.

3. What diagnostic criteria are used to identify participants with DLD? To answer this question, we summarized the diagnostic tools, cut scores, and psychometric properties of the tools, if available, used in the included studies.

4. How is DLD manifested in Chinese languages? What are the areas and magnitude of deficits? To answer these questions, we provided a qualitative description of the study foci in Chinese DLD research. A meta-analysis to report the size of deficits was not conducted in light of the methodological quality appraisal findings in question 1.

2. Method

2.1. Searches

To identify potential articles, we searched from the first available date to present using the databases and search terms specified in Table 2. Initial searches were conducted in June 2020 and a new round of updated and more comprehensive searches were performed simultaneously in July 2022 in three sets of databases: 1) English databases (ProQuest, which contains 62 databases such as ERIC, PsycINFO, and LLBA; and PubMed), 2) Chinese databases from Mainland China (CNKI, hereafter referred to as simplified Chinese), and 3) Chinese databases from Taiwan (ITPLS and Airiti; hereafter referred to as traditional Chinese). Even though Chinese (Cantonese or Mandarin) is an official language in Hong Kong and Singapore, we were unable to identify any relevant academic journals published in Chinese at these two locations. We found a virtual library of Chinese journals published in Macau. But the search terms "語言損傷 (language impairment) OR 語言障礙 (language disability) OR 語言遲緩 (language delay)"

returned zero result in the virtual library. We also conducted searches targeting other major Chinese dialects (e.g., Wu dialect, Southern Min dialect) in the simplified and traditional Chinese databases. These attempts also resulted in zero entry.

Following the searches, a primary coder screened the titles and abstracts of all non-duplicated results, and a secondary coder independently screened one-third of the results, to determine whether or not the articles would potentially meet the inclusion and exclusion criteria (specified in the next section). Reliability was above 90% for all three sets for title screening and ranged from 88% to 93% for abstract screening. Full text articles were assessed by two independent coders regarding the article's eligibility. Each included article was coded on participant characteristics and outcome measures by two coders. Participant characteristics included age, gender, geographic location, language spoken, other dialect/language exposure, diagnostic criteria, DLD and TD matching criteria, standardized language test names and scores, nonverbal IQ test names and scores, reliability and validity information of the language and nonverbal IQ tests. Outcome measures were listed one by one, and each measure was coded in regard to type (neural/behavioral), modality (expressive/ receptive), task procedure, group means and SDs, and effect size (Cohen's *d*). Disagreements were resolved by consensus at each stage.

2.2. Inclusionary and exclusionary criteria

Included studies must meet the following criteria:

- Published in a peer-reviewed journal. We set this criterion so that only studies with relatively high methodological rigor are included.
- Included Chinese-speaking children with DLD as the participants.
- Used a conventional definition of DLD, including a specific reference to low receptive/expressive language skills based on at least one standardized test and/or

language sample (e.g., 1 SD – 1.5 SD below the mean), concerns by familiar adults (i.e., parents, teachers), clinical diagnosis, or a combination of these[2].

- Used a group comparison design by including both a group of participants with DLD and a typically developing comparison group that was matched to the clinical group on chronological age or language level.
- Contained participants whose age fell into the range from preschoolers to high school students (i.e., 3-18 years of age).
- Contained at least one outcome measure that assessed speech, language, literacy, or cognition at the behavioral and/or neurological level.

Studies were excluded if they:

- Were master's or doctoral theses, conference proceedings, or book chapters.
- Contained only participants who were outside of the 3-18 years age range.
- Contained only parent outcomes, genetic outcomes, and social-emotional outcomes.

2.3. Quality rating

We used the JBI systematic reviews critical appraisal checklist for analytical cross-sectional studies to rate the methodological quality of the studies (Moola et al., 2020). This tool includes eight criteria to evaluate the internal validity (e.g., "*were objective, standard criteria used for measurement of the condition*?") and reporting quality (e.g., "*were the study subjects and the setting described in detail*?") of a study. Based on the explanation of the criteria in the scoring manual, we added specific descriptors relevant to the study of DLD (see Table 3 for the checklist). Each study that passed full article screening was appraised by two reviewers, who

[2] A conventional definition of DLD would also require that the study include participants whose nonverbal intelligence is within normal limit (e.g., within 2 SD of the mean). However, a number of studies did not mention nonverbal IQ scores. We did not exclude studies that are missing this information from our review. But we documented whether or not nonverbal IQ scores were reported.

achieved consensus on all the scores through discussion and consultation with the first author. Studies that failed to meet more than three of the eight criteria (i.e., has > 3 "*no*"s) would not be included in the systematic review.

3. Results

The results from the searches are detailed in Figure 1. The searches yielded a total of 1764 records. After removing 672 duplicates, 1092 results went through title screening. Nine hundred thirty-five titles were eliminated after title screening. The remaining 157 results went through abstract screening, which further eliminated 92 results. Sixty-five full text articles were each read by two independent coders to assess the articles' eligibility. Eight articles were eliminated after reading the full text for not meeting the inclusion and exclusion criteria. After title, abstract, and full text screening, 57 articles remained. In addition, one article received more than three "no"s on the quality appraisal checklist and was excluded from the review. Parallel measures were taken to seek out additional articles by contacting authors who publish on this topic, searching the reference lists of published studies as well as the publication lists of authors, and setting auto-alerts on relevant research platforms. Three additional articles published in Chinese were identified from these sources. The following sections report results on the 59 included articles.

3.1 Quality appraisal

Quality appraisal results are presented in Table 3. For the first appraisal question "*were the criteria for inclusion in the sample clearly defined*?", a majority of the studies (52/59) clearly defined their inclusionary criteria for both DLD and TD groups, three studies only provided inclusionary criteria for the DLD group, and four studies did not describe criteria for either group. For question 2 "*were the study subjects and the setting described in detail?*", we coded

whether or not information on participant age, sex, language assessment scores and recruitment site was provided. A majority of the studies (44/59) received an "incomplete". Specifically, one study was missing participant age, 17 studies were missing language assessment scores, 19 studies were missing recruitment sites, and 24 were missing sex distribution of the participants. For question 3 "*was language ability measured in a valid and reliable way?*", the majority of the studies (56/59) received a "yes" as they either used published standardized tests or reported the validity and reliability of self-designed tests. The three remaining studies relied on clinician referrals and did not use any language tests. For question 4 "*were standard criteria used for measurement of the condition?*", given the wide variability noted in previous literature, we adopted a lenient criterion and assigned a "yes" as long as the study mentioned at least -1SD below the mean on one or more language measures and nonverbal IQ of > 70. A majority of the studies (49/59) received a "yes. Seven articles fulfilled the language measure but did not provide the nonverbal IQ criterion. Three studies did not state how the condition was measured and received a "no".

For question 5 "*were confounding factors identified?*", we coded whether or not the study considered three potential confounding factors that could lead to group differences unrelated to the diagnosis: age, socioeconomic status (SES), and amount of dialect/language exposure. Only two studies statistically compared the two groups on all three variables. A majority of the studies (54/59) compared the two groups on at least one but not all three variables. Among them, one study missed age, 48 studies missed SES, and 54 studies missed dialect/language exposure. Three studies missed all three factors. For question 6 "*were strategies to deal with confounding factors stated?*", of the two studies that considered all three potential confounds, one dealt with the unmatched factor through statistical means, and one study did not. For the remaining 57

articles, this item was "not applicable" because the study did not report any information on these factors in the first place (3 studies), or the study considered some of the potential confounds (e.g., age) and found them to be well-matched.

For question 7 "*were the outcomes measured in a valid and reliable way*?", the outcomes were divided into subtypes. Language production measures (e.g., sentence completion, language sampling) would require inter-rater reliability check of transcribing/coding/scoring, whereas measures that do not require a verbal responses (e.g., picture pointing, EEG), the description of the measures should show face validity. Forty-three articles received a "yes" and 16 articles received a "no". For question 8 "*was appropriate statistical analysis used*?", 15 articles checked the assumptions and reported them to support the selected statistical analyses, 43 articles appeared to have used appropriate analyses but did not report the check of assumptions, and one study used wrong analysis method.

To summarize, a majority of the articles received "*yes*" for four of the appraisal items (#1, 3, 4, 7); for three of the items (#2, 5, 8), the majority of the articles were rated as "incomplete" or "unclear". Only six articles received *yes* on 6 or more appraisal items. The numbers of studies receiving three, four, and five *yes*es were 18, 18, and 13, respectively. In light of these findings, we decided to not accompany the qualitative summary of the literature with a meta-analysis. Table 4 provides a full list of the 59 included articles and a description of the bibliometric and sample characteristics of each.

3.2 Bibliometric characteristics

3.2.1 Number and language of publications

Figure 2 depicts the growth trajectory of this literature, presenting the number of articles published in roughly 5-year periods from three databases. Stokes and So (1997) published the

first study of Chinese DLD in the *Asia Pacific Journal of Speech, Language, and Hearing*. The scholarly attention to Chinese DLD was limited until 2015, when there was a sharp uptick in the number of peer-reviewed publications that more than doubled the size of the literature from 26 to 59 between 2016 and now. Over the past 25 years, studies on Chinese children with DLD were mainly published in English journals (66%). The predominance of Chinese DLD studies published in English is especially robust between 1997 and 2010 (73%). Articles from simplified and traditional Chinese databases have also risen in the last five years. These findings indicate increasing attention to Chinese DLD in both English-speaking and Chinese-speaking communities.

3.2.2 Location of and language spoken by the participants

Table 4 presents the geographical locations of and language spoken by the participants. Though there are no new studies of Cantonese DLD that fit our criteria since 2017, research conducted in Hong Kong still takes up a large share of this literature, contributing to 32% of the publications (19 articles), all of which were published in English journals. Studies of Mandarin DLD conducted in Taiwan comprise 25% of the literature, appearing in English (8 articles) and traditional Chinese (7 articles) journals. Finally, studies of Mandarin DLD conducted in Mainland China first appeared in 2012. Despite the relatively short publication history, there is a rapid catch-up, with studies featuring participants from Mainland China comprising 42% of the literature. These articles appeared in English (12 articles) and simplified Chinese (13 articles) journals.

Another metric to examine research activity by region is to count the number of unique participants from Hong Kong, Mainland China, and Taiwan. So far, studies from Taiwan have

included the greatest number of unique participants with DLD (n = 331), followed by those from Mainland China (n =292) and Hong Kong (n = 290)[3].

3.3 Participant selection criteria

3.3.1 Sample size and age

Table 4 presents the sample size, age, and nonverbal IQ scores of the study participants. The average sample size per study is 20 children with DLD and 38 age-matched TD controls[4]. A majority of the articles (41 articles, 69%) included between 11 and 20 participants in the DLD group. Four studies (7%) included fewer than 10 participants with DLD; six studies (10%) included between 21 and 30 DLD participants, and eight studies (14%) included more than 31 participants with DLD. Age-wise, the mean age of the DLD group ranged from 39 months to 139 months. We used the mean age of the group to divide the samples into school-age (i.e., mean age $\geq$ 72 months) or preschool-age (i.e., mean age < 72 months), because 6 years of age is the typical age at which children enroll in elementary school in mainland China (Compulsory Education Law of the People's Republic of China). According to this definition, 17 of the studies (29%) included primarily school-age children, whereas the remainder included primarily preschool-age children.

3.3.2 Comparison group

[3] Among the Hong Kong-based studies, participants from Wong et al (2017) and Wong et al. (2015) overlapped, as did Stokes & Fletcher (2003) and Stokes & So (1997). Participant overlap also exists among Fletcher et al. (2005), Leonard et al. (2006, 2007), Stokes et al. (2006), Wong et al. (2004), and Wong et al. (2010b). Among the Mainland China-based studies, Kim et al. (2019) and Wang et al. (2020) reported the performance of the same group of 15 children with DLD. He & Dai (2012), He & Yu (2013), He et al. (2013), Xu et al. (2020), Yu (2016), Yu et al. (2017), Zeng et al. (2013) and Zeng et al. (2018) reported the performance of the same group of 12 children with DLD. Participant overlap also exists among Wang & Yu (2021), Yu et al. (2021), and Yu et al. (2022). Taiwan-based studies appeared to have no overlap in participants.

[4] Two articles included a language-matched control group only and were not included in this calculation. To et al. (2010) included the normative sample (n= 1080) of the Hong Kong Cantonese Oral Language Assessment. If this study is excluded, the average sample size is 20 for the DLD group and 20 for the TD group.

With regard to the TD comparison group selection, 35 studies (59%) included chronological age-matched controls only; 19 studies (32%) included both an age-matched control group and a language-matched (e.g., matched on mean length of utterance) control group; three studies (5%) included both an age-matched control group and a younger TD control group that was on average 18 months younger than the DLD group; and two (3%) study included a language-matched control group only.

3.3.3 Tests and cut scores

Figure 3 addresses how DLD and TD groups were selected in the studies reviewed, specifically, the number of language tests (including subtests from a comprehensive test battery) on which the participants must score poorly to be included in the DLD group. Twenty-two studies (37%) categorized a child into the DLD group if they scored below the cutpoint on one test; 31 studies (53%) required that the child scored below the cutpoint on at least two tests; and three studies (5%) required that the child scored below the cutpoint on at least three tests. In addition, three studies (5%) did not report standardized language test scores but relied on clinical referrals from speech-language pathologists or pediatricians as the basis for DLD inclusion.

Table 5 shows the tests used for diagnosing participants in Hong Kong, Mainland China, and Taiwan, respectively. Eighteen of the 19 studies conducted in Hong Kong used standardized language tests. To et al. (2010) was an exception: this study examined narrative performance of the normative sample of the *Hong Kong Cantonese Oral Language Scale* (HKCOLAS, T'sou et al., 2006) and included 50 children with DLD, whose diagnosis was made by experienced Speech-Language Pathologists using "an informal assessment checklist that observed aspects of semantics, morphosyntax, and pragmatics" because "there was no standardized reference test in Hong Kong at the time of this study (hence, the development of the HKCOLAS)" (p.654).

Twelve of the 14 Hong Kong-based studies published in or before 2010 used the *Reynell Developmental Language Scales* (RDLS, Hong Kong Society for Child Health and Development, 1987), a test adapted from English into Cantonese and standardized for local use (as described in Klee et al., 2009). However, nine studies used the receptive subscale and three studies used the expressive subscale. Klee et al. (2004) used both the Hong Kong Cantonese Receptive Vocabulary Test (Lee et al., 1996) and the RDLS receptive subscale and required that the participants with DLD score lower than 1SD below the mean on either test. Diagnostic accuracy information is not available for the RDLS. Five studies published in or after 2010 used the HKCOLAS and the diagnostic criterion recommended by the test developer, which is 1.25SD below the mean on at least two out of the six subtests. According to the test developers (T'sou et al., 2006), this criterion resulted in a sensitivity of 98.2% and a specificity of 94.6% in a sample of 56 children with DLD referred by clinicians and 56 TD age-matched peers from the norming sample of the HKCOLAS.

The 25 published studies conducted in Mainland China used a total of nine different tests, most of which were not designed to be diagnostic tools for language disorders. Specifically, 80% of the studies used a standardized test of receptive vocabulary – the *Peabody Picture Vocabulary Test-Revised* (PPVT-R, Sang & Miao, 1990); 14 studies used either the *Revised Language Disorder Scale for Preschool Children* (RLDS-preschool, Lin et al., 2008a), or the *Revised Language Disorder Scale for School-Age Children* (RLDS-school age, Lin et al., 2008b), both of which were designed for Taiwan Mandarin and normed on children from Taiwan; seven studies used the verbal subtests of published IQ tests such as the *McCarthy Scale of Children's Abilities* (The team of MSCA-CR, 1991) and the *Wechsler Intelligence Scale for Children Revised* (WISC-R, Gong & Cai, 1994); five studies used the *Diagnostic Receptive and Expressive*

Assessment of Mandarin (DREAM, Ning et al., 2014); one study used the *Criterion-Referenced Diagnostic Test of Mandarin-Speaking Preschool Children with SLI* (Ning, 2012, 2013); one study used the *Children's Communication Checklist* (Bishop, 1998), and two studies used a set of self-designed language battery. Of the nine tests used, we were able to find diagnostic accuracy values for only one: the DREAM was validated against a combination of pediatricians' judgment and spontaneous language samples. When a cutoff score of 80 on any one of the five DREAM components was applied, an optimal level of accuracy was achieved, yielding sensitivity of 95% and specificity of 82% in differentiating children with and without DLD (Liu et al., 2017).

The 15 studies conducted in Taiwan used a total of 10 different tests. The RLDS-Preschool (Lin et al., 2008a) and the RLDS-School age (Lin et al., 2008b) were the most popular choices and were used by more than half of the studies. The PPVT-R (Lu & Liu, 1998) and the *Wechsler Intelligence Scale for Children* (WISC, Chen, 1997; Chen & Chen, 2013) were each used by three different studies. The *Children's Oral Language Comprehension Test* (Lin & Chi, 2002) was used by two studies. The five remaining instruments were each used by one article. To the best of our knowledge, diagnostic accuracy information was unavailable for these instruments.

In regard to the cut score adopted by research studies when determining the DLD/TD status of their participants, variability is noted. A majority of the studies (36 articles, 61%) used 1.2-1.3 SD below the mean (roughly the 10^{th} percentile) as the cutoff, seven studies (12%) adopted 1.5 SD below the mean (roughly the 7^{th} percentile), and 13 studies (22%) adopted 1SD below the mean (roughly the 16^{th} percentile). Three studies (5%) did not use standardized tests and one study (2%) did not report the cutoff.

The last column of Table 5 provides information on the cut scores applied to each individual test. For some of the most popular tests (e.g., RDLS receptive subscales, PPVT, RLDS), variable cut scores are adopted across studies. This happens even when the test developer has recommended a specific optimal cut score (e.g., DREAM).

3.4 Areas of deficits: Qualitative description

As summarized in Table 6 and Supplementary Materials Table 1, the 59 articles included a large number of outcome measures spanning various domains of oral language and some elements of written language. These outcomes also vary in granularity: some are broad-based measures of language ability (e.g., number of different words in spontaneous language samples); others are precise measures of specific skills (e.g., discrimination between /ba/ and /da/). In the following section, we give a qualitative summary of the areas of studies, arranged by language domain (e.g., grammar, phonology) or construct (e.g., productivity). Within a domain/construct and across domains, the description is ordered according to the size of the evidence base of the outcome measures.

3.4.1 Grammar

Table 6 summarizes the findings by outcome measure. The grammatical domain has attracted the most attention, with 39 articles investigating elements of grammar. These articles have included both generic measures of grammar and fine-grained measures targeting the comprehension and production of specific grammatical constructions. Three of the generic measures of grammar are derived from conversational or narrative language samples. Among them, mean length of utterance (MLU) is the most common measure and was included in 10 studies (see Table 6 for the study references). The second most common is a set of complexity measures such as proportion of complex clauses, subordination index, and syntactic complexity,

which were included in six independent investigations. Grammaticality is the third generic measure from language samples and an outcome measure in three studies. The other three generic grammar measures were sentence/syntactic comprehension, sentence repetition[5], and artificial grammar learning.

Of the fine-grained grammatical measures, aspect markers are the most studied element, with nine independent investigations targeting this as the sole outcome measure or one of the outcome measures. These studies used sentence-picture matching (Chen et al., 2022; Chen & Durrleman, 2022), elicited production (e.g., Chen et al., 2021; He et al., 2013; Fletcher et al, 2005; Stokes & Fletcher., 2003), languages sampling (Hao et al., 2018; Stokes & Fletcher, 2003; Wong et al., 2003), and self-paced reading (Yu et al., 2019) paradigms. The comprehension and/or production of passive construction and relative clauses each garnered research attention in four studies. The production of classifier phrases, a construction at the interface of grammar and semantics, is the outcome measure in three studies. The rest of the grammatical outcomes were included in one or two studies, including Ba-sentence (a non-canonical sentence that follows the Subject-Ba-Object-Verb word order) comprehension and production, Wh- question comprehension and production, negation production, modal auxiliary production, and topic structure comprehension and production.

In summary, of the 51 grammar outcome/study combinations listed in this table, group-level differences between DLD and age-matched TD controls were detected in 42 (82%). Eleven grammar outcome measures were included in at least two studies. Eight of the 11 measures revealed largely consistent group differences across studies (i.e., $\geq$ 75% of the studies yielded

[5] Sentence repetition taps on memory, semantic, and grammatical skills. Here we categorize it under the grammatical domain for the sake of simplicity and following the argument that it is primarily a measure of linguistic skills rather than memory abilities (Archibald & Joanisse, 2009).

significant group differences). Grammaticality was an exception as two of the three studies did not find a significant group difference on this measure. In the discussion, we will revisit this finding. Ba-sentence comprehension and production and negation production also yielded ambiguous results: of the two studies under each construction, one study of preschool-age children reported significant group differences, the other study of school-age children did not report significant group differences.

3.4.2 Vocabulary

The second most popular domain of focus is vocabulary, with 13 studies contributing to the understanding of vocabulary deficits, and four studies contributing to the understanding of novel word learning. Twelve of the 13 studies measured lexical diversity from conversational or narrative language samples, though the actual measures differed and included the number of different words (NDW), D (a lexical diversity measure that is relatively less affected by sample size), type token ratio (TTR), corrected TTR, moving average TTR, measure of textual lexical diversity (MTLD), and sophisticated vocabulary count. Only one study examined standardized vocabulary test scores (Wong et al., 2009). Table 6 shows that of the 17 outcome/study combinations listed, group-level differences between DLD and age-matched TD controls were detected in 15 (88%). All four studies on word learning outcomes found word learning to be an area of deficit in Chinese DLD.

3.4.3 Phonology

The third most studied domain in the Chinese DLD literature is phonology which is included in 13 articles. Nine articles utilized measures of phonological memory. Other phonological domains such as speech perception, phonological awareness, phonological retrieval, phonological representation, and speech production were included in one to five

articles. According to Table 6, 18 out of 24 outcome/study combinations (75%) reported significant group differences between DLD and age-matched TD groups on phonology.

3.4.4 Productivity

Nine studies examined the amount of language produced by children in narrative (Hao et al., 2018; Lai & Wang., 2017; Sheng et al., 2020; Torng & Sah, 2020; Tsai & Chang, 2008; Xue et al., 2022) and conversational (Stokes & Fletcher, 2000; Wong et al., 2010b; Tseng & Liu, 2017) sampling contexts. These studies relied on two measures that respectively count the total number of utterances (TNU) or the total number of words (TNW) generated by children. Table 6 indicates that group differences between DLD and age-matched TD controls were only observed in four out of 13 outcome/study combinations (31%).

3.4.5 Narrative discourse

Nine studies examined group differences in narrative discourse measures. Seven studies investigated the production of story macrostructure, but differences existed in how stories were elicited and analyzed. Six studies used fictional stories based on pictorial stimuli (Chi et al., 2012; Hao et al., 2018; Lai & Wang, 2017; Sheng et al., 2020; Torng & Sah, 2020; Xue et al., 2022), and one study (Tsai & Chang, 2008) elicited personal narratives. Five of the six fictional narrative studies used Stein and Glenn's (1979) system of story grammar analysis; Xue et al. (2022) used the high-point analysis model (Peterson & McCabe, 1983). The personal narrative study (Tsai & Chang, 2008) adopted Labov's (1972) story analysis model. Five studies of narrative discourse explored measures of pragmatics, encompassing the use of referential devices, connectives, and evaluative comments. Table 6 indicates consistent DLD-TD discrepancies across studies on measures of story macrostructure and discourse pragmatics (both 100%). One study included a story comprehension outcome as a part of a word learning task that

embedded novel words into stories. This study found that children with DLD were poorer in understanding the story contents than TD age peers.

3.4.6 Literacy

Four studies delved into written language. These studies measured word reading, reading comprehension and its strategies, word dictation, morphological awareness, and orthographic skills. According to Table 6, differences in literacy measures between DLD and TD groups were detected in five of the nine comparisons (56%).

3.4.7 Miscellaneous areas

The remaining areas of focus could not be subsumed under other headings and did not have subcategories. These outcome measures were included in three or fewer studies. The outcomes are general language abilities as measured by standardized language tests, executive functions, verbal working memory, verbal and spatial memory, and fluency. Table 6 indicates that two studies showed general language deficiencies in Chinese children with DLD.

4. Discussion

4.1 Methodological and reporting quality

The growing recognition for reproducibility has propelled the development of reporting standards in the scientific community. For instance, the EQUATOR (Enhancing the Quality and Transparency of Health Research) network provides a free online library of reporting guidelines for many types of studies. More and more journals now require authors to follow these guidelines when reporting their methodology and writing up the findings. In this spirit, we note that many articles included in the current review are missing a number of details in the participant inclusion and group matching criteria. Improper participant selection could threaten the validity of study results. At a minimum, studies on DLD should present hearing status, and

scores on nonverbal IQ and standardized language tests to ascertain the clinical status of the participants. However, many articles are missing this information. Moreover, participant background, including caregiver's educational level, and the use of Chinese dialects and second languages in the child's environment, are influential in language development. Given the pervasiveness of bidialectalism and early English exposure in Mainland China, Hong Kong, and Taiwan, future studies should include information about the use of Mandarin and other dialects/languages in and outside the home to document the linguistic status of their participants. This information is currently missing in a majority of the Chinese DLD studies. Recruitment methods are also usually not reported, further affecting the replicability of the studies. Moreover, reliability of measurement and assumption checking of statistical tests are often missing. This type of information is essential to boost the readers' confidence in the study results and needs to be reported.

4.2. Bibliometric characteristics

Our searches yielded 59 empirical studies on the manifestations of DLD in Chinese children published between 1997 and July of 2022. A strength of the current review is that the "tower of babel bias" (Gregoire et al., 1995) was potentially reduced by retrieving pertinent work published beyond English. There is a clear and urgent need to raise awareness of DLD worldwide so that more research can be conducted to increase our understanding of the linguistic and cognitive manifestations, underlying neurobiological mechanisms, and social emotional consequences of this disorder. Although there is an upward trend in the growth of the literature in recent years, the output rate of Chinese DLD (33 articles from 2016-2022) is dwarfed by that of the English DLD literature (e.g., 90 empirical studies conducted in the United States alone between 2015 and 2019, Nitido & Plante, 2020).

When viewed by participant geographic region, the three regions each made important contributions to this literature. Hong Kong leads the Chinese DLD research considering the publication history, the number of publications, the breadth of topics (e.g., grammar, phonology, reading, statistical learning), and the expanded scope of investigations that compared DLD to other clinical groups such as DLD + dyslexia. Articles featuring participants from Mainland China, despite the shortest publication history, showed the most rapid growth. The majority of the articles were published in Chinese language journals and may reduce the language barriers faced by frontline clinicians who wish to access this literature. Finally, Taiwan has contributed the largest sample of children with DLD to this literature. Also, there appears to be a higher number of research groups publishing on this topic. Both observations suggest a greater awareness of DLD in research and clinical practice in Taiwan than in Mainland China.

4.3 Diagnostic criteria

With regard to DLD diagnostic criteria, most studies used one to two tests/subtests and 1-1.5 SDs below the mean as the cut score. These criteria are generally in line with what have been used in the English literature (Nitido & Plante, 2020; Spaulding et al., 2006). As for the actual tests used for diagnosing the study sample, studies on Cantonese DLD in Hong Kong showed the most uniformity, especially after the development of the HKCOLAS. The HKCOLAS has good (> 90%) sensitivity and specificity (T'sou et al., 2006). Another recently developed test "*The Hong Kong Test of Preschool Oral Language*" (TOPOL, Wong et al., 2019) also reported acceptable sensitivity (85%) and specificity (83%) in a sample of 54 preschoolers with DLD (aged 2;6 to 5;11) and 54 TD age- and gender-matched peers. These tests are sound options for researchers studying Cantonese-speaking children with DLD between the age of 2;6 and 12;1.

In studies conducted in Mainland China, the most popular test is the PPVT-R, a standardized test of receptive vocabulary normed on 600 children from Shanghai between 3.5 and 9 years of age (Sang & Miao, 1990). There are two potential issues with this choice. First, existing empirical evidence from English does not support the use of vocabulary tests for clinical screening and identification. Gray et al. (1999) examined the diagnostic accuracy of four vocabulary tests. They found that even though the DLD group scored lower than the TD control group on each test, the individual scores of children with DLD typically fell within the normal range. In other words, although these tests may accurately reflect a child's skills in a certain language domain (i.e., have good construct validity), they are not good identifiers of DLD. To know whether these concerns are true in Chinese, we would need direct evidence from diagnostic accuracy studies on these single-word vocabulary tests. Second, even if the diagnostic power of the test is not in question, the normative sample of the PPVT-R is not representative of current-day Mandarin speakers in Mainland China as the norm was published in 1990 and consists of children from Shanghai only. After the PPVT-R, the next popular test choice is the RDLS, a test of receptive and expressive language normed on 735 children from Taiwan (Lin et al., 2008a, 2008b). Again, diagnostic accuracy information is unavailable. While the norm could be representative of the population in Taiwan, its application to Mandarin speakers in Mainland China may be inapt. Lastly, 30% of the studies conducted in Mainland China relied on the verbal subtests of general IQ tests. While these tests may meet psychometric requirements for construct validity, their utility in diagnosing DLD is questionable. These concerns in test selection indicate a severe lag in test development. We note that two new tests are emerging for Mandarin-speaking children in Mainland China: the Mandarin Clinical Evaluation of Language for Preschooler's Core Scale (MCELP-CS, Wu et al., 2020), and the Diagnostic Assessment of

Mandarin SLI on Primary School Students (DAMSLIPSS, Xue et al., 2022). Wu et al. (2020) reported satisfactory diagnostic accuracy of the MCELP-CS in diagnosing language disorders in children with hearing impairment, autism, and cerebral palsy. These are welcoming additions and could to some extent ameliorate the test shortage. Nevertheless, the current uses and misuses of test suggest a need for researchers and practitioners to develop more sophisticated knowledge in evaluating tests, interpreting test scores, and using tests wisely to match the purpose of administration. A highly relevant introductory text on this topic can be found in Klee et al. (2009).

By contrast, researchers in Taiwan appeared to have more choices available in their toolkit, since as many as 10 tests were documented in our review. To the best of our knowledge, though many of these tests reported significant group level differences between impaired and typical groups in the test manuals, none of them reported diagnostic accuracy. Despite the fact that there are more test options, instances of using receptive vocabulary and general IQ tests also existed, suggesting a continuous need for diagnostic sophistication.

To summarize, as in the English literature (Nitido & Plante, 2020; Spaulding et al., 2006), we found variability in cut scores across studies, including when the same standardized tests were used. These variabilities could be because of a lack of information on the optimal cut score for a specific test, differences in opinion as for where to draw the line between typical and disordered language, or a deliberate choice to alter the criterion based on the characteristics of the sample.

4.4 Areas of deficits

The qualitative analysis revealed that Chinese children with DLD show wide-ranging deficits in basic auditory perception, phonological processing, vocabulary diversity, sentence

length and complexity, narrative content and organization, literacy, and word learning. At the same time, there were also areas wherein significant deficits were not consistently found: grammaticality and total number of utterances produced in narrative tasks did not clearly differentiate children with DLD from their TD peers (see Table 6).

How do these findings inform us on the nature of DLD? Does the literature provide consistent patterns on the behavioral traits of DLD that might serve as clinical markers for this disorder in Chinese? In English, one of the most established clinical markers of DLD is finite verb morphology (Archibald & Joanisse, 2009; Rice & Wexler, 1996). In Chinese, though tense and number agreement are not grammatically marked on the verb, aspect, or the temporal contour of the event, is marked on the verb via a small class of bound morphemes. The nine studies that examined aspect markers found reduced and less flexible use of aspect markers in the DLD group than age-matched controls but comparable use as younger MLU controls (e.g., Fletcher et al., 2005; Wong et al., 2003). Thus, verb morphosyntax appears to be disadvantaged in Chinese DLD. However, unlike in English wherein poorer performance in verb finiteness marking is reported relative to younger MLU-matched group (Rice et al., 2009), the comparable performance between DLD and MLU control groups suggests that verb aspect marking is not an area of extraordinary difficulty. It is important to note that finite verb marking is an accuracy-based measure because failures to use English verb finiteness markers result in the utterances becoming ungrammatical. By contrast, aspect marking in Chinese is rate-based because though aspect markers are preferred by mature speakers in certain contexts, these morphemes are rarely obligatory. This could make the difficulties with aspect marking harder to detect because one must know the range of performance TD children of the same age would demonstrate to the same stimuli.

Grammaticality of narrative language sample is an accuracy-based measure available in both Chinese and English. Crosslinguistic differences in diagnostic potential is observed for this measure. In English, percent grammatical utterance in narrative discourse was found to have the best sensitivity and specificity when pitted against two other discourse measures: finite verb morphology composite, which calculates the percent accurate production of verb tense, and clausal density, which measures children's ability in integrating multiple clauses into sentences (Guo et al., 2019, 2020, 2021). In Chinese, three studies have examined grammatical accuracy and they yielded mixed findings: Lai and Wang (2017) found a higher percentage of ungrammatical utterances (M = 35.4%) in their sample of school-age children with DLD (age range = 7;2 to 7;11, mean age = 7;6) in comparison to TD controls (M = 18.7%). Hao et al. (2018) and Sheng et al. (2020) included children of a wider age range (Hao et al: age range = 4;3 to 7;11, mean age = 6;2; Sheng et al: age range 4 to 6 years, mean age = 5;8). Both studies found low occurrence of ungrammatical utterances in their DLD samples (Ms < 7%); but in Sheng et al. (2020) the range was sizable (0-29%). This puzzling discrepancy calls for additional evidence related to the utility of this accuracy-based measure in identifying DLD in Chinese.

Phonological memory as assessed in nonword repetition tasks is also an established clinical marker in English and many other languages. Our qualitative analysis suggest that this may also be the case in Chinese as all but one of the nine studies found group differences. Sentence repetition is widely recognized as a clinical marker of DLD across many languages (e.g., Armon-Lotem et al., 2015; Conti-Ramsden et al., 2001). So far only two studies have examined sentence repetition in Chinese and converged on finding significant deficits in Cantonese and Mandarin-speaking children with DLD. The effect sizes were large to very large and ranged from .83 to 1.73 in Cantonese (Stokes et al., 2006, participants aged 4;2 to 5;7) and

3.25 to 3.47 in Mandarin (Wang et al., 2022, participants aged 4;0 to 5;11) when different scoring approaches were employed. Replication studies are needed for both languages, though the evidence thus far suggests that sentence repetition is a promising clinical marker in Chinese.

4.5 Areas of research needs

The current review revealed good breadth in the research topics, but there is still a clear need to enhance quality and build depth in the knowledge base for Chinese DLD. First, comprehension and production of complex syntax is a known area of deficit across languages and has received some attention in Chinese (e.g., relative clause). Continued attention in this area, including studies that focus on less-studied but highly frequent constructions (e.g., serial verb construction) and studies that use time-sensitive process-based measures (e.g., eye-tracking) is warranted. Second, several studies have elicited narrative samples to examine the functional communication abilities of children with DLD. This is a much-needed complement to decontextualized discrete-skill tasks because narrative more closely mimics daily communication demands and requires discourse level rather than utterance level planning. However, a widely accessible large-scale narrative corpus by children with and without DLD remains a gap that needs to be filled for Chinese. This line of research will also need to tackle methodological hurdles and stimulate transparency and consensus in basic conventions regarding the segmentation of utterances and words in Chinese (Sheng et al., 2020). Third, studies of the learning processes of children with DLD using word learning, statistical learning, artificial grammar learning paradigms, studies that delineate the cognitive profiles of DLD including working memory and executive functions, and studies that examine the neurological underpinning of speech and language processing in DLD are rare and warrant more attention in future investigations. Fourth, the study of real-time semantic processing and sentence processing

is missing in the Chinese DLD literature and could be a fruitful avenue for future research. Fifth, the majority of the studies focused on preschool age children and almost all studies measured performance at one time point. But DLD is a lifelong disability that does not spontaneously resolve with age. Studies of older participants (secondary school students, young adults) and studies of a longitudinal nature should be conducted to build an evidence base on the language, academic, social-emotional, and occupational outcomes of these individuals as they mature.

The ultimate goal for researchers who study clinical populations is to generate a high-quality evidence base to support effective clinical practice. The Chinese DLD literature is also in need of studies that more directly inform assessment and treatment practice. Examples of clinical practice research can be found in Wang et al. (2022)'s diagnostic accuracy study of Mandarin sentence repetition and To et al. (2015)'s treatment study of Cantonese grammar.

4.6 Limitations

This review excluded studies that are theses, dissertations, conference proceedings, and book chapters. Thus, we have inevitably missed some studies on this topic, including ones with less significant results (Conn et al., 2003). In addition, it is beyond the scope of the current study to review and summarize the theoretical perspectives of the literature. This worthy goal should be pursued in future reviews. Also, we stopped short of conducting a meta-analysis in light of the quality appraisal results and the heterogeneity of outcomes studied in this literature. It is our hope that a meta-analysis could be pursued in the future with the continuous accrual of high-quality research evidence.

5. Conclusion

A systematic review of peer-reviewed publications found 59 studies that compared Chinese children with DLD with typically-developing peers on speech, language, literacy, and

cognitive outcomes. These articles are published in both English language and Chinese language journals. While issues with methodological and reporting rigor are present in this young literature, the studies contain valuable information for researchers who wish to study this population and clinicians who serve these individuals. Findings from the retrieved articles corroborate our existing understanding of DLD as a significant and heterogeneous disorder that poses great challenges to language learning. A qualitative summary of the literature suggests that nonword repetition and sentence repetition are potential clinical markers of Chinese DLD. Lastly, features of the Chinese languages may potentially shift the utility of accuracy-based versus rate-based measures of expressive morphosyntax in identifying DLD.

Reference

Archibald, L. M., & Joanisse, M. F. (2009). On the sensitivity and specificity of nonword repetition and sentence recall to language and memory impairments in children. *Journal of speech, language, and hearing research: JSLHR*, *52*(4), 899-914. https://doi.org/10.1044/1092-4388(2009/08-0099)

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

Bishop, D. V. M. (1998). Development of the Children's Communication Checklist (CCC): A Method for Assessing Qualitative Aspects of Communicative Impairment in Children. *Journal of child psychology and psychiatry*, *39*(6), 879-891. https://doi.org/https://doi.org/10.1111/1469-7610.00388

Bishop, D. V. M., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & consortium, a. t. C.-. (2017). Phase 2 of CATALISE: a multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology. *Journal of child psychology and psychiatry*, *58*(10), 1068-1080. https://doi.org/https://doi.org/10.1111/jcpp.12721

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons.

Chen, L., An, S., & He, X. (2021). A comparative analysis of the production of aspect markers by Mandarin-speaking children with developmental language disorders and by their typically developing peers. *Journal of Child Language Acquisition and Development: JCLAD*, *9*(1), 189-208. https://science-res.com/index.php/jclad/article/view/24

Chen, L.-Y., & Liu, H.-M. (2010). 學齡期特定型語言障礙兒童聽知覺區辨能力初探 [Auditory Processing in School-Aged Children with Specific Language Impairments]. *特殊教育研究學刊 [Bulletin of Special Education]*, *35*(1), 1-18. https://doi.org/10.6172/BSE201003.3501001

Chen, R. (1997). *魏氏兒童智力量表第三版（WISC-III）中文版 [The Wechsler Intelligence Scale for Children - Third Edition, Chinese version]*. Chinese Behavioral Science Corporation.

Chen, X., & Chen, R. (2013). *魏氏幼兒智力量表第四版（WPSSI-IV）中文版指導手冊[Administration and scoring manual of Wechsler preschool and primary scale of intelligence-fourth edition (Chinese version)]*. Chinese Behavioral Science Corporation.

Chen, Y., & Liu, H.-M. (2018). 華語學齡前特定型語言障礙兒童的非詞複誦表現與臨床診斷之應用 [Nonword Repetition Performance in Mandarin-Speaking Preschool Children with Specific Language Impairment—A Study of Diagnostic Accuracy]. *特殊教育研究學刊 [Bulletin of Special Education]*, *43*(3), 57-81. https://doi.org/10.6172/BSE.201811_43(3).0003

Chen, Y., & Liu, H. M. (2014). Novel-word learning deficits in Mandarin-speaking preschool children with specific language impairments. *Research in developmental disabilities*, *35*(1), 10-20. https://doi.org/10.1016/j.ridd.2013.10.010

Cheng, Y. Y., Wu, H. C., Shih, H. Y., Yeh, P. W., Yen, H. L., & Lee, C. Y. (2021). Deficits in Processing of Lexical Tones in Mandarin-Speaking Children With Developmental Language Disorder: Electrophysiological Evidence. *Journal of speech, language, and hearing research: JSLHR*, *64*(4), 1176-1188. https://doi.org/10.1044/2021_JSLHR-19-00392

Cheng, R. L. (1985). A Comparison of Taiwanese, Taiwan Mandarin, and Peking Mandarin. *Language*, *61*(2), 352–377. https://doi.org/10.2307/414149

Cheung, H. (2003). Memory capacity in school-age Mandarin-speaking children with specific language impairment. *Asia Pacific Journal of Speech, Language and Hearing*, *8*(3), 160-164. https://doi.org/10.1179/136132803805576156

Chi, P.-H. (2007). 特定型語言障礙兒童音韻短期記憶能力之初探 [Phonological Short-Term Memory in Children with Specific Language Impairment]. *特殊教育研究學刊 [Bulletin of Special Education]*, *32*(4), 19-45. https://doi.org/ 10.6172/BSE200712.3204002

Chi, P.-H. (2008). 特定型語言障礙檢核表之編製 [Development of the Specific Language Impairment Checklist]. *測驗學刊 [Psychological Testing]*, *55*(2), 247-286. https://doi.org/10.7108/PT.200808.0247

Chi, P.-H., Zhang, X.-Z., & Hung, S.-T. (2012). 學前特定型語言障礙兒童進入小學的追蹤研究：語言、識字表現之探討 [A Longitudinal Investigation of Language and Reading Outcomes in Mandarin Chinese-Speaking Children with Specific Language Impairments]. *特殊教育學報 [Journal of Special Education]*, *36*, 61-91. https://doi.org/10.6768/JSE.201212.0061

Conn, V. S., Valentine, J. C., Cooper, H. M., & Rantz, M. J. (2003). Grey Literature in Meta-Analyses. *Nursing Research*, *52*(4), 256-261. https://journals.lww.com/nursingresearchonline/Fulltext/2003/07000/Grey_Literature_in_Meta_Analyses.8.aspx

Conti-Ramsden, G., Botting, N., & Faragher, B. (2001). Psycholinguistic markers for specific language impairment (SLI). *Journal of child psychology and psychiatry*, *42*(6), 741-748.

Compulsory Education Law of the People's Republic of China. 中华人民共和国义务教育法_法律法规_中国政府网 (www.gov.cn)

Dai, H., & He, X. (2021). 特殊型语言障碍儿童与高功能自闭症儿童对汉语特殊疑问句的理解 [Comprehension of Wh-questions by Chinese Children with Specific Language Impairment and Those with High-Functioning Autism]. *现代外语（双月刊）[Modern Foreign Languages (Bimonthly)]*, *44*(4), 522-535. https://kns.cnki.net/kcms/detail/detail.aspx?FileName=XDWY202104009&DbName=DKFX2021

Directorate General of Budget, Accounting and Statistics, Executive Yuan. (2020). *109 年人口及住宅普查初步統計結果提要分析* [*Preliminary Results of Population and Housing Census of the Year 109*]. https://www.stat.gov.tw/public/Attachment/1112143117MKFOK1MR.pdf

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455-463. https://doi.org/10.1111/j.0006-341x.2000.00455.x

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj*, *315*(7109), 629-634.

Fletcher, P., Leonard, L., Stokes, S. F., & Wong, A. (2009). Morphosyntactic deficits in Cantonese-speaking children with specific language impairment. In S.-P. Law, B. S. Weekes, & A. M.-Y. Wong (Eds.), *Language disorders in speakers of Chinese* (pp. 75-88). Multilingual Matters.

Fletcher, P., Leonard, L. B., Stokes, S. F., & Wong, A. M.-Y. (2005). The expression of aspect in Cantonese-speaking children with specific language impairment. *Journal of speech, language, and hearing research: JSLHR*, *48*(3), 621-634. https://doi.org/10.1044/1092-4388(2005/043)

Fung, R. S.-Y. (2009). Characteristics of Chinese in relation to language disorders. In S.-P. Law, B. S. Weekes, & A. -M.-Y. Wong (Eds.), *Language disorders in speakers of Chinese* (pp. 1–18). Multi-lingual Matters.

Gong, Y., & Cai, T. (1994). 中国修订韦氏儿童智力量表 [The Wechsler Intelligence Scale for Children Revised in China (C-MISC)]. *中国临床心理学杂志 [Chinese Journal of Clinical Psychology]*, *2*(1), 1-6.

Gray, S., Plante, E., Vance, R., & Henrichsen, M. (1999). The diagnostic accuracy of four vocabulary tests administered to preschool-age children. *Language, Speech, and Hearing Services in Schools*, *30*(2), 196-206.

Grégoire, G., Derderian, F., & Le Lorier, J. (1995). Selecting the language of the publications included in a meta-analysis: Is there a tower of babel bias? *Journal of Clinical Epidemiology*, *48*(1), 159-163. https://doi.org/https://doi.org/10.1016/0895-4356(94)00098-B

Guo, L.-Y., Eisenberg, S., Schneider, P., & Spencer, L. (2019). Percent Grammatical Utterances Between 4 and 9 Years of Age for the Edmonton Narrative Norms Instrument: Reference Data and Psychometric Properties. *American Journal of Speech-Language Pathology*, *28*(4), 1448-1462. https://doi.org/doi:10.1044/2019_AJSLP-18-0228

Guo, L.-Y., Eisenberg, S., Schneider, P., & Spencer, L. (2020). Finite Verb Morphology Composite Between Age 4 and Age 9 for the Edmonton Narrative Norms Instrument: Reference Data and Psychometric Properties. *Language, Speech, and Hearing Services in Schools*, *51*(1), 128-143. https://doi.org/doi:10.1044/2019_LSHSS-19-0028

Guo, L.-Y., Schneider, P., & Harrison, W. (2021). Clausal Density Between Ages 4 and 9 Years for the Edmonton Narrative Norms Instrument: Reference Data and Psychometric Properties. *Language, Speech, and Hearing Services in Schools*, *52*(1), 354-368. https://doi.org/doi:10.1044/2020_LSHSS-20-00043

Hao, Y., Sheng, L., Zhang, Y., Jiang, F., de Villiers, J., Lee, W., & Liu, X. L. (2018). A Narrative Evaluation of Mandarin-Speaking Children With Language Impairment. *Journal of speech, language, and hearing research: JSLHR*, *61*(2), 345-359. https://doi.org/10.1044/2017_JSLHR-L-16-0367

He, X. (2020). 国外特殊型语言障碍的语言学研究及思考 [A Review of Overseas Linguistic Studies on Specific Language Impairment and Some Reflections]. *语言战略研究 [Chinese Journal of Language Policy and Planning]*, *26*(2), 61-70. https://doi.org/10.19689/j.cnki.cn10-1361/h.20200205

He, X., & Dai, H. (2012). 普通话特殊型语言障碍儿童否定结构习得研究 [The Acquisition of Negation in Mandarin-speaking Children with SLI] *广东外语外贸大学学报 [Journal of Guangdong University of Foreign Studies] 4*, 16-20. https://kns.cnki.net/kcms/detail/detail.aspx?FileName=GDWY201204003&DbName=CJFQ2012

He, X., Sun, L., & Tian, L. (2013). 汉语特殊型语言障碍儿童体标记“了”和“在”的产出研究 [The Production of Aspect Markers “Le” and “Zai” in Chinese-speaking Children with Specific Language Impairment]. *外语教学 [Foreign Language Education]*, *34*(3), 27-32. https://doi.org/10.16362/j.cnki.cn61-1023/h.2013.02.012

He, X., & Yu, H. (2013). 汉语特殊型语言障碍儿童关系从句理解研究 [The grammar impairment of Mandarin Chinese SLI children: Evidence from relative clause comprehension]. *现代外语（季刊）[Modern Foreign Languages (Quarterly)]*, *36*(4), 340-346+437. https://kns.cnki.net/kcms/detail/detail.aspx?FileName=XDWY201304002&DbName=CJFQ2013

Hong Kong Society for Child Health and Development (1987). *Reynell Development of Language Scales, Cantonese (Hong Kong) version, Manual*.

Huang, R. J., Jian, S. Y., Jhu, L. S., & Lu, L. (2011). *Receptive and expressive vocabulary test (REVT)*. Psychological Publishing.

Huang, R.-Z., Cai, Yun-Chun, Lin, Jia-Rong, Zhang, Yi-Yu, & Wang, Yi-Qun (2014). *華語學齡兒童溝通及語言能力測驗 [Test of communication and language ability for school-age children in Chinese]*. 心理出版社.

Iao, L. S., Ng, L. Y., Wong, A. M., & Lee, O. T. (2017). Nonadjacent Dependency Learning in Cantonese-Speaking Children With and Without a History of Specific Language Impairment. *Journal of speech, language, and hearing research: JSLHR*, *60*(3), 694-700. https://doi.org/10.1044/2016_JSLHR-L-15-0232

Kidd, E., & Garcia, R. (2022). How diverse is child language acquisition research? *First Language*, *42*(6), 703–735. https://doi.org/10.1177/01427237211066405

Kidd, J. C., Shum, K. K., Wong, A. M., Ho, C. S., & Au, T. K. (2017). Auditory perception and word recognition in Cantonese-Chinese speaking children with and without Specific Language Impairment. *Journal of child language*, *44*(1), 1-35. https://doi.org/10.1017/S0305000915000604

Kim, H.-k., Wang, Y.-x., & Shen, S. (2019). 发声复述及延迟时间对特定型语言障碍儿童新词命名学习的影响 [Impacts of Vocal Rehearsal and Retention Duration on Novel Word Naming in Children with Specific Language Impairment]. *中国康复理论与实践 [Chinese Journal of Rehabilitation Theory and Practice]*, *25*(4), 448-454. https://doi.org/10.3969/j.issn.1006-9771.2019.04.015

Klee, T., Stokes, S. F., Wong, A. M. Y., Fletcher, P., & Gavin, W. J. (2004). Utterance length and lexical diversity in Cantonese-speaking children with and without specific language impairment. *Journal of speech, language, and hearing research: JSLHR*, *47*(6), 1396-1410. https://doi.org/10.1044/1092-4388(2004/104)

Klee, T., Wong, A., Stokes, S., Fletcher, P., & Leonard, L. (2009). Assessing Cantonese-speaking Children with Language Difficulties from the Perspective of Evidence-Based Practice: Current Practice and Future Directions. In S.-P. Law, B. S. Weekes, & A. M.-Y. Wong (Eds.), *Language disorders in speakers of Chinese* (pp. 89-111). Multilingual Matters. https://doi.org/10.21832/9781847691170-008

Kmet, L. M., Cook, L. S., & Lee, R. C. (2004). *Standard quality assessment criteria for evaluating primary research papers from a variety of fields*.

Labov, W. (1972). *Language in the inner city: Studies in the Black English vernacular*. University of Pennsylvania Press.

Lai, H.-C., & Wang, C.-L. (2017). 有效鑑別國小二年級語言障礙兒童和典型發展兒童的敘事指標 [Narrative Indicators for Effectively Distinguishing between Second-Grade Students with Language Disorders and their Typically Developing Peers]. *特殊教育研究學刊 [Bulletin of Special Education]*, *42*(3), 33-62. https://doi.org/10.6172/BSE.201711_42(3).0002

Lee, K. Y. S., Lee, L. W. T., & Cheung, P. S. P. (1996). *Hong Kong Cantonese Receptive Vocabulary Test*. Hong Kong: Child Assessment Services.

Leonard, L. B., Wong, A. M., Deevy, P., Stokes, S. F., & Fletcher, P. (2006). The Production of Passives by Children with Specific Language Impairment Acquiring English or Cantonese. *Applied psycholinguistics*, *27*(2), 267-299. https://doi.org/10.1017/S0142716406060280

Leonard, L. B. (2014). *Children with specific language impairment*. MIT press.

Leonard, L. B. (2020). A 200-Year History of the Study of Childhood Language Disorders of Unknown Origin: Changes in Terminology. *Perspectives of the ASHA Special Interest Groups*, *5*(1), 6-11. https://doi.org/10.1044/2019_pers-sig1-2019-0007

Leonard, L. B., Deevy, P., Wong, A. M., Stokes, S. F., & Fletcher, P. (2007). Modal verbs with and without tense: a study of English- and Cantonese-speaking children with specific language impairment.

International journal of language & communication disorders, *42*(2), 209-228. https://doi.org/10.1080/13682820600624240

Li, C. N., & Thompson, S. A. (1989). *Mandarin Chinese: A functional reference grammar* (Vol. 3). Univ of California Press.

Lin, B., & Chi, P.-H. (2002). *兒童口語理解測驗 [Children's Oral Language Comprehension Test]*. 國立臺灣師範大學特殊教育學系.

Lin, B., Huang, Y., Huang, G., & Xuan, C. (2008a). *修訂學前兒童語言障礙評量表 [Revised Language Disorder Scale for Preschool Children]*. 臺灣師範大學特殊教育中心.

Lin, B., Huang, Y., Huang, G., & Xuan, C. (2008b). *修訂學齡兒童語言障礙評量表 [Revised Language Disorder Scale for School-Age Children]*. 臺灣師範大學特殊教育中心.

Lin, B., Huang, Y., Huang, G., & Xuan, C. (2018b). *修訂學齡兒童語言障礙評量表 [Revised Language Disorder Scale for School-Age Children]*. 臺灣師範大學特殊教育中心.

Lin, B. G., & Lin, M. S. (2007). *Language Disorder Scale of Preschoolers–Revised*. National Taiwan Normal University.

Lin, Y., & Zhang, Y. (2021). 学龄前语言障碍儿童的执行功能特征研究 [Characteristics of executive function in preschoolers with developmental language disorder]. *教育生物学杂志 [Journal of Bio-education]*, *9*(3), 179-183. https://doi.org/10.3969/j.issn.2095-4301.2021.03.003

Liu, H. M., & Chien, H. Y. (2020). Speech sound errors of Mandarin-speaking preschool children with co-occurring speech sound disorder and developmental language disorder. *Clinical linguistics & phonetics*, *34*(12), 1130-1148. https://doi.org/10.1080/02699206.2020.1724334

Liu, X. L., de Villiers, J., Ning, C., Rolfhus, E., Hutchings, T., Lee, W., Jiang, F., & Zhang, Y. W. (2017). Research to Establish the Validity, Reliability, and Clinical Utility of a Comprehensive Language Assessment of Mandarin. *Journal of speech, language, and hearing research: JSLHR*, *60*(3), 592-606. https://doi.org/10.1044/2016_jslhr-l-15-0334

Lu, L., & Liu, H.-H. (1998). *修訂畢保德圖畫詞彙測驗 [The Peabody Picture Vocabulary Test: Revised in Chinese]*. Psychological Publishing.

Lv, H.-H., & Tsao, F.-M. (2018). 遲語兒幼兒期至學齡前期的語言發展型態：兩年縱貫研究 [Late-talking Children's Language Development Pattern in the Early Childhood: A Longitudinal Study for Two Years]. *教育心理學報 [Bulletin of Educational Psychology]*, *49*(4), 611-636. https://doi.org/10.6251//BE P.201806_49(4).0005

Ma, X., & Liang, D. (2019). 汉语SLI 儿童快速映射词语学习缺陷的原因: 语音记忆还是语义表征? [The Reasons for the Deficiencies of Chinese Children with Specific Language Impairment in Fast Mapping Word Learning: Phonological Working Memory or Semantic Representation]. *当代语言学 [Contemporary Linguistics] 21*(1), 104-118. https://kns.cnki.net/kcms/detail/detail.aspx?FileName=DDYX201901006&DbName=CJFQ2019

Matthews, S., & Yip, V. (1994). *Cantonese: a comprehensive grammar*. Routledge.

McGregor, K. K., Goffman, L., Van Horne, A. O., Hogan, T. P., & Finestack, L. H. (2020). Developmental Language Disorder: Applications for Advocacy, Research, and Clinical Service. *Perspectives of the ASHA Special Interest Groups*, *5*(1), 38-46. https://doi.org/10.1044/2019_persp-19-00083

Moola S, Munn Z, Tufanaru C, Aromataris E, Sears K, Sfetcu R, Currie M, Qureshi R, Mattis P, Lisy K, Mu P-F. (2020). Chapter 7: Systematic reviews of etiology and risk . In: Aromataris E, Munn Z (Editors). *JBI Manual for Evidence Synthesis.* JBI. Available from https://synthesismanual.jbi.global. https://doi.org/10.46658/JBIMES-20-08

MSCA-CR课题组 [The team of MSCA-CR]. (1991). 麦卡锡幼儿智能量表 (MSCA) 简介及其在上海试测结果报告 [The revision of trait norm about McCarthy Scale (MSCA) of Children's Abilities in Shanghai proper]. *心理科学*, *5*, 32-36.

Murza, K. A., & Ehren, B. J. (2020). Considering the Language Disorder Label Debate From a School Speech-Language Pathology Lens. *Perspectives of the ASHA Special Interest Groups*, *5*(1), 47-54. https://doi.org/10.1044/2019_persp-19-00077

Ning, C. Y. (2012). *The criterion-referenced diagnostic test of Mandarin-speaking preschool children with SLI*. Tianjin University Press.

Ning, C. Y. (2013). *The language assessment for Mandarin-speaking preschool children*. Tianjin University Press.

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

Ministry of Education of the People's Republic of China. (2021). *中国语言文字概况 [Overview of Chinese languages]*.http://www.moe.gov.cn/jyb_sjzl/wenzi/202108/t20210827_554992.html

Nitido, H., & Plante, E. (2020). Diagnosis of Developmental Language Disorder in Research Studies. *Journal of speech, language, and hearing research: JSLHR*, *63*(8), 2777-2788. https://doi.org/10.1044/2020_JSLHR-20-00091

Norbury, C. F., Gooch, D., Wray, C., Baird, G., Charman, T., Simonoff, E., Vamvakas, G., & 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, 57*(11), 1247–1257. https://doi.org/10.1111/jcpp.12573

Peterson, C., & McCabe, A. (1983). Three ways of looking at a child's narrative: A psycholinguistic analysis. *New York: Plenum. ROSCH, E.(1975). Cognitive representations of semantic categories. Journal of Experimental Psychology*, *104*, 192-233.

Plante, E., & Vance, R. (1994). Selection of preschool language tests: A data-based approach. *Language, Speech, and Hearing Services in Schools*, *25*(1), 15-24.

Reetzke, R., Zou, X., Sheng, L., & Katsos, N. (2015). Communicative Development in Bilingually Exposed Chinese Children With Autism Spectrum Disorders. *Journal of Speech, Language, and Hearing Research.*, *58*(3), 813–825. https://doi.org/10.1044/2015_JSLHR-L-13-0258

Reilly, S., Tomblin, B., Law, J., McKean, C., Mensah, F. K., Morgan, A., Goldfeld, S., Nicholson, J. M., & Wake, M. (2014). Specific language impairment: a convenient label for whom? *Int J Lang Commun Disord*, *49*(4), 416-451. https://doi.org/10.1111/1460-6984.12102

Rice, M. L., Hoffman, L., & Wexler, K. (2009). Judgements of omitted BE and DO in questions as extended finiteness clinical markers of specific language impairment (SLI) to 15 years: A study of growth and asymptote. *Journal of speech, language, and hearing research*, *52*(6), 1417-1433.

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), 1239-1257.

Sang, B., & Miao, X. (1996). 皮博迪图片词汇测验修订版（PPVT-R）上海市区试用常模的修订 [The Revision of Trail Norm of Peabody Picture Vocabulary Test Revised (PPVT-R) in Shanghai Proper]. *心理科学通讯 [Information on Psychological Sciences]*(5), 22-27+67+65-66. https://doi.org/10.16719/j.cnki.1671-6981.1990.05.004

Sheng, L., Shi, H., Wang, D., Hao, Y., & Zheng, L. (2020). Narrative Production in Mandarin-Speaking Children: Effects of Language Ability and Elicitation Method. *Journal of speech, language, and hearing research: JSLHR*, *63*(3), 774-792. https://doi.org/10.1044/2019_JSLHR-19-00087

Siu, E., & Man, D. W. K. (2006). Working memory and sentence comprehension of Hong Kong Chinese children with specific language impairment. *International journal of rehabilitation research*, *29*, 267–269. https://doi.org/10.1097/01.mrr.0000210058.06989.f6

Spaulding, T. J., Plante, E., & Farinella, K. A. (2006). Eligibility criteria for language impairment: Is the Low End of Normal Always Appropriate?. *Language, Speech, and Hearing Services in Schools*, *37*(1), 61-72. https://doi.org/10.1044/0161-1461(2006/007)

Stein, N. L., & Glenn, C. G. (1979). An analysis of story comprehension in elementary school children. In R. O. Freedle (Ed.), *Advances in discourse processes: New directions in discourse processing (Vol. 2)*. NJ: Ablex.

Stokes, S. F., Wong, A. M., Fletcher, P., & Leonard, L. B. (2006). Nonword repetition and sentence repetition as clinical markers of specific language impairment: the case of Cantonese. *Journal of speech, language, and hearing research: JSLHR*, *49*(2), 219–236. https://doi.org/10.1044/1092-4388(2006/019)

Stokes, S. F., & Fletcher, P. (2000). Lexical diversity and productivity in Cantonese-speaking children with specific language impairment. *International journal of language & communication disorders*, *35*(4), 527-541. https://doi.org/10.1080/136828200750001278

Stokes, S. F., & Fletcher, P. (2003). Aspectual forms in Cantonese children with specific language impairment. *Linguistics*, *41*(2), 381-405. https://doi.org/10.1515/ling.2003.013

Stokes, S. F., & So, L. K. H. (1997). Classifier use by language-disordered and age-matched Cantonese-speaking children. *Asia Pacific Journal of Speech, Language and Hearing*, *2*(2), 83-101. https://doi.org/10.1179/136132897805577413

T'sou, B., Lee, T., Tung, P., Chan, A., Man, Y., & To, C. (2006). *Hong Kong Cantonese Oral Language Assessment Scale*. Hong Kong: Language Information Sciences Research Centre, City University of Hong Kong.

To, C., Lui, H., Li, X. X., & Lam, G. Y. H. (2015). A Randomized Controlled Trial of Two Syntactic Treatment Procedures With Cantonese-Speaking, School-Age Children With Language Disorders. *Journal of Speech Language and Hearing Research*, *58*. https://doi.org/10.1044/2015_JSLHR-L-14-0140

To, C. K., Stokes, S. F., Cheung, H. T., & T'sou, B. (2010). Narrative assessment for Cantonese-speaking children. *Journal of speech, language, and hearing research: JSLHR*, *53*(3), 648–669. https://doi.org/10.1044/1092-4388(2009/08-0039)

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), 1245-1260.

Torng, P. C., & Sah, W. H. (2020). Narrative abilities of Mandarin-speaking children with and without specific language impairment: macrostructure and microstructure. *Clinical linguistics & phonetics*, *34*(5), 453-478. https://doi.org/10.1080/02699206.2019.1655097

Tsai, W., & Chang, C.-j. (2008). "But I first... and then he kept picking" Narrative skill in Mandarin-speaking children with language impairment. *Narrative Inquiry*, *18*(2), 349-377. https://doi.org/10.1075/ni.18.1.09tsa

Tseng, Y.-H., & Liu, H.-M. (2017). 國小高年級語言障礙學生口語說明與對話能力之初探 [Examining Performance of Expository and Conversational Discourse in Mandarin-Speaking Children with Language Impairment]. *特殊教育學報 [Journal of Special Education]*, *46*, 1-30. https://doi.org/10.6768/JSE

Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software*, *36*(3), 1 - 48. https://doi.org/10.18637/jss.v036.i03

von Hippel, P. T. (2015). The heterogeneity statistic I2 can be biased in small meta-analyses. *BMC Medical Research Methodology*, *15*(1), 1-8. https://doi.org/10.1186/s12874-015-0024-z

Wang, D., Zheng, L., Lin, Y., Zhang, Y., & Sheng, L. (2022). Sentence Repetition as a Clinical Marker for Mandarin-Speaking Preschoolers With Developmental Language Disorder. *Journal of speech, language, and hearing research: JSLHR*, *65*(4), 1543-1560. https://doi.org/10.1044/2021_JSLHR-21-00401

Wang, H., & Yu, H. (2021). The Production of Relatives in Mandarin Children With Specific Language Impairment–From the Perspective of Edge Feature Underspecification Hypothesis [Original Research]. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.705526

Wang, X., & Huang, J. (2016). Non-Word Repetition Performance in Mandarin-Speaking Preschool Children With and Without Specific Language Impairment (SLI). *Chinese Journal of Applied Linguistics (Quarterly)*, *39*(3), 337-352+375. https://doi.org/10.1515/cjal-2016-0022

Wang, Y.-X., Kim, H.-K., Zhou, W.-P., Chao, T.-T., & Yang, Y. (2020). 特定型语言障碍儿童新异名词快速附带性学习能力研究 [Quick Incidental Novel Words Learning Ability Study in Children with Specific Language Impairment]. *中国听力语言康复科学杂志 [Chinese Scientific Journal of Hearing and Speech Rehabilitation]*, *93*, 120-124. https://doi.org/10.3969/j.issn.1672-4933.2019.02.011

Wong, A. (2023). *Understanding development and disorder in Cantonese using language sample analysis.* Routledge: New York, USA

Wong, A., Stokes, S., & Fletcher, P. (2003). Collocational diversity in perfective aspect zo2 use in Cantonese children with SLI. *Journal of Multilingual Communication Disorders*, *1*(2), 132-140. https://doi.org/10.1080/1476967031000090962

Wong, A. M., Leonard, L. B., Fletcher, P., & Stokes, S. F. (2004). Questions without movement: A study of Cantonese-speaking children with and without specific language impairment. *Journal of speech, language, and hearing research: JSLHR*, *47*(6), 1440–1453. https://doi.org/10.1044/1092-4388(2004/107)

Wong, A. M., Ciocca, V., & Yung, S. (2009). The perception of lexical tone contrasts in Cantonese children with and without specific language impairment (SLI). *Journal of speech, language, and hearing research: JSLHR*, *52*(6), 1493–1509. https://doi.org/10.1044/1092-4388(2009/08-0170)

Wong, A. M., Klee, T., Stokes, S. F., Fletcher, P., & Leonard, L. B. (2010b). Differentiating Cantonese-speaking preschool children with and without SLI using MLU and lexical diversity (D). *Journal of speech, language, and hearing research: JSLHR*, *53*(3), 794–799. https://doi.org/10.1044/1092-4388(2009/08-0195)

Wong, A.M.Y, Leung, C, Ng, A.K.H, Cheung P.S.P., Siu, E.K.L., To, K.S., Sam, S.K.L., Cheung, H.T., Lo, S.K., Lam, C.C.C.S. (2019). Technical Manual of the Hong Kong Test of Preschool Oral Language (Cantonese) (TOPOL). Hong Kong: Department of Health, Hong Kong SAR Government. http://hdl.handle.net/10722/261022

Wong, A. M. Y., Ho, C. S. H., Au, T. K. F., Kidd, J. C., Ng, A. K. H., Yip, L. P. W., & Lam, C. C. C. (2015). (Dis)connections between specific language impairment and dyslexia in Chinese. *Reading and Writing*, *28*(5), 699-719. https://doi.org/10.1007/s11145-015-9546-3

Wong, A. M. Y., Ho, C. S. H., Au, T. K. F., McBride, C., Ng, A. K. H., Yip, L. P. W., & Lam, C. C. C. (2017). Reading comprehension, working memory and higher-level language skills in children with SLI and/or dyslexia. *Reading and Writing*, *30*(2), 337-361. https://doi.org/10.1007/s11145-016-9678-0

Wong, A. M. Y., Kidd, J. C., Ho, C. S. H., & Au, T. K. F. (2010a). Characterizing the Overlap Between SLI and Dyslexia in Chinese: The Role of Phonology and Beyond. *Scientific Studies of Reading*, *14*(1), 30-57. https://doi.org/10.1080/10888430903242043

Wu, H., Liu, Q., Yu, B., Zhang, Y., Ren, D., Serdarevic, M., Liang, Z., Wang, Y., Chen, S., Zhang, K., & Chen, S. (2020). Psychometric properties of the mandarin clinical evaluation of language for preschooler's core scale. *Journal of Communication Disorders*, *87*, 106000. https://doi.org/https://doi.org/10.1016/j.jcomdis.2020.106000

Wu, S.-Y. (2020). Mean length of utterance among Mandarin-speaking children with and without DLD. *Child Language Teaching and Therapy*, *36*(3), 165-179. https://doi.org/10.1177/0265659020945366

Wu, S. Y., Huang, R. J., & Tsai, I. F. (2019). The applicability of D, MTLD, and MATTR in Mandarin-speaking children. *Journal of Communication Disorders*, *77*, 71-79. https://doi.org/10.1016/j.jcomdis.2018.10.002

Xu, L., He, X., Xu, J., & Su, L. (2020). 汉语特殊型语言障碍儿童量词习得研究 [The Acquisition of Classifiers in Mandarin-speaking Children with Specific Language Impairment]. *外语教学 [Foreign Language Education]*, *41*(4), 63-68. https://kns.cnki.net/kcms/detail/detail.aspx?FileName=TEAC202004012&DbName=CJFQ2020

Xue, J., Zhuo, J., Li, P., Liu, J., & Zhao, J. (2022). Characterizing macro- and micro-structures of narrative skills for Mandarin-speaking school-age children with specific language impairment. *Journal of Communication Disorders*, *96*, 106199. https://doi.org/10.1016/j.jcomdis.2022.106199

Yu, H. (2016). The Grammar Impairment of Mandarin Chinese SLI Children: Evidence from Topic-comment Structures. *Journal of Language Teaching and Research*, *7*(2). https://doi.org/10.17507/jltr.0702.08

Yu, H., Du, J., & He, X. (2022). 汉语特殊型语言障碍儿童被动句理解研究 [Comprehension of Passives by Mandarin Children with Specific Language Impairment]. *现代外语（双月刊）[Modern Foreign Languages (Bimonthly)]*, *45*(1), 53-65. https://kns.cnki.net/kcms/detail/detail.aspx?FileName=TEAC202004012&DbName=CJFQ2020

Yu, H., He, X., & Wang, H. (2017). 普通话特殊型语言障碍儿童关系从句产出研究 [The production of relative clauses by Mandarin preschool children with specific language impairment]. *现代外语（双月刊）[Modern Foreign Languages (Bimonthly)]*, *40*(4), 495-506+584. https://kns.cnki.net/kcms/detail/detail.aspx?FileName=XDWY201704006&DbName=CJFQ2017

Yu, H., Wang, H., & He, X. (2021). The comprehension of relative clauses in Mandarin Children with suspected specific language impairment. *Journal of child language*, 1-26. https://doi.org/10.1017/S0305000921000660

Yu, J., Wang, Q., & Liang, D. (2019). 学龄期普通话特异性语言损伤儿童体习得——来自词汇体和语法体组合加工的证据 [The Acquisition of Aspect in School-age Mandarin-specific Language Impairment Children: Evidence from Processing the Combination of Lexical and Grammatical Aspects]. *心理发展与教育 [Psychological Development and Education]*, *35*(2), 203-209.

Zeng, T., Li, H., Li, K., & He, X. (2013). 汉语普通话特殊语言障碍儿童“把” 字句习得研究 [On Acquisition of ba (把) -construction in Mandarin-speaking Children with Specific Language Impairment]. *华文教育与研究 [TCSOL Studies]*, *1*, 10-18. https://doi.org/10.16131/j.cnki.cn44-1669/g4.2013.01.001

Zeng, T., Zhu, T., Li, X., & Zhu, R. (2018). Passive Structure Features in Mandarin-Speaking Children with Specific Language Impairment: Optional Movement. *Journal of Language, Linguistics and Literature*, *4*(1), 8-18.

Zhang, X., & Tomblin, J. B. (2000). The association of intervention receipt with speech-language profiles and social-demographic variables. *American Journal of Speech-Language Pathology*, *9*(4), 345-357.

Zhang, Y., Xu, X., Jiang, Y., Sun, W., Wang, Y., Song, Y., Dong, S., Zhu, Q., Jiang, F., & Sheng, L. (2021). Early language and communication development in Chinese children: Adaption and validation of a parent report instrument. *International Journal of Speech-Language Pathology*, *23*(4), 372-381. https://doi.org/10.1080/17549507.2020.1817558

Zhou, J., & Zhou, S. (2019). A study on differences between Taiwanese mandarin and mainland mandarin in vocabulary. In *3rd International Conference on Culture, Education and Economic Development of Modern Society (ICCESE 2019)* (pp. 212-215). Atlantis Press.

Table 1. Countries and regions that use Chinese as an official language

	Spoken	Written
Hong Kong	Cantonese	Traditional Chinese Character
Macau	Cantonese	Traditional Chinese Character
Mainland China	Mandarin	Simplified Chinese Character
Singapore	Mandarin	Simplified Chinese character
Taiwan	Mandarin	Traditional Chinese Character

Note. In addition to Chinese, English is an official language in Hong Kong, Portuguese is an official language in Macau, and English, Malay, and Tamil are official languages in Singapore.

Table 2. Search terms

English database	Simplified Chinese database	Traditional Chinese databas
ProQuest and PubMed	CNKI	ITPLS, and Airiti library
(Chinese OR Mandarin OR Cantonese) AND (Pediatric OR child OR adolescent) AND (language impairment OR language disorder OR SLI OR DLD OR language delay OR language disability)	语言损伤 (language impairment) OR 语言障碍 (language disability) OR 语言迟缓 (language delay) AND 儿童 (child) OR 学生 (student)	語言損傷 (language impairment) OR 語言障礙 (language disability) OR 語言遲緩 (language delay) AND 兒童 (child) OR 學生 (student)

Note. For CNKI searches, we limited the journals to those that are included in CSSCI, CSCD (中国科学引文数据库 Chinese Science Citation Database), or *A Guide to the Core Journals of China* published by Peking University (北京大学《中文核心期刊要目总览》).

CNKI: China National Knowledge Infrastructure, 中国知网; ITPLS: Index to Taiwan Periodical Literature System, 臺灣期刊論文索引系統; Airiti Library: 華藝線上圖書館.

Table 3 Summary of quality appraisal results

Appraisal question	Option	n of articles
1. Were the criteria for inclusion in the sample clearly defined? [a]	**Yes**: specified cutoff scores on language and nonverbal IQ tests for both groups. The actual numbers were not scrutinized as long as clear criteria were provided,	52
	Incomplete: only one group is defined	3
	No: neither is defined	4
2. Were the study subjects and the setting described in detail? [b]	**Yes**: report age, sex, language assessment scores and recruitment site for both TD and DLD	15
	Incomplete: mention at least one, but not all four criteria	44
	No: mention none of the 4 criteria	0
3. Was language ability measured in a valid and reliable way?	**Yes**: used published standardized tests or reported validity (construct validity or diagnostic validity) and reliability (test retest reliability or internal consistency) of self-designed test	56
	Unclear: did not report validity and reliability of any test	0
	No: no language measure	3
4. Were standard criteria used for measurement of the condition? [c]	**Yes**: at least -1SD below the mean on one or more language measures and nonverbal IQ $\geqslant$ 70.	49
	Incomplete: meet at least one but not both criteria	7
	No: did not meet any criteria	3
5. Were confounding factors identified? [d]	**Yes**: compared the two groups on all of the three variables: age, socioeconomic status and dialect/language exposure	2
	Incomplete: compared the two groups on at least one but not all three variables	54
	No: did not compare the two groups on any of the three variables	3
6. Were strategies to deal with confounding factors stated?	**Yes**: identified the three variables in Q5; the variables (or some of variables) were NOT comparable; dealt with them.	1
	No: identified the three variables in Q5; the variables (or some of variables) were NOT comparable; but did NOT deal with them.	1
	Not applicable: identified the three variables in Q5 and all the variables were comparable (no need to address); did not compare any of the variables (no way to address); Incomplete in Q5.	57
7. Were the outcomes measured in a valid and reliable way?	**Yes**: if outcomes were production measures or existing instruments, did the authors report inter-rater reliability for transcribing/coding/scoring? OR if outcomes were comprehension or EEG measures, did the description of the measure show face validity?	43
	No: did not fit above descriptions	16

8. Was appropriate statistical analysis used?	**Yes**: check assumptions and report them to support the selected statistical analysis	15
	Unclear: the analysis seems to be appropriate but did not report the check of assumptions	43
	No: clearly wrong analysis methods	1

a. unfulfilled criteria: TD group is not clearly defined (n= 3);
b. unfulfilled criteria: age (n=1), sex (n =24), language assessment scores (n=17);
c. unfulfilled criteria: nonverbal IQ (n=7) ;
d. unfulfilled criteria: age (n=1), socioeconomic status (n =48), dialect/language exposure (n=57) .

Table 4 Individual study and sample summary

Reference	Study Characteristics				Sample Characteristics					
	Location	Language	N (total)	Type and N of TD comparison group	TD			DLD		
					N	Age (months)	NVIQ	N	Age (months)	NVIQ
Chen & Durrleman (2022)	Mainland China	Mandarin	28	1 Age, nonverbal IQ matched group	14	63.42	101.14[c]	14	61.11	97.5[c]
Chen & Liu (2010)	Taiwan	Mandarin	40	1 Age, gender, nonverbal IQ, maternal education matched group	20	110	100.65[l]	20	109	96.5[l]
Chen & Liu (2014)	Taiwan	Mandarin	70	1 Age, gender, maternal education matched group	33	65.6	102.7[a]	37	65.4	96.3[a]
Chen & Liu (2018)	Taiwan	Mandarin	54	1 Age, gender, maternal education matched group	29	65.8	101.1[a]	26	65.2	95.5[a]
Chen et al. (2021)	Mainland China	Mandarin	34	1 AM group	17	62.31	111.06[o]	17	61.38	96.18[o]
Chen et al. (2022)	Mainland China	Mandarin	40	1 AM group *Also includes an HFA group	20	63.24	106.95[c]	20	63	91.3[c]
Cheng et al. (2021)	Taiwan	Mandarin	24	1 AM group	12	57.92	71.17[m]	12	58.42	28.75[m/r/s]
Cheung (2003)	Taiwan	Mandarin	30		AM: 10	88	NR	10	86	NR

				1 AM group & 1 LM group using Preschool Language Disorder Test	LM: 10	67				
Chi (2007)	Taipei; Taiwan	Mandarin	36	1 Age, gender, SES matched group	18	Primary grade 1	23.50[b]	18	Primary grade 1	22.78[b]
Chi et al. (2012)	Taiwan	Mandarin	42	1 Age, gender, SES matched group	21	84	99.33[a]	21	86.04	93.76[a]
Dai & He (2021)	Mainland China	Mandarin	40	1 AM group Also includes an HFA group	26	61.63	WNL[c]	14	64.53	WNL[c]
Fletcher et al. (2005)	Hong Kong	Cantonese	45	1 AM group & 1 LM group using MLU	AM: 15 LM: 15	60.47 37.73	110.67[d] 115.13[d/t]	15	60.67	97.47[d]
Hao et al. (2018)	Shanghai; Mainland China	Mandarin	36	1 Age, maternal education, and nonverbal IQ matched group	18	74	98.89[e]	18	74	90.44[e]
He & Dai (2012)	Changsha, Guangzhou, Hefei, Xinxiang; Mainland China	Mandarin	36	1 AM group & 1 LM group using MLU; Both gender matched	AM: 12 LM: 12	59.5 NR	WNL[f]	12	59.33	≥80[f]
He & Yu (2013)	Changsha, Guangzhou, Hefei, Xinxiang; Mainland China	Mandarin	36	1 Age and gender matched group & 1 LM group using MLU	AM: 12 LM: 12	NR Range: 35 - 59	NR	12	Range: 48-73	NR
He et al. (2013)	Mainland China	Mandarin	36		AM: 12	59.67	NR	12	59.75	95.67[f]

				1 AM group & 1 LM group using MLU	LM: 12	44.08				
Iao et al. (2017)	Hong Kong	Cantonese	32	1 Age and nonverbal IQ matched group	16	116.25	109[g]	16	119	109.25[g]
Kidd et al. (2017)	Hong Kong	Cantonese	110	1 AM group	53	70.2	110.8[g]	57	70.1	105.2[g]
Kim et al. (2019)	Shanghai; Mainland China	Mandarin	45	1 AM group & 1 LM group matching using PPVT-R; Both gender matched	AM:15 LM: 15	69.47 57.13	30.07[h] 25.47[h]	15	69.47	28.4[h]
Klee et al. (2004)	Hong Kong	Cantonese	45	1 AM group & 1 LM group using RDLS-C	AM: 15 LM: 15	56.87 35.93	NR	15	56.4	WNL[d/i]
Lai & Wang (2017)	Taipei, Taichung; Taiwan	Mandarin	60	1 Age and gender matched group	30	91.2	NR	30	89.6	≥85[a]
Leonard et al. (2006)	Hong Kong	Cantonese	45	1 AM group & 1 LM group using MLU	AM:15 LM:15	61 38	WNL[d] WNL[d/j]	15	61	97.47[d]
Leonard et al. (2007)	Hong Kong	Cantonese	45	1 AM group & 1 LM group using MLU	AM: 15 LM: 15	60.67 37.73	WNL[d] WNL[d/j]	15	60.67	≥83[d]
Lin & Zhang (2021)	Shanghai; Mainland China	Mandarin	40	1 Age and gender matched group	20	60.51	NR	20	60.64	NR
Liu & Chien (2020)	Taiwan	Mandarin	44	1 AM group Also includes a SSD only group	24	49.33	114.4[a]	20 (SSD+LI)	51.5	96.8[a]

						Longitudinal study from 2-4 years:			Longitudinal study from 2-4 years:	
Lv & Tsao (2018)	Taipei, New Taipei city, Taoyuan, Yilan; Taiwan	Mandarin	65	1 AM group Also includes a late talker group	33	T1: 27.83 T2: 39.44 T3: 51.19	T1: 11.94[k] T2:121.70[j] T3: 102.51[c]	12	T1: 28.37 T2: 39.44 T3: 50.12	T1: 9.42[k] T2: 125.17[j] T3: 96.5[c]
Ma & Liang (2019)	Mainland China	Mandarin	32	1 AM group	16	94.62	≥85[u]	16	92.88	≥85[u]
Sheng et al. (2020)	Nanjing; Mainland China	Mandarin	42	1 Age, nonverbal IQ, maternal education, and Mandarin exposure matched group	21	67.5	120.1[e]	21	67.6	120.6[e]
Siu & Man (2006)	Hong Kong	Cantonese	46	1 Age and gender matched group	23	67.3	NR	23	68.3	NR
Stokes & Fletcher (2000)	Hong Kong	Cantonese	30	1 LM group using MLU	15	31	NR	15	48	NR
Stokes & Fletcher (2003)	Hong Kong	Cantonese	27	1 AM group	14	52.7	NR	13	53.5	NR
Stokes & So (1997)	Hong Kong	Cantonese	28	1 AM group	14	52.78	NR	14	53.83	NR
Stokes et al. (2006)	Hong Kong	Cantonese	44	1 AM group & 1 LM group using MLU	AM: 15 LM: 15	60 39	110.67[d] 112.47[d/j]	14	59	97.43[d]

To et al. (2010)	Hong Kong	Cantonese	50	1 AM group (normative sample)	1080	58 - 145	≥85[l]	50	66 - 144	≥85[l]
Torng & Sah (2020)	Taiwan	Mandarin	36	1 Age, gender matched group	18	63.6	98.06[a]	18	64.44	89.56[a]
Tsai & Chang (2008)	Taiwan	Mandarin	12	1 Age, SES matched group	6	103	NR	6	103	WNL[m] /n
Tseng & Liu (2017)	Taipei, Keelung; Taiwan	Mandarin	20	1 Age, gender, SES matched group	10	142	99.8[a]	10	139	87.4[a]
Wang & Huang (2016)	Mainland China	Mandarin	46	1 Age and gender matched group	23	54.87	NR	23	54.49	NR
Wang & Yu (2021)	Mainland China	Mandarin	45	1 Age and gender matched group & younger TD group	AM: 15 YTD: 15	62.1 45	NR	13	61.77	WNL[m]
Wang et al. (2020)	Mainland China	Mandarin	45	1 AM group & 1 LM group using PPVT-R; Both gender matched.	AM: 15 LM:15	69.47 57.13	30.07[h] 25.47[h]	15	69.47	28.4[h]
Wang et al. (2022)	Shanghai; Mainland China	Mandarin	32	1 Age, maternal education and Mandarin exposure matched group	16	60.8	120.9[m]	16	60.3	101.9[m]
Wong et al. (2003)	Hong Kong	Cantonese	12	1 LM group using MLU	6	36	NR	6	60	WNL[d]
Wong et al. (2004)	Hong Kong	Cantonese	31	1 AM group & 1 LM group using MLU	AM: 11 LM: 9	60.36 40.11	WNL[d] WNL[d/j]	11	60.64	99.55[d]
Wong et al. (2009)	Hong Kong	Cantonese	41		AM: 14	64.86	111.29[d]	14	64	

				1 AM group & 1 LM group using CRVT	LM: 13	51	112.92[d]			103.86[d]
Wong et al. (2010a)	Hong Kong	Cantonese	39	1 AM group *Also includes a SLI-History group	9	112.11	WNL[g]	DLD-only: 7	106.29	WNL[p/g]
								DLD+dyslexia:13	106.54	WNL[p/g]
Wong et al. (2010b)	Hong Kong	Cantonese	29	1 AM group	14	55.71	108.93[d]	15	55.27	102.8[d]
Wong et al. (2015)	Hong Kong	Cantonese	94	1 AM group	40	84.93	110.78[g]	DLD-only: 19	84.66	109.32[g]
								DLD+dyslexia: 25	86.8	100.80[g]
Wong et al. (2017)	Hong Kong	Cantonese	74	1 AM group	34	96.68	110.76[g]	DLD-only: 18	95.83	110[g]
								DLD+dyslexia: 22	99.36	101.41[g]
Wu (2020)	Taipei, New Taipei city, Miaoli; Taiwan	Mandarin	124	1 AM group	T5: 54	64	NR	D5: 18	64	NR
					T6: 35	75		D6: 17	75	
Wu et al. (2019)	Taipei; Taiwan	Mandarin	134	1 AM group	T3: 40	41	NR	D3: 22	40	NR
					T4: 38	52		D4: 34	52	
Xu et al. (2020)	Mainland China	Mandarin	36		AM: 12	59.67	NR	12	59.75	95.67[f]

				1 AM group & 1 LM group using MLU; Both gender matched	LM:12	44.08				
Xue et al. (2022)	Mainland China	Mandarin	11	1 Age and nonverbal IQ matched group	Lower grades: 26	98.28	-0.54[q]	Lower grades: 29	96.6	-0.35[q]
					Higher grades: 29	132	-0.44[q]	Higher grades: 26	132	-0.69[q]
Yu (2016)	Mainland China	Mandarin	36	1 AM group & 1 LM group using MLU; Both gender matched	AM: 12	54	NR	12	58	≥80[f]
					LM: 12	42				
Yu et al. (2017)	Mainland China	Mandarin	36	1 Age and gender matched group & LM group using MLU	AM: 12	59.5	NR	12	59.3	≥80[f]
					LM: 12	NR				
Yu et al. (2019)	Mainland China	Mandarin	48	1 AM group & 1 LM group using comprehension scores of RLDS-school	AM: 16	105.59	112[g]	16	104.24	103[g]
					LM: 16	87.82	105[g]			
Yu et al. (2021)	Mainland China	Mandarin	44	1 Age and gender matched group & younger TD group	AM: 14	62.27	NR	15	62.83	WNL[m]
					YTD: 15	45.01				
Yu et al. (2022)	Mainland China	Mandarin	45	1 Age and gender matched group & younger TD group	AM: 15	62.18	NR	15	61.77	WNL[m]
					YTD: 15	45.01				

Zeng et al. (2013)	Changsha, Guangzhou, Hefei, Xinxiang; Mainland China	Mandarin	36	1 AM group & 1 LM group using MLU	AM: 12 LM: 12	NR Range: 35 - 59	NR	12	59.33	≥80[f]
Zeng et al. (2018)	Changsha, Guangzhou, Hefei, Xinxiang; Mainland China	Mandarin	36	1 AM group & 1 LM group using MLU	AM: 12 LM:12	59.67 44.08	NR	12	59.75	95.7[f]

Note. NVIQ: nonverbal IQ; RLDS-school: Revised Language disorder Scale for School-age Children; RDLS-C: Reynell Developmental Language Scales-Cantonese (Hong Kong version); CRVT: Cantonese Receptive vocabulary test; WNL: Within normal limits.

a: Test of Nonverbal Intelligence-3rd edition (TONI- III, Chinese version; Wu et al., 2006); b: Raven's Coloured Progressive Matrices (Raven & Court, 1996); c: Nonverbal IQ of Wechsler preschool and primary scale of intelligence-fourth edition (Chinese version) (Li & Zhu, 2012); d: Columbia Mental Maturity Scale (Burgemeister et al., 1972); e: Primary Test of Nonverbal Intelligence (PTONI; Ehrler & McGhee, 2008); f: McCarthy Scale of Children's Abilities –Chinese Revised Version (MSCA-CR 课题组 [The team of MSCA-CR], 1991); g: Raven's Standard Progressive Matrices (Raven SPM; Raven, 1986); h: Combined Raven's Test (Li et al., 1988); i: Bayley Scales of Infant Mental Development—Revised (Bayley, 1993); j: Leiter International Performance Scale—Revised (Roid & Miller, 1997); k: Bayley Scales of Infant and Toddler Development-Third Edition (Bayley-III; Bayley, 2006); l: Performance IQ of Hong Kong Wechsler Intelligence Scales for Children (Chen, 1997); m: Performance IQ of Wechsler Perschool and Primary Scale of Intelligence-Revised (Chinese version; Chen & Chen, 2000); n: Test of Nonverbal Intelligence-2rd edition (TONI-II) (Chinese version; Wu et al., 1996); o: Full IQ of Wechsler Perschool and Primary Scale of Intelligence (WPPSI-IV; Wechsler, 2012); p: Full IQ of Hong Kong Wechsler Intelligence Scales for Children (HK-WISC) (Wechsler, 1981); q: Z-scores of Matrix Analogy Reasoning Set 2 and Set 4 (Naglieri, 1995); r: Stanford-Binet Intelligence Scale-V (Roid, 2003); s: Abbreviated battery IQ (Roid, 2003); t: Leiter International Performance Scale (Leiter, 1979)

Table 5. Tests used in the 59 articles divided by geographic regions

Geographic regions	Test	Test type	Number of articles	Cut-off distribution
Hong Kong	Reynell Developmental Language Scales - receptive subscale	Omnibus	9	1SD: 3 1.20SD: 4 1.25SD: 1 1.50SD: 1
	Hong Kong Cantonese Oral Language Assessment Scales	Omnibus	5	1.25SD: 5
	Reynell Developmental Language Scales - expressive subscale	Omnibus	3	1.50SD: 3
	Hong Kong Cantonese Receptive Vocabulary Test	Vocabulary	1	1SD:1
Mainland China	Peabody Picture Vocabulary Test-Revised version	Vocabulary	20	1SD: 6 1.25SD: 14
	Revised Language Disorder Scale for Preschool Children	Omnibus	12	1.25SD: 12
	McCarthy Scale of Children's Abilities (Chinese Revised Version) - verbal subtests	Verbal IQ	5	1.25SD: 5
	Diagnostic Receptive and Expressive Assessment of Mandarin	Omnibus	5	1SD: 3 1.30SD: 2
	Revised Language Disorder Scale for School-age Children	Omnibus	2	1.25SD: 2
	Wechsler Intelligence Scale for Children _verbal subtests	Verbal IQ	2	1.25SD: 2
	Self-design tasks	Omnibus	2	1SD/1.5SD: 1 1.25SD: 1
	Children's Communication Checklist	Omnibus	1	1.25SD: 1
	The Criterion-Referenced Diagnostic Test of Mandarin-speaking Preschool Children with SLI	Omnibus	1	1.50SD: 1
Taiwan	Revised Language Disorder Scale for Preschool Children	Omnibus	7	1SD: 4 1.25SD: 3
	Wechsler Intelligence Scale for Children _verbal subtests	Verbal IQ	4	1SD: 1 1.25SD: 2 1.30SD: 1
	Peabody Picture Vocabulary Test-Revised version	Vocabulary	4	1SD: 1 1.25SD: 2 1.50SD: 1
	Revised Language Disorder Scale for School-age Children	Omnibus	2	1SD: 1 1.50SD: 1
	Children's Oral Language Comprehension Test	Omnibus	2	1.25SD: 2
	Language Disorder Scale of Preschooler-Revised	Omnibus	1	1.30SD: 1

	Test of School-aged Children's Auditory Comprehension	Omnibus	1	1.25SD: 1
	The Specific Language Impairment Checklist	Omnibus	1	1.25SD: 1
	Test of Communication and Language Ability for School-Age Children in Chinese	Omnibus	1	1.30SD: 1
	The Receptive and Expressive Vocabulary Test	Vocabulary	1	1.25SD: 1

Table 6 Outcomes measures

Domain	Measure	Study	Group difference
Grammar: Generic	MLU	Hao et al. (2018)	1[#]
		Klee et al. (2004)*	1
		Lai & Wang (2017)	1[#]
		Sheng et al. (2020)	0
		Tsai & Chang (2008)	1[#]
		Tseng & Liu (2017)	1[#]
		Wong et al. (2009)*	1
		Wong et al. (2010b)*	1
		Wu (2020)	1[#]
		Xue et al. (2022)	1[#]
	complexity: % of complex clauses/ sentences	Hao et al. (2018)	1[#]
		Sheng et al. (2020)	1
		Torng & Sah (2020)	0
		Xue et al. (2022)	1[#]
	complexity: subordination index	Lai & Wang (2017)	1[#]
	complexity: syntactic complexity	To et al. (2010)*	1[#]
	grammaticality	Hao et al. (2018)	0[#]
		Lai & Wang (2017)	1[#]
		Sheng et al. (2020)	0
	sentence repetition	Stokes et al. (2006)*	1
		Wang et al. (2022)	1
	sentence comprehension	Siu & Man (2006)*	1
	artificial grammar learning	Iao et al. (2017)*	1[#]
Grammar: Fine-grained	aspect markers	Chen & Durrleman (2022)	1
		Chen et al. (2021)	1
		Chen et al. (2022)	1
		Fletcher et al. (2005)*	1
		Hao et al. (2018)	1[#]
		He et al. (2013)	1
		Stokes & Fletcher (2003)*	1
		Wong et al. (2003)*	0
		Yu et al. (2019)	1[#]
	passive sentence/ construction	Hao et al. (2018)	1[#]
		Leonard et al. (2006)*	0
		Yu et al. (2022)	1
		Zeng et al. (2018)	1
	Relative clause	He & Yu (2013)	1
		Yu et al. (2017)	1
		Wang & Yu (2021)	1
		Yu et al. (2021)	1

	classifier phrases	Hao et al. (2018)	1[#]
		Stokes & So (1997)*	1
		Xu et al. (2020)	1
	Ba sentence	Hao et al. (2018)	0[#]
		Zeng et al. (2013)	1
	Wh-question	Dai & He (2021)	1
		Wong et al. (2004)*	1
	negation production	Hao et al. (2018)	0[#]
		He & Dai (2012)	1
	modal auxiliary production	Leonard et al. (2007)*	NR
	topic sentence structure	Yu (2016)	1
Vocabulary	NDW	Hao et al. (2018)	1[#]
		Lai & Wang (2017)	1[#]
		Sheng et al. (2020)	1
		Stokes & Fletcher (2000)*	NR
		Torng & Sah (2020)	1
		Tsai & Chang (2008)	1[#]
		Wong et al. (2010b)*	1
		Xue et al. (2022)	1[#]
	D	Klee et al. (2004)*	1
		Wong et al. (2010b)*	1
		Wu et al. (2019)	1
	TTR	Tseng & Liu (2017)	0[#]
	corrected TTR	Tseng & Liu (2017)	1[#]
	MLTD	Wu et al. (2019)	1
	MATTR	Wu et al. (2019)	1
	sophisticated vocabulary	To et al. (2010)*	1[#]
	standardized vocabulary test	Wong et al. (2009)*	1
	word learning	Chen & Liu (2014)	1
		Kim et al. (2019)	1
		Ma & Liang (2019)	1[#]
		Wang et al. (2020)	1
Phonology	phonological memory/ nonword repetition	Chen & Liu (2018)	1
		Chi (2007)	1[#]
		Kidd et al. (2017)*	1
		Ma & Liang (2019)	1[#]
		Siu & Man (2006)*	1
		Stokes et al. (2006)*	0
		Wang & Huang (2016)	1
		Wong et al. (2010a)*	1[#]
		Wong et al. (2015)*	1[#]
	speech perception	Chen & Liu (2010)	1[#]
		Cheng et al. (2021)	1
		Chi (2007)	NR
		Kidd et al. (2017)*	1
		Wong et al. (2009)*	1
	phonological awareness	Chi (2007)	NR

		Kidd et al. (2017)*	1
		Wong et al. (2010a)*	1[#]
		Wong et al. (2015)*	1[#]
	phonological retrieval	Kidd et al. (2017)*	0
		Wong et al. (2010a)*	1[#]
		Wong et al. (2015)*	0[#]
	phonological representation	Kidd et al. (2017)*	1
		Wong et al. (2010a)*	0[#]
	speech production	Liu & Chein (2020)	1
Productivity	TNW	Hao et al. (2018)	0[#]
		Lai & Wang (2017)	1[#]
		Sheng et al. (2020)	0
		Stokes & Fletcher (2000)*	NR
		Tsai & Chang (2008)	1[#]
		Tseng & Liu (2017)	0[#]
		Xue et al. (2022)	1[#]
		Wong et al. (2010b)*	1
	TNU	Hao et al. (2018)	0[#]
		Sheng et al. (2020)	0
		Torng & Sah (2020)	0
		Tsai & Chang (2008)	0[#]
		Xue et al. (2022)	0[#]
Narrative discourse	story macrostructure	Chi et al. (2012)	1[#]
		Hao et al. (2018)	1[#]
		Lai & Wang (2017)	1[#]
		Sheng et al. (2020)	1
		Torng & Sah (2020)	1
		Tsai & Chang (2008)	1[#]
		Xue et al. (2022)	1[#]
	discourse pragmatics	Lai & Wang (2017)	1[#]
		To et al. (2010)*	1[#]
		Torng & Sah (2020)	1
		Tsai & Chang (2008)	1[#]
		Xue et al. (2022)	1[#]
	story comprehension	Chen & Liu (2014)	1
Literacy	word reading	Chi et al. (2012)	1[#]
		Wong et al. (2010a)*	0[#]
		Wong et al. (2017)*	1[#]
	morphological awareness	Wong et al. (2010a)*	1[#]
	morphological awareness	Wong et al. (2015)*	1[#]
	orthographic skills	Wong et al. (2010a)*	0[#]
	orthographic skills	Wong et al. (2015)*	0[#]
	reading comprehension	Wong et al. (2017)*	1[#]
	word dictation	Wong et al. (2010a)*	0[#]
Miscellaneous	general language abilities	Chi et al. (2012)	1[#]
		Lin & Zhang (2021)	1
		Lv & Tsao (2018)	NR

	executive function	Lin & Zhang (2021)	1
	verbal working memory	Wong et al. (2017)*	1#
	verbal memory	Cheung (2003)	0#
	spatial memory	Cheung (2003)	1#
	fluency	Tseng & Liu (2017)	0#

Note. * Cantonese-speaking participants; 0: group differences were not found; 1: group differences were found; NR: the original study did not provide direct group comparison, or the TD control was not age-matched; #the sample consists of primarily school-age children. The absence of # indicates that the study included primarily preschool-aged children