Research Article

Late Talkers' Language, Metaphor, Theory of Mind, and Reading Skills at 9 Years of Age

Camilla E. Crawshaw,[a] Carina Lüke,[b] and Ute Ritterfeld[a]

[a] Research Unit of Language and Communication, Department of Rehabilitation Sciences, TU Dortmund University, Germany [b] Special Education and Therapy in Language and Communication Disorders, Institute of Special Education, Faculty of Human Sciences, Julius-Maximilians-University of Würzburg, Germany

ARTICLE INFO

Article History:
Received November 12, 2023
Revision received April 19, 2024
Accepted November 4, 2024

Editor-in-Chief: Julie A. Washington
Editor: Raúl Rojas

https://doi.org/10.1044/2024_JSLHR-23-00703

ABSTRACT

Purpose: Prior work has found that "late talkers" (LTs) as a group continue to demonstrate lower language and reading outcomes compared to their typically developing (TD) peers even into young adulthood. Others identified that children diagnosed with developmental language disorder (DLD) show difficulties later with theory of mind (ToM) tasks and metaphor comprehension, but there is a shortage of research specifically investigating these advanced skills in LTs. The current study therefore compared language-related skills of former LTs with their TD peers at school age.

Method: A longitudinal sample (N = 35) of monolingual German-speaking children was observed from age 1 until 9, comprising TD children (n = 27) and children identified as LTs at age 2 (n = 8), of which two met criteria for DLD between ages 3 and 6. Children's language (productive vocabulary, productive and receptive grammar), reading, metaphor comprehension, and ToM skills (ToM scale and Strange Stories) were investigated, and group comparisons were conducted.

Results: Former LTs performed worse than the TD children on measures of productive vocabulary, receptive grammar, metaphor comprehension, and the ToM Strange Stories task at the age of 9, but not on measures of productive grammar, reading, or the ToM scale.

Conclusions: The findings indicate that LTs can catch up with their TD peers in some areas of language and ToM but that subtle differences remain across other complex areas. Further research is needed to pinpoint possible explanations for why certain skills are more strongly impacted and the potential developmental interactions between these competencies.

In contrast to children who are typically developing (TD) or who go on to manifest a developmental language disorder (DLD), less focus has been placed on the outcomes of "late talkers" (LTs) who seem to "catch up" with their TD peers, especially regarding their abilities beyond structural language. LT status refers to children who have no known cognitive impairment or hearing loss, demonstrate a comparably small productive vocabulary between the ages of 12–24 months, do not start to combine words at the age of 24 months, and demonstrate differences in both noun and verb acquisition (cf. Horvath et al., 2022; Perry et al., 2023; Rescorla, 2009, 2011; Sansavini et al., 2021). Other differences can persist in speech processing and visual attention (Perry et al., 2023). Estimates of prevalence vary between 10% and 20% of all 2-year-olds (cf. Zubrick et al., 2007). Many LTs go on to reach expectations set for their TD peers, although for some, this may not be until they are 5 years old, and up to 40% of LTs will not potentially going on to receive a diagnosis of DLD (e.g., Bates et al., 1995; Bishop, 2017; Norbury et al., 2016; Rescorla, 2011). At first, those children who do appear to catch up were believed to resume a typical developmental trajectory. However, more recent work has observed that they experience persistently lower outcomes in language and language-

Correspondence to Camilla E. Crawshaw: camilla.crawshaw@tu-dortmund.de. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

related skills, even into early adulthood (cf. Bates et al., 1995; Horvath et al., 2022; Perry et al., 2023; Rescorla, 2009, 2011; Sansavini et al., 2021). This suggests that TD children, LTs, and children with DLD may fall—albeit differing quantitatively from one another—along dimensional spectra of language(-related) abilities (Rescorla, 2009, 2011; Thal et al., 2013).

Early findings (e.g., Bates et al., 1995; Rescorla et al., 1997) were published more than 25 years ago. However, since then relatively few studies have analyzed the development of language and reading alongside other language-related skills (e.g., theory of mind [ToM] or figurative language) in older children with a history of being LTs but who do not necessarily meet criteria for DLD. We will investigate where LTs might fall compared to their peers along linguistic (productive vocabulary, productive grammar, receptive grammar) and language-related developmental spectra at school age. This includes reading comprehension and two other complex skills bridging language and social cognition: ToM and metaphor comprehension. With reference to prior work, the following introductory sections will briefly address each skill and present the current picture of their development in LTs. Where gaps exist, we will consider those with DLD for analogical potential.

Language

A number of longitudinal studies have followed LTs until the ages of 4, 5, or 6 years (e.g., Hammer et al., 2017), but only a scarce few have continued beyond (for a discussion, see Rescorla, 2009). One longitudinal cohort of 56 LTs (Bishop & Edmundson, 1987) were subdivided at age 5;6 (years;months) into those whose language issues had resolved (*n* = 26) versus those whose had not (*n* = 30) and followed up until the age of 15 years (Stothard et al. 1998). Compared to age-matched TD controls, the children with resolved language issues did not differ significantly on vocabulary and language comprehension skill tests but did perform significantly less well on phonological processing and literacy skill tests (Stothard et al., 1998). The children with persistent language issues demonstrated significantly lower performance across all measured areas of spoken and written language (Stothard et al., 1998). Rescorla (2009) followed the development of 26 children identified as LTs at intake (24–31 months) who had typical nonverbal cognitive ability and receptive language, alongside 23 TD children matched at intake on age, socioeconomic status, and nonverbal ability, until they were 17 years old. Rescorla's (2009) LTs demonstrated performance in the average range on all language and reading tasks at 17 years of age but still achieved significantly lower vocabulary, grammar, and verbal memory scores than their TD peers, despite no significant differences in their reading and writing scores. Thus, evidence from prior work indicates that LTs' performance on language and reading measures might change over time, differ at both the individual and sample level, and depend upon how these skills are measured. Even so, it has also revealed that many LTs do experience long-lasting, residual challenges in their language outcomes.

Reading

Many studies have already demonstrated that children's early oral language skills can predict their later reading comprehension abilities (e.g., Hjetland et al., 2019; Language and Reading Research Consortium & Chiu, 2018). However, a distinction between LT and TD reading abilities seems less clear. In Bishop and Edmundson's (1987) and Stothard et al.'s (1998) longitudinal sample, LTs performed significantly worse at 15 years of age than their TD peers on reading skills comprising single-word reading, single-word spelling, and reading comprehension. In Rescorla's (2009) sample at 17 years of age, however, there were no significant differences across a set of subtests evaluating ability to decode words on a list, timed reading and comprehension of short statements, and timed writing of short statements using target words. Interestingly, when Bishop and Edmundson's sample was assessed at age 8;6, the "recovered" LTs (non-DLD) displayed no reading or spelling difficulties and performed within the normal range on tasks tapping phonological strategies: nonword reading and spelling (cf. Stothard et al., 1998). In contrast, although Rescorla's sample mostly performed within the average range on reading measures when they were aged 8, 9, and 13 years, they still demonstrated significantly lower performance than their TD peers. These inconsistencies might suggest that longitudinal findings regarding LTs are influenced by differences at both the group and individual levels within children's language and literacy development (Bates et al., 1995).

More recent, fine-grained approaches to literacy research may support this lack of a clear-cut distinction in LTs' comparative reading skills. A longitudinal study by Psyridou et al. (2018) with 200 children aged 2–16 years found that LTs who developed dyslexia had experienced both expressive and receptive vocabulary delay as well as a family risk for dyslexia. These children struggled with reading comprehension but not reading fluency, sustaining difficulties into adolescence, while LTs without receptive vocabulary difficulties generally became typical readers (Psyridou et al., 2018). Psyridou et al. argued that being an LT was not a sufficient risk index for developing reading comprehension difficulties. This would align with prior work, signifying different types of and severities within

LTs (e.g., Thal et al., 2013), which do not facilitate simple binary classifications (Dollaghan, 2013), and highlighting a potential need for the building of subgroups.

ToM

ToM refers to the ability to understand and attribute one's own and others' mental states and representations; taps perspective-taking skills; and is closely linked to other linguistic, pragmatic, cognitive, and social skills (cf. Devine & Lecce, 2021). Within the TD population, previous work has already established that language and ToM are co-developing skills, with earlier language skills predicting later ToM performance (for a meta-analysis, see Milligan et al. 2007). However, researchers are still divided on exactly which components of language are facilitative. Some authors have specified grammatical constructions such as sentential complementation (e.g., J. de Villiers, 2007; Durrleman et al., 2017). Some have highlighted the impact of receptive grammar and sentence comprehension skills (De Mulder et al., 2019; Ebert, 2020). Others have instead emphasized the role of more general language and communicative skills (e.g., Ensor et al., 2014; Milligan et al., 2007). It has been demonstrated that vocabulary can also independently predict ToM performance (De Mulder et al., 2019; P. A. de Villiers, 2005; Devine et al., 2016; Ebert, 2020; Happé, 1995), and some have underlined the importance of mental-state verbs (Astington & Baird, 2005).

ToM in LTs

To our knowledge, no existing work has gone beyond expressive and receptive language or reading skills to consider metacognitive skills in LTs who do not necessarily go on to manifest DLD. This gap should be addressed since extensive prior work has shown early linguistic skills are formed by and chronologically aligned with the emergence of nonlinguistic skills (cf. Thal et al., 2013). Thal et al.'s (2013) longitudinal study followed LTs until the age of 7 years, identifying a spectrum spanning TD to LT with delayed production but typical comprehension (late producer) and on further to LT with both delayed production and comprehension (late comprehender). Group differences were evident at as early as 10 months of age and across this spectrum, representational skills (in terms of gesture use) were shown to be progressively poorer and lowest in late comprehenders (Thal et al., 2013). Compared to their TD peers, LTs appear to engage in less symbolic play, a precursor to ToM (Paul & Ellis Weismer, 2013). LTs followed over a 15-year longitudinal study used fewer cognitive mental-state terms (e.g., "think," "know") at 5 years of age than their TD peers (Rescorla, 2013); these terms are important for the linguistic embedding of ToM concepts. Early differences in representational ability could be associated with subsequent delays in ToM development (a metarepresentational ability). Alongside TD or autistic populations, prior work has only addressed ToM skills in LTs who received a diagnosis of DLD. Since LTs who appear to have "recovered" likely fall somewhere between their TD peers and peers with DLD along dimensional spectra of performance (Rescorla, 2009, 2011; Thal et al., 2013), we will next discuss existing work on the DLD population to provide a situative context.

ToM in Children With DLD

Children with DLD face issues with language production, comprehension, and processing. They experience varying degrees of impairment across broad-ranging areas of language, including lexical, morphosyntactic, pragmatic, as well as both oral and written skills (Sansavini et al., 2021). Children with DLD often have long-term learning difficulties and can struggle with behavioral, psychological, emotional, and social adaptation, affecting their ability to work and form relationships in adulthood (for a recent review of predictors and outcomes for DLD, see Sansavini et al., 2021). On the whole, prior work has shown that ToM skills are also a delayed area in DLD (Andrés-Roqueta et al., 2013; Durrleman & Delage, 2020; Durrleman et al., 2017; Farrant, 2015; Farrant et al., 2006; Farrar et al., 2009; Gillott et al., 2004; Nilsson & de Lopez, 2016; Rakhlin et al., 2011; Smit et al., 2019; Spanoudis, 2016; Vissers & Koolen, 2016). ToM ability covers many different aspects of mentalizing skills, and a variety of tasks exist to test it (for a recent discussion and overview, see Devine & Lecce, 2021). It is therefore plausible that the type of task utilized might influence whether researchers identify differences in performance across TD children, those who are LTs, and those with DLD. In Table 1, we present an overview of recent studies addressing ToM performance in children with DLD. As can be seen from the table, the studies vary widely in terms of participant ages and ToM tasks used.

In a meta-analysis of ToM skills in DLD, Nilsson and de Lopez (2016) evaluated 17 studies covering a total of 745 children between ages 4 and 12 years and found that children with DLD performed substantially worse on ToM tasks than their age-matched TD peers. Their findings reinforce the idea that (early) language and ToM skills are associated with language facilitating age-appropriate ToM development. They also indicate a potential interface, with impairment in one domain extending into the other (Nilsson & de Lopez, 2016). Vissers and Koolen (2016) conducted a review of studies investigating social–emotional functioning and ToM abilities in children with DLD aged 2;3–6;2. They concluded that preschoolers with DLD experienced social–emotional difficulties and

Table 1. Overview of recent studies considering theory of mind (ToM) performance in children with developmental language disorder (DLD).

Authors	Participants	ToM task(s) used	Findings re. ToM
Andrés-Roqueta et al. (2013)	Children with DLD (*n* = 31) aged 3;5–7;5 (years;months), age- and gender-matched typically developing (TD) controls (*n* = 31), younger language- and gender-matched TD controls (*n* = 31)	2 false belief (FB) tasks: unexpected transfer and unexpected contents	The DLD group performed worse than the age-matched group but similarly to the language-matched group
Durrleman et al. (2017)	Children with DLD (*n* = 20) aged 6;5–11;7, children on autistic spectrum (*n* = 34) aged 6;9–14;4, TD children (*n* = 30) aged 4;9–11;8	Low-verbal picture-sequencing task with a story requiring FB attribution	Children across groups performed similarly depending on their sentential complementation ability
Farrant (2015)	Children with DLD (*n* = 30) aged 4;0–6;2, TD children (*n* = 30) matched for nonverbal ability, gender, and age	Diverse desires, diverse beliefs, knowledge access, FB: unexpected contents, low verbal FB, visual perspective-taking (VPT), emotional perspective-taking	The DLD group performed worse than the TD group on all measures except for diverse beliefs (difference marginally significant)
Farrant et al. (2006)	Children with DLD (*n* = 20) aged 4;10–5;8, TD children (*n* = 20) matched for nonverbal ability, gender, and age	Diverse desires, diverse beliefs, knowledge access, FB: unexpected contents, real–apparent emotions, VPT	The DLD group performed worse than the TD group on more complex VPT, knowledge access, and FB: unexpected contents
Farrar et al. (2009)	Children with DLD (*n* = 34) aged 3;6–5;5	FB: unexpected contents, FB: unexpected transfer, 3 appearance–reality: 2 mistaken attribute, 1 mistaken identity	General grammatical development and vocabulary predicted ToM ability; sentential complementation no unique role
Gillott et al. (2004)	Children with DLD (*n* = 15), children with high-functioning autism (*n* = 15), and TD children (*n* = 15): aged 8–12, age- and gender matched	12 Strange Stories (lie, white lie, joke, pretence, misunderstanding, persuasion, appearance–reality, figure of speech, sarcasm, forgetting, double bluff, contrary emotion)	The DLD and autistic groups gave fewer correct mental-state answers than the TD group; the autistic group gave more inappropriate answers than DLD or TD
Rakhlin et al. (2011)	Children with DLD and IQ above 85 (*n* = 21), DLD and IQ below 85 (*n* = 4), IQ below 85 but no DLD (*n* = 5), TD (*n* = 22), aged 5;0–12;9, non-DLD groups older on average	8 story scenarios for FB: unexpected transfer	The DLD groups performed worse than the non-DLD groups; language development scores related to FB performance even after controlling for IQ and short-term memory
Spanoudis (2016)	Children with DLD (*n* = 20) aged 8;9–12;2, age- and gender-matched TD children (*n* = 20), gender- and language-matched controls (*n* = 18)	18 short stories incorporating faux pas recognition and the Strange Stories: 12 social indiscretions, 6 control stories	The DLD group performed worse than age-matched controls; language and ToM skills were related: syntactic and pragmatic abilities predicted ToM performance

impairments in both cognitive (imitation, joint attention, and false belief [FB] understanding) and affective (recognizing and understanding emotions) ToM (Vissers & Koolen, 2016). Smit et al. (2019) reviewed studies on social emotional and ToM ability in adolescents (ages 10–24 years) with DLD or who were deaf or hard of hearing, examining parallels and establishing a framework that was mediated by limited linguistic competence or restricted language exposure.

Metaphor Comprehension

Metaphor comprehension requires making a nonliteral mapping between concepts and linguistic forms to access and understand an interlocutor's intended meaning. This is often ambiguous and may require perspective-taking skills. Metaphor comprehension has thus been linked with both semantic skills and ToM ability (Deckert et al., 2019; Kalandadze et al., 2018; Lecce et al., 2019; Norbury, 2005). This indicates that metaphor might function as a form of interface between ToM and language (cf. Pronina et al., 2023). Bidirectional longitudinal associations have been found between TD 9-year-old children's ability to understand metaphors and their peer acceptance or rejection outcomes (Del Sette et al., 2021). Work highlighting ToM as a predictor of social functioning ability has also suggested that adolescents with DLD's difficulties forming peer relationships stem from their pragmatic language impairments (Smit et al., 2019). These findings underline that the ability to comprehend metaphors is not merely a useful academic skill but instead crucially important for children's social outcomes. As a pragmatic, metalinguistic skill interfacing ToM and language abilities, metaphor comprehension may thus be especially relevant to consider in LTs. It could also provide a potential opportunity to narrow down challenging developmental domains and help pinpoint underlying, causal deficits. Building on the work of Thal et al. (2013), if some (late comprehender) but not all (late producer) LTs struggle with early representational skills, then metaphor comprehension might be a useful domain to help build and distinguish LT subgroups. Since no prior work appears to have considered metaphor comprehension in LTs without DLD, next we present the findings regarding the DLD population.

Metaphor Comprehension in DLD

In contrast to ToM, only very few studies have considered how well children with DLD comprehend and process metaphors or other types of figurative or abstract language (e.g., Bühler et al., 2018; Lorusso et al., 2015; Norbury, 2005; Spanoudis, 2016). Those addressing metaphor comprehension also incorporated measures of ToM. Norbury (2005) compared TD children (n = 34), language-impaired children (n = 28), language-impaired autistic children (n = 31), and non–language-impaired autistic children (n = 29) across semantic knowledge, ToM, and metaphor comprehension measures. Results showed specifically that both groups of children with language impairment (with and without autistic status) performed significantly worse across all three measures (Norbury, 2005). However, this "language impairment" classification comprised both children with DLD and another language disorder associated with autism. Spanoudis (2016) found that children with DLD performed significantly worse than TD controls on novel metaphor and simile comprehension. Bühler et al. (2018) investigated novel metaphor comprehension in children with DLD (n = 15) aged 3;6–4;1 compared to age-matched TD peers (n = 15) and language-matched younger TD children (n = 15). Children with DLD performed less well than age-matched controls but similarly to language-matched controls, suggesting their delay in metaphor comprehension arose from general, overall linguistic competence rather than difficulties in pragmatic inference making (Bühler et al., 2018). Both Norbury and Spanoudis found correlations between language, ToM, and metaphor comprehension abilities. However, given (a) the paucity of studies on metaphor comprehension in DLD, (b) work suggesting that metaphor comprehension and ToM are linked, and (c) findings indicating that ToM skills are impaired or at least delayed in DLD, it remains difficult to conclude whether children with DLD's metaphor comprehension struggles are purely caused by linguistic issues or whether there may be other contributing factors.

Addressing Potential (Meta) Representational, Inferential, and Decoding Deficits in LTs

Children develop an understanding of pictorial and linguistic representation, enabling them to acquire vocabulary, as a precursor to understanding the metarepresentation—representing one's own and others' representations—required for explicit ToM (Doherty, 2008). Being able to read also requires understanding that written letters, words, and text refer to and represent some object, concept, condition, or context in the real or a possible world. Thus, reading, ToM, and metaphor comprehension may all tap semantic, inferential, or decoding skills in potentially analogous ways. Moreover, to understand written narratives, children actively require ToM skills to attribute mental states, thoughts, and emotions to characters (Gordon Pershey, 2000). Some prior research had identified direct connections between early and advanced ToM skills and later reading comprehension (Atkinson et al., 2017; Boerma et al., 2017), but more recent longitudinal work has indicated this connection is largely mediated by early language skills (Ebert, 2020). On a potential spectrum or even complex continuum of interrelated competencies, reading ability and figurative language comprehension might arguably interact more closely. In a study with 199 TD children, Levorato et al. (2004) found that the ability to understand a text predicted children's understanding of idioms in context. In summary, connections between reading comprehension and ToM or metaphor comprehension may be complex, may draw on similar underlying skills, and may not be easy to disentangle. However, early language skills clearly impact these later skills, and their development should therefore be addressed in LTs as well. ToM is an example of a metarepresentational, metacognitive ability, while metaphor comprehension is a metalinguistic skill that also taps representational abilities. Investigating these skills might reveal important associations between very early and later complex language development in LTs.

Research Aim

Prior work (e.g., Rescorla, 2002, 2009; Rescorla et al., 1997; Stothard et al., 1998) has found that LTs with and without a diagnosis of DLD continue to demonstrate differences from their TD peers in language outcomes even into adolescence. Findings are mixed regarding reading-related skills; these may be impacted at different developmental stages (Bishop & Edmundson, 1987; Rescorla, 2009; Stothard et al. 1998). More recent work suggests reading problems may only occur in combination with a family risk for dyslexia (Psyridou et al., 2018). Both ToM and metaphor comprehension seem delayed in children with DLD (e.g., Nilsson & de Lopez, 2016; Norbury, 2005), but no study has yet investigated whether these skills are also impacted in LTs who appear to "recover."

Earlier language ability predicts later performance on ToM (e.g., Milligan et al., 2007) as well as metaphor processing tasks (e.g., Deckert et al., 2019; Kalandadze et al., 2018), and ToM has been associated with metaphor comprehension (e.g., Lecce et al., 2019; Norbury, 2005). If language ability is conceptualized as a spectrum ("language endowment spectrum"; cf. Ellis Weismer, 2007, p. 84), then it is likely that LTs would also fall somewhere between TD children and children with DLD along spectra or continua of language-associated skills such as ToM and metaphor comprehension. The current study thus sought to explore two research questions:

1. Are language and reading abilities impacted in (German-speaking) LTs at age 9?
2. Are metacognitive and metalinguistic (ToM and metaphor comprehension) abilities affected in 9-year-old children with a history of being an LT?

Method

Participants

The children participating in this study were 35 (18 boys, 17 girls) monolingual German-speaking children who had previously taken part in a wider longitudinal study between the ages of 1 and 6 (for further details, see Crawshaw et al., 2024; Lüke et al., 2020). To address the current research goals, we worked with the children at age 9 (age at first testing session: M = 9 years 0 months 16 days, SD = 11 days; age at second session approximately a week later: M = 9 years 0 months 24 days, SD = 12 days; age at third session approximately a month after first testing: M = 9 years 1 month 18 days, SD = 31 days). Of the 35 children, eight were characterized as LTs at age 2 (four boys, four girls). Of these eight children, two (both girls) received a diagnosis of DLD between ages 3 and 6 but no longer met diagnostic criteria at age 9, so we decided to keep them under the grouping of LTs.

Criteria for LT

The criteria for definition as an LT (at age 2) or as having DLD (ages 3–6) stemmed from the children's scores on standardized language measures and information received via a parental report questionnaire. At age 2, this consisted of seven measures:

- FRAKIS: *Fragebogen zur frühkindlichen Sprachentwicklung* (Questionnaire for Early Language Acquisition), German version of the MacArthur–Bates Communicative Development Inventories for children between 18 and 30 months, including the subscales Vocabulary Size, Morphological Skills, and Syntactic Skills (Szagun et al., 2009).
- SETK-2: *Sprachentwicklungstest für zweijährige Kinder* (Language Acquisition Test for 2-Year-Old Children), including the subtests Word Comprehension, Sentence Comprehension, Word Production, and Sentence Production (Grimm, 2000).

At age 3, three measures were used:

- SETK-3–5: *Sprachentwicklungstest für drei- bis fünfjährige Kinder* (Language Acquisition Test for 3- to 5-Year-Old Children), including the subtests Sentence Comprehension and Sentence Production (Grimm, 2001).
- PDSS: *Patholinguistische Diagnostik bei Sprachentwicklungsstörungen* (Patholinguistic Diagnostics for Developmental Language Disorder), containing the subtest Word Production (Kauschke and Siegmüller, 2010).

At age 4, two measures were used:

- P-ITPA: *Potsdam-Illinois Test für Psycholinguistische Fähigkeiten* (Potsdam-Illinois Test of Psycholinguistic Abilities), including the subtests Word Production and Grammar Production (Esser et al., 2010).

At ages 5 and 6, three measures were used:

- the same two tests as at age 4; and
- TROG-D: *Test zur Überprüfung des Grammatikverständnisses* (German version of the Test for Reception of Grammar; Fox, 2013).

To meet criteria for LT status, children had to score results in at least one language subtest that were 1.5 SD below the mean (i.e., a T score of $\leq$ 35) and 1 SD below the mean in at least one further language subtest (i.e., a T score of < 40). These diagnostic criteria were defined in the early stages of this study, many years prior to the work of the CATALISE consensus across English-speaking countries (Bishop et al., 2017) and are more aligned with the results of a Delphi study later conducted across German-speaking countries (Lüke et al., 2023). At age 3;6, children's nonverbal IQ was assessed using the Snijders–Oomen Non-Verbal Intelligence Test (Tellegen et al., 2007), indicating no differences between LT (Mdn = 98.0, IQR = 25.0) and TD (Mdn = 108.0, IQR = 12.0, U = 56.0, Z = 1.44, p = .151) groups.

Via a questionnaire at age 9, we asked parents to inform us of any diagnoses their children had received since the previous round of data collection at age 6. In the intervening 3 years, four of the children (three TD, one LT) had been officially diagnosed with dyslexia. The task used in this study to measure reading performance would not be appropriate for confirming or issuing a diagnosis of dyslexia (prevalent in the LT population). Since we did not systematically assess dyslexia, confirm any outside diagnoses, or seek potential new ones, we have not addressed it further in our analyses. However, parents were informed and advised that they may wish to seek further support for their children when their performance had been lower than might be expected.

Participants' socioeconomic status as well as their demographic and family background data have already been published in full elsewhere (cf. Lüke et al., 2017), but they came from a WEIRD (Western, Educated, Industrialized, Rich, and Democratic) society. Before commencing the study, all procedures, measures, and assessment of participants were evaluated for ethical considerations and granted approval by the internal review board of TU Dortmund University (reference code: GEKTUDO-2020-13). Informed consent was sought from the parents or caregivers of the participating children, and they were able to withdraw their child along with all relevant data from the study at any time. As each testing session began, the children were also asked to verbally confirm that they wanted to participate. They were informed that they could withdraw from the testing sessions whenever they wished at no disadvantage to themselves or their families.

Procedure

Each child took part in three testing sessions: The first two sessions were conducted online using a digital conferencing tool (Zoom), and the final session was conducted in-person within our lab. During the first session, we utilized a measure of productive vocabulary and the ToM scale. In the second session, we employed a metaphor comprehension task, a reading task, and the Strange Stories tasks. During the third session, we executed measures of productive and receptive grammar. Two research assistants, oblivious to the specific research questions, carried out the testing sessions. Each child consistently interacted with the same experimenter across the three sessions, except for one child who experienced the first two online sessions with one experimenter and the final in-person session with the other experimenter. This scheduling conflict resulted from an unfortunately unavoidable delay caused by the ongoing COVID-19 pandemic. During the online sessions, each child was tested from their home using a laptop or tablet that belonged to their family. All experimental interactions were audio- and video-recorded.

Measures

Productive Vocabulary

We measured the children's vocabulary skills using the Productive Vocabulary subtest from the standardized German-language version of the P-ITPA (Esser et al., 2010). This task taps both semantic and lexical skills. The experimenter presents the child with descriptors to elicit progressively more difficult target words, for example, "I'm thinking about something with feathers, what could that be?" Children are able to score 0, 1, or 2 points for each item, and the test is ended early if a child gives a 0-point answer 6 times in a row. Final raw scores were converted to *T* scores. Taking the example item "I'm thinking about something that has minutes, what could that be?", a 2-point answer could be "a clock," "time (of day)," "hour," and "time (abstract)." These refer to general classifications or superordinates. A 1-point answer could be any specific or concrete example of clocks, time aspects with respect to minutes, or activities and events where measuring time (in minutes) plays an important role, for example, "a stopwatch" or "alarm clock."

Productive Grammar

The children's productive grammar skills were measured with the Expressive Grammar subtest of the standardized P-ITPA (Esser et al., 2010). In this subtest, the experimenter points to a supporting image and begins a sentence for the child to verbally complete, for example, "This apple is big, and this is even … [bigger]" or "This is one flower, and these are four … [flowers]." Children are able to score 0 or 1 points for each item, and the test is ended early if a child gives a 0-point answer 6 times in a row. Final raw scores were converted to *T* scores. Syntactic and morphological abilities are tested across items, requiring children to build:

- past forms of verbs (simple, perfect, and passive forms; separable verbs are included and correct sentential word order is required);
- comparatives and superlatives of adjectives;
- plurals of nouns with examples of masculine, feminine, and neuter genders, as well as both regular and irregular plurals;
- the genitive form of nouns using possessive determiners; and
- case declension of personal pronouns (both accusative and dative forms).

Receptive Grammar

Receptive grammar skills were assessed using the standardized German-language test of grammar comprehension (TROG-D; Fox, 2013). In this test, the experimenter presents the child with a set of four images and then reads a sentence aloud to the child, and the child must then point to the corresponding target image. This test examines sentence comprehension of a variety of grammatical structures marked by inflection, function words, and sentence structure. These include adjectives, nouns, plurals, verbs, negation, use of the perfect tense and passive voice, topicalization, double object construction, disjunctive conjunctions, and coordination with "and." Further items cover

- the prepositions "in," "on," "over," and "under";
- personal pronouns (nominative, accusative and dative);

- two-element sentences (subject–predicate construction, noun phrase with article and adjective);
- three-element sentences (subject–predicate–object);
- subordination with “during,” “after,” and “that”; and
- relative clauses (including pronoun in the accusative or dative cases).

The test contains 84 items, scored in 21 blocks each containing four items. All four items in the block must be correctly answered to “achieve” a grammatical structure. Their final, total number of correct blocks is then converted into a *T* score. The blocks ascend in difficulty, and the test is ended early if at least one out of four test items is incorrect across five consecutive blocks. An example test item, “The cats look at the ball,” includes three distractors alongside the target image: Lexical Distractor 1 (subject and verb), “The boys are playing with the ball”; Lexical Distractor 2 (object), “The cats are looking at the butterfly”; and Grammatical Distractor 3, “The cats are looking at the balls.”

Reading

Children’s reading comprehension skills were assessed using a German-language standardized reading screening measure (*Salzburger Lese-Screening für die Schulstufen 2–9* [Salzburg Reading-Screening-Test for School Levels 2–9]; Wimmer & Mayringer, 2014). This measure focuses on reading speed and accuracy. It consists of a battery of sentences (ceiling: 100 sentences) that gradually increase in length and complexity. The child must indicate whether each sentence’s content is true or false, for example, “Trees can speak.” The sentence booklet was posted to each participating child’s home, and their caregivers were instructed to hide it until the appointed moment during its online testing session. Before the target test items, the experimenter conducts two short practice sessions with two separate sets of practice items. The first practice session is not timed, and answers are discussed together with the experimenter. The second practice session demonstrates how the timing of the test works. For the scored test, the children are given 3 min to complete as many sentences as they can. The items are scored as correct or incorrect and counted. Answers are not scored if they follow 10 mistakes in a row (including not providing an answer). Final raw scores on the test are evaluated relative to a standardized sample of the applicable age. These are then converted to “reading quotients” (RQs) expressing how far the score deviates from the standardization sample’s average. The RQ uses the same scaling as IQ tests, but we converted the children’s RQs into standard *T* scores to compare reading with oral language skills across the descriptive results.

Metaphor Comprehension

We tested the children’s comprehension of a total of 22 age-appropriate metaphors embedded within short sentences to provide context (e.g., *Mein Herz ist gebrochen.* [My heart is broken]). A context-embedded approach is in line with prior work (cf. Babarczy et al., 2019; Özçalışkan, 2005). The metaphors were drawn from a larger set originally developed by Vogt and Indefrey (2017), who shared this list with us, permitting its use. Vogt and Indefrey’s set consisted of 11 subcategories of metaphor, each containing five test items: technomorphic, anthropomorphic, animal, synesthetic (perceptual character), synesthetic (emotional character), pseudosynesthetic, spatial, process metaphors, action metaphors, action metaphors (emotional states), and condition metaphors (emotional states). To shorten this task to fit our study’s time constraints, we conducted a pilot study surveying 30 adult native speakers. We presented the original, full set (11 categories with five metaphors each) and asked respondents to identify the two most common metaphors within each category. Using the data from their responses, we removed the extremes (first and least most common) and selected the second most common and second least common metaphors from each category to form our final set of 22 metaphors (cf. Crawshaw et al., 2024).

For consistency, each metaphor had been prerecorded by a native speaker, and each corresponding audio file was played twice for the children. The experimenter asked the child a sequence of four questions: (a) Could one say such a sentence; (b) have you heard the sentence before; (c) did you understand it; and (d) if the child answered positively, could you explain the sentence in your own words. If the child answered the third question negatively, then they were asked instead what they thought the sentence might mean. During this task, the children were shown a neutral screen consisting of a gray background with a small, black, focusing cross in the center to hold their attention and minimize unnecessary distraction. The children’s answers were recorded, transcribed, and scored as either correct or incorrect on the basis of their explanation of the metaphor in their own words (or what they thought it might mean). A second coder independently scored all responses for 28.6% (a randomly selected 10 of the 35 total participants) of the children for the purpose of calculating interrater reliability (Cohen’s κ = .79).

ToM

The children’s ToM skills were assessed using two separate measures: (a) the final three subtests of the authorized German-language translation of the Extended Theory-of-Mind Scale (originally by Peterson et al., 2012; Wellman & Liu, 2004; translated by Henning et al., 2013).

These covered hidden emotion, sarcasm, and explicit FB (the unexpected transfer task), all of which were scored as pass or fail. (b) Pretested German-language translations of Happé's Strange Stories (following Happé, 1994; White et al., 2009; pretested German translations by Ebert, 2020; Rakoczy et al., 2012, 2018). A total of eight Strange Stories were evenly distributed across four themes: double bluff, white lie, persuasion (deception), and misunderstanding. The target question for each story was either "Why did [they] say this?" or "Why did [they] do this?" Each story was also assigned a neutral (*Kannst du es mir noch ein bisschen genauer sagen?* [Can you say that a bit more precisely for me?]) and mental (e.g., *Was glaubt Simon, was Moritz denkt?* [What does Simon believe Moritz thinks?]) follow-up question. The experimenter used these to clarify the children's responses. Responses could be scored as correct (2 points), partially correct (1 point), or completely incorrect (0 points), with a potential total of 16 points across the eight stories. A second coder independently scored all answers for 28.6% of the children (a randomly selected 10 of the 35 total participants) to calculate interrater agreement (ToM scale: Cohen's κ = 1.0, Strange Stories: Cohen's κ = .92).

The Strange Stories had been pre-recorded by a native speaker, and each audio file was played twice for the children. During this task, the children were shown the same gray neutral screen with a fixation cross as in the metaphor comprehension task. To fit the digital context, we updated the stimuli for both the ToM scale and Strange Stories by making some minor adaptations to the materials. For the ToM scale, we needed to use the screen-sharing function in the digital conferencing tool to display supporting images that would originally have been physically laid out on the table in front of the child. For the Hidden Emotion subtest, two separate pictures had to be merged to form one image on the screen: We centered the picture of the story's main character and positioned it above the depictions of the three facial expressions (happy, in-between, and sad). As the context did not support pointing, we adapted the image so that each of the pictured expressions was surrounded by a different and easily distinguishable colored outline. The children were instead asked to identify the relevant facial expression by the color of its outline. For the Sarcasm subtest, we displayed the original image on the screen but used a spotlight cursor when the experimenter would have pointed to the characters. In the original version of the Explicit FB subtest, a small, toy figurine of a boy would be placed on the table near the picture but between the depictions of the story's two locations. For the digital setting, we altered the original image by positioning a cartoon picture of a boy equidistantly between the two location depictions at the bottom of the screen.

The content of two of the eight Strange Stories (German translations: Ebert, 2020; Rakoczy et al., 2012, 2018) were very minimally updated to the modern context, but the adaptations did not change the stories' target meanings. One persuasion story was changed so that the protagonist instead tried to obtain pieces of pizza (and not mini sausages). One white-lie story was adapted so that the protagonist's disappointing Christmas present was "a pile of books" instead of an old brand of encyclopedia sets that children might not know anymore.

Results

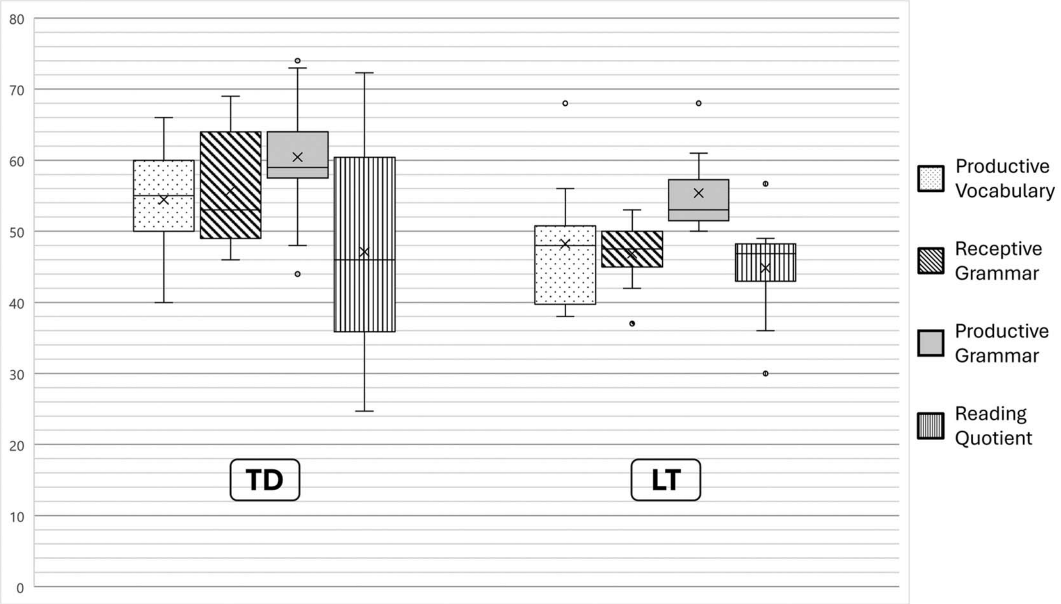
The data were analyzed using IBM SPSS Statistics (Version 29), and the multiple-comparison correction was computed using an online calculator by Hemmerich (2016). We first present the descriptive results of each group across all measures at 9 (see Table 2). Figure 1 depicts box plots for the *T*-score standardized measures across oral language and reading. Table 3 presents the individual performance and language history for each child characterized as an LT at age 2. Looking at Figure 1 as well as the groups' mean performances and score

Table 2. Descriptive results of the typically developing (TD) and late talker (LT) children's mean performance and score range across measures of language, metacognitive, and metalinguistic skills.

Test	TD *M* (score range)	LT *M* (score range)
Productive Vocabulary (TD *n* = 27, LT *n* = 8; *T* score)	54.44 (40–66)	48.25 (38–68)
Productive Grammar (TD *n* = 27, LT *n* = 8; *T* score)	60.44 (44–74)	55.38 (50–68)
Receptive Grammar (TD *n* = 27, LT *n* = 8; *T* score)	55.74 (46–69)	46.88 (37–53)
Reading Score (TD *n* = 26, LT *n* = 8; reading quotient *T* score)	47.12 (24.70–72.30)	44.84 (30.00–56.70)
ToM scale (TD *n* = 26, LT *n* = 8; maximum of 3)	2.12 (1–3)	2.00 (1–3)
Strange Stories (TD *n* = 26, LT *n* = 7; maximum of 16)	12.65 (9–16)	10.00 (4–13)
Metaphor Comprehension (TD *n* = 27, LT *n* = 8; maximum of 22)	10.07 (6–15)	7.00 (1–19)

Note. ToM = theory of mind.

Figure 1. Results of TD and LT children's performance and score range across the standardized measures of language and reading (*T* scores on the *y*-axis, means, medians, interquartile ranges, and outliers presented). TD = typically developing; LT = late talker.



ranges in Table 2, we can see overlaps across all of the domains. Group performances appear more comparable across productive grammar, the RQ, and the ToM scale, but there are clearer gaps in the domains of productive vocabulary, receptive grammar, Strange Stories, and metaphor comprehension. Many measures were collected during the testing sessions, and there are unfortunately a small number of cases with missing data. Sample sizes are therefore reported for each analysis.

We investigated language (productive vocabulary, productive grammar, receptive grammar), reading, ToM (ToM scale and Strange Stories), and metaphor comprehension skills in LT versus TD children in our sample. We conducted nonparametric group comparisons due to the uneven and small group sizes. The results presented in Table 4 show that LTs in our sample performed significantly less well than their TD peers on measures of productive vocabulary, receptive grammar, and metaphor comprehension and in the Strange Stories task. The effect sizes of these four Mann–Whitney *U* tests, measured by the rank-biserial correlation *r* reported in Table 4, were all found to be > 0.4, indicating moderate effects. After applying the Bonferroni-type Benjamini and Hochberg (1995) method to control the false discovery rate when conducting multiple comparisons, we no longer achieved significant group differences on the productive vocabulary measure (see Table 4). The LTs trended slightly toward lower performance on the productive grammar measure, but this did not reach statistical significance. There were also no statistically significant differences in their performance on the ToM scale tasks or the reading measure. We ran post hoc power analyses of the Mann–Whitney tests using the software G*Power (Version 3.1.9.4; Faul et al., 2007) for the subtests where the descriptive data indicated lack of power might conceal group differences (productive vocabulary and productive grammar) as well as for subtests showing clearer differences (receptive grammar, the Strange Stories task, and metaphor comprehension). The following sample sizes would have been required: productive vocabulary (N = 68, TD n = 52, LT n = 16), productive grammar (N = 98, TD n = 76, LT n = 22), receptive grammar (N = 46, TD n = 35, LT n = 11), the Strange Stories task (N = 52, TD n = 41, LT n = 11), and metaphor comprehension (N = 44, TD n = 34, LT n = 10).

Discussion

There is a paucity of work addressing where "recovering" LTs might fall between TD children and children with DLD on potential spectra or continua of language-related competencies. Thus, we aimed to address two questions within our investigation: (1) Are language and reading abilities impacted in (German-speaking) LTs at age 9? (2) Are metacognitive and metalinguistic (ToM and metaphor comprehension) abilities affected in 9-year-old children with a history of being an LT? To answer Research Question 1, our results indicated lower

Table 3. Background and performance of eight children with a history of late talker (LT).

	Child No.							
Variable	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Sex	Male	Female	Female	Male	Female	Female	Male	Male
Family history of language delay	Yes	Yes	No	No	Yes	Yes	No	No
Size of productive vocabulary at 2[a]	23	69	123	95	41	152	28	3
Word comprehension at 2[b]	44	41	69	48	54	38	38	48
Sentence comprehension at 2[b]	26	26	35	35	54	26	35	54
Word production at 2[b]	26	32	33	32	30	36	30	26
Sentence production at 2[b]	30	34	Missing	34	35	36	39	30
Parental guidance[c]	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
Status 2;6	LT	LT	LT	TD	TD	LT	TD	TD
Speech and language therapy	20 units	No	Breakup	—	—	10 units	—	—
Status 3–6	TD	DLD[d]	DLD	TD	TD	TD	TD	TD
Corollary diagnosis: dyslexia	No	No	No	Yes	No	No	No	No
Status 9	TD	TD	TD	TD	TD	TD	TD	TD
Productive vocabulary at 9	48	38	39	49	68	48	56	40
Productive grammar at 9	52	50	52	50	56	68	61	54
Receptive grammar at 9	49	46	37	49	53	42	46	53
Reading score at 9 (converted to *T* score)	45.3	30.0	49.0	36.0	47.7	48.0	46.0	56.7
ToM scale at 9	2	2	1	3	3	2	2	1
False belief	1	1	0	1	1	1	1	1
Hidden emotion	1	1	1	1	1	1	1	0
Sarcasm	0	0	0	1	1	0	0	0
Strange Stories at 9	8	11	4	13	12	11	11	Missing
Metaphor comprehension at 9	5	5	3	6	19	10	7	1

Note. TD = typical development; DLD = developmental language disorder; ToM = theory of mind.

[a]Size of productive vocabulary measured with *Fragebogen zur frühkindlichen Sprachentwicklung* (Szagun et al., 2009). [b]Standard *T* values, measured with *Sprachentwicklungstest für zweijährige Kinder* (Grimm, 2000). [c]Taking part in two units with a speech-language pathologist who explained and demonstrated language beneficial behavior. [d]Did not participate at 5 years old.

performance on the productive vocabulary and receptive grammar measures but not on the productive grammar and reading measures. To answer Research Question 2, performance appeared lower on the metaphor comprehension and the advanced ToM Strange Stories tasks but not on the ToM scale tasks (hidden emotions, sarcasm, and explicit FB). Correcting for multiple comparisons slightly weakened our findings: Group differences in productive vocabulary no longer reached statistical significance. This is likely due to low statistical power resulting from the small number of children with an LT background participating in the study. Despite this limitation,

Table 4. Results of the Mann–Whitney *U* tests comparing language, metacognitive, and metalinguistic skills between late talkers (LTs) and typically developing (TD) children including performance on each theory of mind (ToM) scale task.

	TD		*LTs*						
Variable	***Mdn***	***IQR***	***Mdn***	***IQR***	***U***	***Z***	***p***	***r***	***p****
Productive Vocabulary (TD *n* = 27, LT *n* = 8)	55.0	10	48.0	15	55.0	−2.089	.037	.352	.065
Productive Grammar (TD *n* = 27, LT *n* = 8)	59.0	8	53.0	9	63.5	−1.762	.078	.296	.109
Receptive Grammar (TD *n* = 27, LT *n* = 8)	53.0	15	47.5	9	45.0	−2.514	.012	.418	.042
Reading Score, original RQ (TD *n* = 26, LT *n* = 8)	94.0	43.0	95.25	15.6	102.5	−0.061	.951	.011	.951
ToM scale (TD *n* = 26, LT *n* = 8)	2.0	1	2.0	2	95.0	−0.408	.683	.063	.797
Strange Stories (TD *n* = 26, LT *n* = 7)	13.0	3	11.0	4	38.5	−2.339	.019	.402	.044
Metaphor Comprehension (TD *n* = 27, LT *n* = 8)	10.0	3	5.5	6	44.5	−2.514	.012	.422	.042

Note. Bonferroni-type Benjamini and Hochberg (1995) corrections for multiple comparisons and controlling the false discovery rate are reported under *p**. For the highest level of accuracy, the original reading quotient for the reading score is used instead of our *T*-score conversion. RQ = reading quotient.

our findings still provide an interesting starting point; studies with larger sample sizes should further investigate ToM and metaphor comprehension alongside language skills in LT children who appear to have "caught up" with their TD peers. We will now address each skill and aim to contemplate (a) why we might have found these differences and (b) how our findings might fit within the existing body of research and its relevant theoretical considerations.

Language

We found that LTs seem to catch up with their TD peers in some language skills (productive grammar) but that subtle differences potentially remain across others (vocabulary, receptive grammar). These findings are partially in line with prior work. Rescorla (2002, 2005, 2009) found that LTs performed in the average range on all language tasks at ages 9, 13, and 17 years but that their factored scores continued to be significantly lower in vocabulary and grammar, indicating that early delays can have an enduring impact on developmental outcomes. It is not immediately clear why we found no differences in productive grammar, but it may be due to the task design of our measures. Rescorla (2009) found that nonlanguage skills measured at 2 years of age explained some of the variance in her longitudinal sample's performance on vocabulary, grammar, and verbal memory measures. These nonlanguage skills included nonverbal cognitive abilities tapped by tests involving blocks, puzzles, pegs, drawing, and object hiding. They also covered general test-taking skills such as attention, cooperation, and persistence. When these skills were measured at 17 years of age, the LTs scored in the average range but still significantly lower than TD peers (Rescorla, 2009). This could potentially explain our difference with prior work. Our productive grammar task may have tapped different underlying skills than (or may not have been as sensitive as) those used in Rescorla's longitudinal study. Future studies comparing LTs to their TD peers and peers with DLD should carefully consider their task designs and ideally test multiple measures for each language skill to ensure greater sensitivity.

Language Within a TD–LT–DLD Spectrum

Given our very small LT sample size, individual differences in this particular group might have meant that they happened to be slightly stronger than other LTs in productive grammar skills, leading to no statistically significant difference with the TD group. This would align with work identifying different subtypes or severities within LTs (Thal et al., 2013) and problems with using LT as a binary clinical category (Dollaghan, 2013). It would also fit with prior theories (Bates et al., 1995; Dollaghan, 2013; Rescorla, 2009, 2011; Thal et al., 2013) regarding a dimensional account of language delay where individual LT and TD children differ quantitatively along a language ability spectrum. Considering the post hoc power analyses, it is, however, also possible that the group differences in productive grammar might be so small that our sample size was insufficient to identify it.

Reading

We found no significant difference in reading comprehension between LT and TD groups at 9 years of age. This contrasts with some prior work: LTs in Rescorla's (2002, 2005) longitudinal sample scored significantly lower on reading comprehension at 13 years of age but showed no difference at ages 6 or 7 years and only began to show significant differences when they were tested at ages 8 and 9 years. Differences in task design and demands could be a factor here, but another point also worth considering is cross-linguistic difference in orthographic transparency. English is an orthographically less transparent language than German (the language of the participants in our study). It is possible that differences in reading comprehension performance may become evident earlier in English readers than German readers (Borleffs et al., 2019; Diamanti et al., 2018). Prior cross-linguistic work with English- and German-speaking dyslexic children aged 10–12 years found that the English-speaking group was more greatly impaired in reading accuracy than the German-speaking one (Landerl et al., 1997). This might explain why we found no significant difference in our sample's performance at 9 years of age. Our participants might not have yet reached a developmental point where underlying differences in language(-adjacent) skills had begun to accelerate disparities in reading comprehension outcomes. Alternatively, the task design might need to be more complex and possess enhanced sensitivity to identify any differences between German-speaking TD and LT children. These possibilities should definitely be investigated further in future work with longitudinal samples of German-speaking LTs.

Dyslexia

Some prior work (Psyridou et al., 2018) has identified that a family history of dyslexia is a stronger predictor than LT status for reading comprehension performance. We would like to note here that our analyses could have been based on partial information. In our sample of 34 children, one of our eight LTs and three of our 26 TD children had been diagnosed with dyslexia. However, six (only one of whom was an LT) of the 30 nondyslexic children performed more poorly than the highest performing dyslexic child. The other six nondyslexic LTs' reading scores were clustered at or above the median. Some of the children might possibly have had undiagnosed dyslexia and might go on to receive a diagnosis in future. Once the study finished, all of the participating children's parents were informed about their child's comparative performances across the tasks. They were given advice for how to support their child's development in these areas so that they could seek further assistance or assessment.

Reading Within a TD–LT–DLD Spectrum

Psyridou et al. (2018) noted that only LTs who had experienced both expressive and receptive language delay (late comprehender group in Thal et al., 2013) went on to face difficulties with reading comprehension, and not LTs who had experienced expressive without receptive language delay (late producer group in Thal et al., 2013). Bishop and Edmundson's (1987) and Stothard et al.'s (1998) longitudinal sample performed similarly. The "good-outcome" LT group (language difficulties appeared to resolve) had a lower median performance than the controls on verbal comprehension and reading, while the "poor-outcome" LT group had an even lower median performance (Stothard et al., 1998). This could support the idea that one subgroup of LTs is more likely to experience challenges in meaning-based decoding (Thal et al., 2013). Unfortunately, this potential split cannot be addressed within our study due to the small LT sample size; however, future work should account for these subgroupings when considering measures that tap meaning-based decoding.

Advanced Language–Related Skills: ToM and Metaphor Comprehension

We found that LTs appeared to perform less well than their TD peers on the Strange Stories and metaphor comprehension tasks but not on the ToM scale (hidden emotions, sarcasm, and explicit FB). The three ToM scale subtests could only be scored as pass or fail, so it is difficult to build up a more nuanced picture of the abilities tested and their age-appropriateness. The explicit FB and hidden emotions tasks were actually designed so that much younger children (around 4–5 years of age) could pass them. Our main reason for including them in our test battery was to conduct a baseline assessment of children's abilities. Prior research (cf. Peterson et al., 2012) has also shown that sarcasm is still emerging at 9 years of age, so we may have assessed it slightly too early in our sample. Testing it a little later on when it is more established might better reflect each group's abilities.

ToM and Metaphor Comprehension Within a TD–LT–DLD Spectrum

To our knowledge, no other studies have investigated ToM and metaphor comprehension abilities in LTs who appear to "recover." We can, therefore, only attempt to fit our findings somewhere between and within the existing work on TD children and children with DLD. There are different possibilities to explore in seeking an explanation for the LTs' comparatively lower performance on the Strange Stories and metaphor comprehension tasks. One starting point could be the LTs' lower performance on vocabulary and receptive grammar measures. Prior work has associated these abilities with performance on ToM and metaphor comprehension tasks (see e.g., Deckert et al., 2019; Milligan et al., 2007; Osterhaus & Koerber, 2021). Combining the ideas of a language endowment spectrum (Ellis Weismer, 2007) and language ability as a dimensional spectrum along which individuals differ (Rescorla, 2009, 2011), we could further incorporate language-associated ToM and metaphor comprehension skills. It would be logical for LTs without a DLD diagnosis to fall somewhere along the spectrum (or continuum) between their TD peers and peers with DLD in these competencies too.

Potential Differences in ToM Task Demands

LTs who do go on to receive a DLD diagnosis have been found to struggle with language processing (Sansavini et al., 2021). It is therefore also possible that our audio-based Strange Stories and metaphor comprehension tasks may have been harder for the LTs to process than the ToM scale tasks, which were supported visually. However, Nilsson and de Lopez (2016) have reviewed numerous studies in which children with DLD showed ToM impairments even when completing tasks with lower verbal demands. This suggests that at least for DLD, verbal processing demands cannot necessarily explain away ToM impairments. Nevertheless, rather than the "social–perceptual" component of ToM (Tager-Flusberg and Joseph, 2005), Nilsson and de Lopez proposed that children with DLD's ToM delays are rooted in the "socio-cognitive" ToM component (Tager-Flusberg & Joseph, 2005). This component supposedly enables inference making about mental phenomena through the integration of different kinds of information. This could tie into both a processing-deficit perspective on LTs and children with DLD as well as the idea of a language and language-related ability spectrum or continuum. Relevant for DLD-adjacent LTs and irrespective of a ToM deficit versus delay, Nilsson and de Lopez importantly highlight that potential impairments early on could become enduring characteristics that affect continued social development and outcomes.

Metaphor Comprehension and Processing

Work has shown that metaphor comprehension is important for peer social interaction outcomes (Del Sette et al., 2021) and that it is linked to both language (e.g., Deckert et al., 2019) and ToM ability (Norbury, 2005). Understanding ambiguous metaphors can require thinking about a speaker's intended meaning, and some metaphors in themselves contain mental aspects (Lecce et al., 2019). LTs' lower performance on the metaphor comprehension task would be consistent with prior work investigating metaphor comprehension in DLD populations (Bühler et al., 2018; Norbury, 2005; Spanoudis, 2016). While task demands in terms of verbal processing could have been a contributing factor, the findings from Lorusso et al.'s (2015) study using event-related potentials (ERPs) could also provide some interesting insights. Lorusso et al. compared children with DLD, children with nonverbal learning disabilities, and a control TD group (aged 6–15 years, matched for both gender and chronological age) on the processing of literal versus figurative uses of verbs (e.g., *I picked up a flower* vs. *I picked up an idea*). This is analogous to metaphor comprehension. They found that the children with DLD showed the lowest accuracy and the most deviant ERP patterns when processing abstract and figurative expressions (Lorusso et al., 2015). They also performed worse on abstract sentences than the participants with nonverbal learning disabilities (Lorusso et al., 2015). The authors argued that this lent "support to theories assuming that linguistic processes are crucial to the representation and comprehension of abstract language" (Lorusso et al., 2015, p. 17). These results are meaningful for our findings regarding LTs, reinforcing the idea that underlying impairments in language ability may critically disrupt non-literal language processing.

Looking to the Future

As children grow older, metaphor comprehension and ToM skills become increasingly important for academic and social success. Our findings indicate that LTs may need additional support with these competencies to ensure they do not fall behind their TD peers at school. However, despite these potential deficits, the highest scorer on the metaphor comprehension task in our study was actually an LT (late producer but not comprehender; see Tables 2 and 3). This highlights that rather than a binary LT status, more nuance is required for advanced metalinguistic skills. To develop appropriate, tailored support for LTs, future work needs to go beyond basic differences in performance outcomes. It must address remaining questions about whether (a) metalinguistic and metacognitive competencies are impacted simply as a result of impaired language skills, or (b) language and these competencies share common underlying deficits, or even (c) there are associated but differing impairments across both these competencies and language ability.

Limitations

Despite our longitudinal perspective and longstanding interaction with the participating children that strengthens this study's findings, we must address some limitations. First, we recognize that this study is based on a small sample size (N = 35) and that the participating groups of TD and LT children are uneven (N = 35, TD n = 27, LT n = 8) with a very small number of participating LTs. A definitive generalization from our results is consequently not possible. Instead they provide initial insights into whether LTs perform differently compared to their TD peers on metalinguistic (metaphor comprehension) and advanced metacognitive (Strange Stories) tasks. More conclusive work is needed with matched, larger, and even groups of participants. Each target ability should be tested using a variety of measures to ensure that contrasting findings do not arise from task differences between studies.

Second, although all previous testing points in our wider longitudinal study were carried out in person, we had to adapt in response to the COVID-19 pandemic and implement a hybrid testing method. Two sessions were carried out digitally, and one session was carried out in person. Prior to the pandemic, it was fairly uncommon to use digital testing rather than the standard in-person experimental method. This could perhaps weaken our results. There are challenges involved in using online testing methods (Braun et al., 2020) as well as tools for research such as Zoom (Gray et al., 2020). However, there are also early indications that studies carried out via Zoom with young children (Escudero et al., 2021) can still be an appropriate, reliably comparable, alternative measure to testing in-person. Of course, further evaluation is still needed. Some additional, possibly mitigating factors include the highly positive reception of our hybrid method by the participating children and families, as well as the longstanding in-person context and relationship built up with them by the university department. This context afforded us a well-established developmental picture of the children's abilities, lending weight to the potential reliability of our findings.

Implications and Future Opportunities

Our work indicates that there may be a spectrum and even a complex continuum of language ability and associated skills, such as metalinguistic and metacognitive competencies. LTs can move past deficits in performance, which would lead to a classification of "disorder" and appear to catch up, yet still demonstrate discernible differences from TD children many years later. Metalinguistic and metacognitive competencies are important for effective social interaction with peers and enable engagement with more complex academic tasks like the study of literature. Our findings suggest that even LTs who appear to "recover" might benefit from targeted, extra support. This could be provided informally by caregivers or more formally within intervention programs or in a school context.

Future comparative research should investigate these skills within larger-sized samples, integrating children's history of language development and building (sub)groups of TD, LT, and DLD children. More finely grained approaches are needed to identify which aspects of advanced metacognitive and metalinguistic abilities are specifically impacted in these groups, along with their associated linguistic competencies and cognitive functions.

Data Availability Statement

The data set generated for this study cannot be made publicly available as participants did not provide consent for that; however, private enquiries to view an anonymized version of the data set are welcomed and can be addressed to the corresponding author.

Acknowledgments

We thank the German Research Foundation (DFG) for funding the earlier stages of this longitudinal study between the years 2012 and 2018 under Grants RI 898/6-1 and RI 898/6-2 to U.R. and Grant LU 2116/1-2 to C.L., which was then extended by internal money from the authors' university department. We are very grateful to all the families who have continued to participate over these many years as well as to our collaborators Angela Grimminger, Katharina Rohlfing, and Ulf Liszkowski, who were involved in the project during previous stages. We are also very thankful for all the research assistants who over the years have assisted with or have been involved in carrying out our longitudinal study, the preparation of materials, or in the transcription and coding process: Jannika Böse, Annika Büssing, Olivia Espeter, Tabea Musshoff, Katharina Rademacher, Kathrin Gremplewski, Lisa Grundhöfer, Inga Hackbarth, Bianca Henze, Janina Jurjahn, Miriam Krimer, Carina Lucas, Laura Polke, Anika Queißer, Suzan Ryschka, and Annika Biewener.

References

Andrés-Roqueta, C., Adrian, J. E., Clemente, R. A., & Katsos, N. (2013). Which are the best predictors of theory of mind delay in children with specific language impairment? *International Journal of Language & Communication Disorders, 48*(6), 726–737. https://doi.org/10.1111/1460-6984.12045

Astington, J., & Baird, J. (2005). Introduction: Why language matters. In J. Astington & J. Baird (Eds.), *Why language matters for theory of mind* (pp. 3–25). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195159912.003.0001

Atkinson, L., Slade, L., Powell, D., & Levy, J. P. (2017). Theory of mind in emerging reading comprehension: A longitudinal study of early indirect and direct effects. *Journal of Experimental Child Psychology, 164,* 225–238. https://doi.org/10.1016/j.jecp.2017.04.007

Babarczy, A., Balázs, A., & Krizsai, F. (2019). Preschoolers' metaphor comprehension. Methodological issues in experimental pragmatics. *Acta Universitatis Sapientiae, Philologica, 11*(2), 133–150. https://doi.org/10.2478/ausp-2019-0017

Bates, E., Dale, P., & Thal, D. (1995). Individual differences and their implications for theories of language development. In P. Fletcher & B. MacWhinney (Eds.), *Handbook of child language* (pp. 96–151). Basil Blackwell.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bishop, D. V. M. (2017). Why is it so hard to reach agreement on terminology? The case of developmental language disorder (DLD). *International Journal of Language & Communication Disorders, 52*(6), 671–680. https://doi.org/10.1111/1460-6984.12335

Bishop, D. V. M., & Edmundson, A. (1987). Language-impaired 4-year-olds: Distinguishing transient from persistent impairment. *Journal of Speech and Hearing Disorders, 52*(2), 156–173. https://doi.org/10.1044/jshd.5202.156

Bishop, D. V. M., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & Catalise-2 Consortium. (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/10.1111/jcpp.12721

Boerma, I. E., Mol, S. E., & Jolles, J. (2017). The role of home literacy environment, mentalizing, expressive verbal ability, and print exposure in third and fourth graders' reading comprehension. *Scientific Studies of Reading, 21*(3), 179–193. https://doi.org/10.1080/10888438.2016.1277727

Borleffs, E., Maassen, B. A., Lyytinen, H., & Zwarts, F. (2019). Cracking the code: The impact of orthographic transparency and morphological-syllabic complexity on reading and developmental dyslexia. *Frontiers in Psychology, 9,* Article 2534. https://doi.org/10.3389/fpsyg.2018.02534

Braun, R., Blok, V., Loeber, A., & Wunderle, U. (2020). COVID-19 and the onlineification of research: Kick-starting a dialogue on responsible online research and innovation (RoRI). *Journal of Responsible Innovation, 7*(3), 680–688. https://doi.org/10.1080/23299460.2020.1789387

Bühler, D., Perovic, A., & Pouscoulous, N. (2018). Comprehension of novel metaphor in young children with developmental language disorder. *Autism & Developmental Language Impairments, 3,* 1–11. https://doi.org/10.1177/2396941518817229

Crawshaw, C. E., Lüke, C., & Ritterfeld, U. (2024). *Does early gesture usage contribute alongside oral language to later theory of mind performance and metaphor comprehension? Indications from a longitudinal study with children aged 1–9 years* [Manuscript submitted for publication]. Research Unit of Language and Communication, Department of Rehabilitation Sciences, TU Dortmund University, Germany.

Deckert, M., Schmoeger, M., Schaunig-Busch, I., & Willinger, U. (2019). Metaphor processing in middle childhood and at the transition to early adolescence: The role of chronological age, mental age, and verbal intelligence. *Journal of Child Language, 46*(2), 334–367. https://doi.org/10.1017/S0305000918000491

Del Sette, P., Ronchi, L., Bambini, V., & Lecce, S. (2021). Longitudinal associations between metaphor understanding and peer relationships in middle childhood. *Infant and Child Development, 30*(4), Article e2232. https://doi.org/10.1002/icd.2232

De Mulder, H. N. M., Wijnen, F., & Coopmans, P. H. A. (2019). Interrelationships between theory of mind and language development: A longitudinal study of Dutch-speaking kindergartners. *Cognitive Development, 51,* 67–82. https://doi.org/10.1016/j.cogdev.2019.03.006

de Villiers, J. (2007). The interface of language and theory of mind. *Lingua, 117*(11), 1858–1878. https://doi.org/10.1016/j.lingua.2006.11.006

de Villiers, P. A. (2005). The role of language in theory-of-mind development: What deaf children tell us. In J. Astington & J. Baird (Eds.), *Why language matters for theory of mind* (pp. 266–297). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195159912.003.0013

Devine, R. T., & Lecce, S. (Eds.). (2021). *Theory of mind in middle childhood and adolescence: Integrating multiple perspectives*. Routledge. https://doi.org/10.4324/9780429326899

Devine, R. T., White, N., Ensor, R., & Hughes, C. (2016). Theory of mind in middle childhood: Longitudinal associations with executive function and social competence. *Developmental Psychology, 52*(5), 758–771. https://doi.org/10.1037/dev0000105

Diamanti, V., Goulandris, N., Campbell, R., & Protopapas, A. (2018). Dyslexia profiles across orthographies differing in transparency: An evaluation of theoretical predictions contrasting English and Greek. *Scientific Studies of Reading, 22*(1), 55–69. https://doi.org/10.1080/10888438.2017.1338291

Doherty, M. (2008). *Theory of mind: How children understand others' thoughts and feelings*. Psychology Press. https://doi.org/10.4324/9780203929902

Dollaghan, C. (2013). Late talker as a clinical category: A critical evaluation. In L. A. Rescorla & P. S. Dale (Eds.), *Late talkers: Language development, interventions, and outcomes* (pp. 91–112). Brookes.

Durrleman, S., Burnel, M., & Reboul, A. (2017). Theory of mind in SLI revisited: Links with syntax, comparisons with ASD. *International Journal of Language & Communication Disorders, 52*(6), 816–830. https://doi.org/10.1111/1460-6984.12317

Durrleman, S., & Delage, H. (2020). Training complements for belief reasoning in developmental language disorder. *Journal of Speech, Language, and Hearing Research, 63*(6), 1861–1877. https://doi.org/10.1044/2020_JSLHR-19-00075

Ebert, S. (2020). Theory of mind, language, and reading: Developmental relations from early childhood to early adolescence. *Journal of Experimental Child Psychology, 191,* Article 104739. https://doi.org/10.1016/j.jecp.2019.104739

Ellis Weismer, S. (2007). Typical talkers, late talkers, and children with specific language impairment: A language endowment spectrum? In R. Paul (Ed.), *The influence of developmental perspectives on research and practice in communication disorders: A festschrift for Robin S. Chapman* (pp. 83–102). Erlbaum.

Ensor, R., Devine, R. T., Marks, A., & Hughes, C. (2014). Mothers' cognitive references to 2-year-olds predict theory of mind at ages 6 and 10. *Child Development, 85*(3), 1222–1235. https://doi.org/10.1111/cdev.12186

Escudero, P., Pino Escobar, G., Casey, C. G., & Sommer, K. (2021). Four-year-old's online versus face-to-face word learning via eBooks. *Frontiers in Psychology, 12,* Article 450. https://doi.org/10.3389/fpsyg.2021.610975

Esser, G., Wyschkon, A., Ballaschk, K., & Hänsch, S. (2010). *Potsdam-Illinois Test für Psycholinguistische Fähigkeiten (P-ITPA)* [Potsdam-Illinois Test for Psycholinguistic Abilities (P-ITPA)]. Hogrefe.

Farrant, B. M. (2015). Specific language impairment and perspective taking: Delayed development of theory of mind, visual and emotional perspective taking. *Journal of Childhood & Developmental Disorders, 1*(1), 1–8. https://doi.org/10.4172/2472-1786.100008

Farrant, B. M., Fletcher, J., & Maybery, M. T. (2006). Specific language impairment, theory of mind, and visual perspective taking: Evidence for simulation theory and the developmental role of language. *Child Development, 77*(6), 1842–1853. https://doi.org/10.1111/j.1467-8624.2006.00977.x

Farrar, M. J., Johnson, B., Tompkins, V., Easters, M., Zilisi-Medus, A., & Benigno, J. P. (2009). Language and theory of mind in preschool children with specific language impairment. *Journal of Communication Disorders, 42*(6), 428–441. https://doi.org/10.1016/j.jcomdis.2009.07.001

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/bf03193146

Fox, A. V. (2013). *TROG-D. Test zur Überprüfung des Grammatikverständnisses* [Test for Assessing Grammatical Understanding] (6th ed.). Schulz-Kirchner.

Gillott, A., Furniss, F., & Walter, A. (2004). Theory of mind ability in children with specific language impairment. *Child Language Teaching and Therapy, 20*(1), 1–11. https://doi.org/10.1191/0265659004ct260oa

Gordon Pershey, M. (2000). Children's elicited use of pragmatic language functions: How six- and seven-year-old children adapt to the interactional environments of story scenarios. *Language Awareness, 9*(4), 218–235. https://doi.org/10.1080/09658410008667147

Gray, L. M., Wong-Wylie, G., Rempel, G. R., & Cook, K. (2020). Expanding qualitative research interviewing strategies: Zoom video communications. *The Qualitative Report, 25*(5), 1292–1301. https://doi.org/10.46743/2160-3715/2020.4212

Grimm, H. (2000). *SETK-2. Sprachentwicklungstest für zweijährige Kinder* [Language Acquisition Test for 2-Year-Old Children]. Hogrefe.

Grimm, H. (2001). *Sprachentwicklungstest für drei- bis fünfjährige Kinder (SETK 3–5): Diagnose von Sprachverarbeitungsfähigkeiten und auditiven Gedächtnisleistungen* [Language Acquisition Test for 3- to 5-Year-Old Children]. Hogrefe.

Hammer, C. S., Morgan, P., Farkas, G., Hillemeier, M., Bitetti, D., & Maczuga, S. (2017). Late talkers: A population-based study of risk factors and school readiness consequences. *Journal of Speech, Language, and Hearing Research, 60*(3), 607–626. https://doi.org/10.1044/2016_JSLHR-L-15-0417

Happé, F. G. E. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders, 24*(2), 129–154. https://doi.org/10.1007/BF02172093

Happé, F. G. E. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Development, 66*(3), 843–855. https://doi.org/10.1111/j.1467-8624.1995.tb00909.x

Hemmerich, W. (2016). *Rechner zur Adjustierung des α-Niveaus* [StatisticGuru: Calculator for adjusting the α level]. StatistikGuru. https://statistikguru.de/rechner/adjustierung-des-alphaniveaus.html

Henning, A., Hofer, T., & Aschersleben, G. (2013). *Erweiterte "Theory of Mind-Skala" für 3- bis 11-jährige Kinder* [Extended Theory-Of-Mind Scale for 3- to 11-year-old children] (2nd ed.). Universität des Saarlandes Arbeitseinheit Entwicklungspsychologie.

Hjetland, H. N., Lervåg, A., Lyster, S. A. H., Hagtvet, B. E., Hulme, C., & Melby-Lervåg, M. (2019). Pathways to reading comprehension: A longitudinal study from 4 to 9 years of age. *Journal of Educational Psychology, 111*(5), 751–763. https://doi.org/10.1037/edu0000321

Horvath, S., Kueser, J. B., Kelly, J., & Borovsky, A. (2022). Difference or delay? Syntax, semantics, and verb vocabulary development in typically developing and late-talking toddlers. *Language Learning and Development, 18*(3), 352–376. https://doi.org/10.1080/15475441.2021.1977645

Kalandadze, T., Norbury, C., Nærland, T., & Næss, K. A. B. (2018). Figurative language comprehension in individuals with autism spectrum disorder: A meta-analytic review. *Autism, 22*(2), 99–117. https://doi.org/10.1177/1362361316668652

Kauschke, C., & Siegmüller, J. (2010). *Patholinguistische Diagnostik bei Sprachentwicklungsstörungen (PDSS)* [Patholinguistic Diagnostics for Developmental Language Disorder] (2nd ed.). Urban & Fischer.

Landerl, K., Wimmer, H., & Frith, U. (1997). The impact of orthographic consistency on dyslexia: A German–English comparison. *Cognition, 63*(3), 315–334. https://doi.org/10.1016/S0010-0277(97)00005-X

Language and Reading Research Consortium., & **Chiu, Y. D**. (2018). The simple view of reading across development: Prediction of Grade 3 reading comprehension from prekindergarten skills. *Remedial and Special Education, 39*(5), 289–303. https://doi.org/10.1177/0741932518762055

Lecce, S., Ronchi, L., Del Sette, P., Bischetti, L., & Bambini, V. (2019). Interpreting physical and mental metaphors: Is theory of mind associated with pragmatics in middle childhood? *Journal of Child Language, 46*(2), 393–407. https://doi.org/10.1017/S030500091800048X

Levorato, M. C., Nesi, B., & Cacciari, C. (2004). Reading comprehension and understanding idiomatic expressions: A developmental study. *Brain and Language, 91*(3), 303–314. https://doi.org/10.1016/j.bandl.2004.04.002

Lorusso, M. L., Burigo, M., Borsa, V., & Molteni, M. (2015). Processing sentences with literal versus figurative use of verbs: An ERP study with children with language impairments, nonverbal impairments, and typical development. *Behavioural Neurology, 2015,* 1–21. https://doi.org/10.1155/2015/475271

Lüke, C., Grimminger, A., Rohlfing, K. J., Liszkowski, U., & Ritterfeld, U. (2017). In infants' hands: Identification of preverbal infants at risk for primary language delay. *Child Development, 88*(2), 484–492. https://doi.org/10.1111/cdev.12610

Lüke, C., Kauschke, C., Dohmen, A., Haid, A., Leitinger, C., Männel, C., & Neumann, K. (2023). Definition and terminology of developmental language disorders—Interdisciplinary consensus across German-speaking countries. *PLOS ONE, 18*(11), Article e0293736. https://doi.org/10.1371/journal.pone.0293736

Lüke, C., Ritterfeld, U., Grimminger, A., Rohlfing, K. J., & Liszkowski, U. (2020). Integrated communication system: Gesture and language acquisition in typically developing children and children with LD and DLD. *Frontiers in Psychology, 11,* Article 118. https://doi.org/10.3389/fpsyg.2020.00118

Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development, 78*(2), 622–646. https://doi.org/10.1111/j.1467-8624.2007.01018.x

Nilsson, K. K., & de Lopez, K. J. (2016). Theory of mind in children with specific language impairment: A systematic review and meta-analysis. *Child Development, 87*(1), 143–153. https://doi.org/10.1111/cdev.12462

Norbury, C. F. (2005). The relationship between theory of mind and metaphor: Evidence from children with language impairment and autistic spectrum disorder. *British Journal of Developmental Psychology, 23*(3), 383–399. https://doi.org/10.1348/026151005X26732

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

Osterhaus, C., & Koerber, S. (2021). The development of advanced theory of mind in middle childhood: A longitudinal study from age 5 to 10 years. *Child Development, 92*(5), 1872–1888. https://doi.org/10.1111/cdev.13627

Özçalışkan, Ş. (2005). On learning to draw the distinction between physical and metaphorical motion: Is metaphor an early emerging cognitive and linguistic capacity? *Journal of Child Language, 32*(2), 291–318. https://doi.org/10.1017/S0305000905006884

Paul, R., & Ellis Weismer, S. (2013). Late talking in context: The clinical implications of delayed language development. In L. Rescorla & P. S. Dale (Eds.), *Late talkers: Language development, interventions, and outcomes* (pp. 203–217). Brookes.

Perry, L. K., Kucker, S. C., Horst, J. S., & Samuelson, L. K. (2023). Late bloomer or language disorder? Differences in toddler vocabulary composition associated with long-term language outcomes. *Developmental Science, 26*(4), Article e13342. https://doi.org/10.1111/desc.13342

Peterson, C. C., Wellman, H. M., & Slaughter, V. (2012). The mind behind the message: Advancing theory-of-mind scales for typically developing children, and those with deafness, autism, or Asperger syndrome. *Child Development, 83*(2), 469–485. https://doi.org/10.1111/j.1467-8624.2011.01728.x

Pronina, M., Prieto, P., Bischetti, L., & Bambini, V. (2023). Expressive pragmatics and prosody in young preschoolers are more closely related to structural language than to mentalizing. *Language Learning and Development, 19*(3), 323–344. https://doi.org/10.1080/15475441.2022.2074852

Psyridou, M., Eklund, K., Poikkeus, A. M., & Torppa, M. (2018). Reading outcomes of children with delayed early vocabulary: A follow-up from age 2–16. *Research in Developmental Disabilities, 78,* 114–124. https://doi.org/10.1016/j.ridd.2018.05.004

Rakhlin, N., Kornilov, S. A., Reich, J., Babyonyshev, M., Koposov, R. A., & Grigorenko, E. L. (2011). The relationship between syntactic development and theory of mind: Evidence from a small-population study of a developmental language disorder. *Journal of Neurolinguistics, 24*(4), 476–496. https://doi.org/10.1016/j.jneuroling.2011.03.001

Rakoczy, H., Harder-Kasten, A., & Sturm, L. (2012). The decline of theory of mind in old age is (partly) mediated by developmental changes in domain-general abilities. *British Journal of Psychology, 103*(1), 58–72. https://doi.org/10.1111/j.2044-8295.2011.02040.x

Rakoczy, H., Wandt, R., Thomas, S., Nowak, J., & Kunzmann, U. (2018). Theory of mind and wisdom: The development of different forms of perspective-taking in late adulthood. *British Journal of Psychology, 109*(1), 6–24. https://doi.org/10.1111/bjop.12246

Rescorla, L. (2002). Language and reading outcomes to age 9 in late-talking toddlers. *Journal of Speech, Language, and Hearing Research, 45,* 360–371. https://doi.org/10.1044/1092-4388(2002/028)

Rescorla, L. (2005). Age 13 language and reading outcomes in late-talking toddlers. *Journal of Speech, Language, and Hearing Research, 48*(2), 459–472. https://doi.org/10.1044/1092-4388(2005/031)

Rescorla, L. (2009). Age 17 language and reading outcomes in late-talking toddlers: Support for a dimensional perspective on language delay. *Journal of Speech, Language, and Hearing*

Research, 52(1), 16–30. https://doi.org/10.1044/1092-4388(2008/07-0171)

Rescorla, L. (2011). Late talkers: Do good predictors of outcome exist? *Developmental Disabilities Research Reviews, 17*(2), 141–150. https://doi.org/10.1002/ddrr.1108

Rescorla, L. (2013). Late-talking toddlers: A 15-year follow-up. In L. A. Rescorla & P. S. Dale (Eds.), *Late talkers: Language development, interventions, and outcomes* (pp. 219–239). Brookes.

Rescorla, L., Roberts, J., & Dahlsgaard, K. (1997). Late talkers at 2. *Journal of Speech, Language, and Hearing Research, 40*(3), 556–566. https://doi.org/10.1044/jslhr.4003.556

Sansavini, A., Favilla, M. E., Guasti, M. T., Marini, A., Millepiedi, S., Di Martino, M. V., & Lorusso, M. L. (2021). Developmental language disorder: Early predictors, age for the diagnosis, and diagnostic tools. A scoping review. *Brain Sciences, 11*(5), 654–691. https://doi.org/10.3390/brainsci11050654

Smit, L., Knoors, H., Hermans, D., Verhoeven, L., & Vissers, C. (2019). The interplay between theory of mind and social emotional functioning in adolescents with communication and language problems. *Frontiers in Psychology, 10,* Article 1488. https://doi.org/10.3389/fpsyg.2019.01488

Spanoudis, G. (2016). Theory of mind and specific language impairment in school-age children. *Journal of Communication Disorders, 61,* 83–96. https://doi.org/10.1016/j.jcomdis.2016.04.003

Stothard, S. E., Snowling, M. J., Bishop, D. V. M., Chipchase, B. B., & Kaplan, C. A. (1998). Language-impaired preschoolers: A follow-up into adolescence. *Journal of Speech, Language, and Hearing Research, 41*(2), 407–418. https://doi.org/10.1044/jslhr.4102.407

Szagun, G., Stumper, B., and Schramm, S. A. (2009). *FRAKIS. Fragebogen zur frühkindlichen Sprachentwicklung [Questionnaire for Early Language Acquisition]*. Pearson.

Tager-Flusberg, H., & Joseph, R. M. (2005). How language facilitates the acquisition of false belief understanding in children with autism. In J. W. Astington & J. A. Baird (Eds.), *Why language matters for theory of mind* (pp. 298–319). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195159912.003.0014

Tellegen, P. J., Laros, J. A., & Petermann, F. (2007). *SON-R 2 1/2–7: Non-Verbaler Intelligenztest* [Non-Verbal Intelligence Test]. Hogrefe.

Thal, D. J., Marchman, V. A., & Tomblin, J. B. (2013). Late-talking toddlers: Characterization and prediction of continued delay. In L. A. Rescorla & P. S. Dale (Eds.), *Late talkers: Language development, interventions, and outcomes* (pp. 169–201). Brookes.

Vissers, C., & Koolen, S. (2016). Theory of mind deficits and social emotional functioning in preschoolers with specific language impairment. *Frontiers in Psychology, 7,* 157–171. https://doi.org/10.3389/fpsyg.2016.01734

Vogt, S., & Indefrey, P. (2017). *Metaphernerwerb: eine empirische Studie bei Kindern im Alter von sechs bis vierzehn Jahren* [Metaphor acquisition: An empirical study with children between the ages of six to fourteen years]. *Metaphorik.de, 27*(2017), 69–106. https://www.metaphorik.de/de/journal/27/metaphernerwerb-eine-empirische-studie-bei-kindern-im-alter-von-sechs-bis-vierzehn-jahren.html

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development, 75*(2), 523–541. https://doi.org/10.1111/j.1467-8624.2004.00691.x

White, S., Hill, E., Happé, F., & Frith, U. (2009). Revisiting the strange stories: Revealing mentalizing impairments in autism. *Child Development, 80*(4), 1097–1117. https://doi.org/10.1111/j.1467-8624.2009.01319.x

Wimmer, H., & Mayringer, H. (2014). *Salzburger Lese-Screening für die Schulstufen 2–9: SLS 2–9* [Salzburg reading–screening test for school levels 2–9: SLS 2–9]. Huber.

Zubrick, S. R., Taylor, C. L., Rice, M. L., & Slegers, D. W. (2007). Late language emergence at 24 months: An epidemiological study of prevalence, predictors, and covariates. *Journal of Speech, Language, and Hearing Research, 50*(6), 1562–1592. https://doi.org/10.1044/1092-4388(2007/106)