Learning Words from Stories

How Method of Story Delivery and Questioning Styles Influence Children’s Vocabulary Learning

Inaugural-Dissertation

zur Erlangung der Doktorwürde der

Fakultät für Humanwissenschaften

der

Julius-Maximilians-Universität Würzburg

Vorgelegt von

Jan Niklas Peter Lenhart

aus Uffenheim

Würzburg

2019

Erstgutachter: Prof. Dr. Wolfgang Lenhard

Zweitgutachter: Prof. Dr. Wolfgang Schneider

Tag der Disputation: 16. Juli 2019

Danksagung

Die vier Studien, die dieser teilkumulativen Dissertation zugrunde liegen, wurden im Rahmen des DFG-Projekts „Experimentelle und längsschnittliche Untersuchung der Effekte von Sprachinterventionen im Vorschulalter bei Kindern mit einem Risiko für die Entwicklung schulischer Leistungsprobleme“ (LE 2680/4-1) durchgeführt. Ich möchte mich daher an erster Stelle bei meinem Betreuer Prof. Dr. Wolfgang Lenhard für die Möglichkeit bedanken, innerhalb dieses Projekts promovieren zu können. Bei Prof. Dr. Wolfgang Schneider bedanke ich mich für die Bereitschaft, die Zweitbegutachtung meiner Dissertation zu übernehmen.

Mein Dank gilt auch den Projektpartnern an der Universität Regensburg, PD Dr. Sebastian Suggate und Enni Vaahtoranta, die mich – ebenso wie Prof. Dr. Wolfgang Lenhard – durch wertvolles Feedback bei der Abfassung sowie bei diversen Revisionen der vorliegenden Manuskripte unterstützt haben. In diesem Zusammenhang möchte ich mich auch bei Michaela Pirkner für das sorgfältige Korrekturlesen der Manuskripte bedanken. Danken möchte ich zudem allen Hilfskräften, Praktikanten und Praktikantinnen, sowie allen Bachelor- und Masterstudierenden, die mich bei der Datenerhebung unterstützt haben. Weiterhin bedanke ich mich bei den Erzieherinnen und Erziehern der Kindergärten, den Kindern und deren Eltern.

Schließlich gilt mein besonderer Dank meinem ehemaligen Kollegen Dr. Klaus Lingel, der mir stets mit Rat und Tat während unserer langen Bürogemeinschaft zur Seite stand, und meiner Verlobten, die mich auch während schwieriger Phasen der Promotion immer unterstützt hat.

Zusammenfassung

Die Verwendung von Geschichten zur Sprachförderung ist weitverbreitet. Einerseits zielen sie auf eine allgemeine Förderung der Wortschatzentwicklung von Kindern, andererseits sollen mit ihrer Hilfe auch Rückstände in der Wortschatzentwicklung von Risikokindern aufgeholt werden. Während der förderliche Effekt von geschichtenbasierten Interventionen bereits gut dokumentiert ist (Marulis & Neuman, 2010, 2013), besteht eine Forschungslücke zum Einfluss einzelner Faktoren, wie zum Beispiel der allgemeinen Darbietungsart der Geschichten oder dem Einsatz von Fragen (R. L. Walsh & Hodge, 2018).

Das Ziel der vorliegenden Dissertation war es, verschiedene Hypothesen in Bezug auf die Effekte unterschiedlicher Darbietungsarten und Fragestile auf den kindlichen Worterwerb durch Geschichten zu untersuchen. Die Darbietungsart bezieht sich in der vorliegenden Arbeit auf die Frage, ob Geschichten vorgelesen oder frei erzählt vorgetragen werden. Es wird davon ausgegangen, dass sich Vorlesen und freies Erzählen hinsichtlich des Erzählerverhaltens und der sprachlichen Komplexität unterscheiden. Bei den Fragestilen handelt es sich um Unterschiede im kognitiven Anspruchsniveau (niedrig vs. hoch vs. „scaffolding“-artig aufsteigend von niedrig zu hoch) und bei der Platzierung der Fragen (innerhalb der Geschichte vs. nach der Geschichte).

In den ersten beiden Studien der vorliegenden Dissertation (Studien 1 und 2) wurden Vorlesen und freies Erzählen hinsichtlich ihrer Effekte untersucht und verglichen. Studie 1 bestand aus zwei Experimenten und war als Messwiederholungsdesign konzipiert. Drei- bis sechsjährige Kindergartenkinder

($N_{\text{Experiment1}}$ = 83; $N_{\text{Experiment2}}$ = 48) bekamen Geschichten je einmal vorgelesen oder frei erzählt präsentiert. In Studie 2 wurde das Design der ersten Studie durch ein Between-Subjects-Format ersetzt und durch den Einbezug eines Geschichtenverständnismaßes sowie durch zwei weitere Experimentalbedingungen, die aus Audioaufnahmen beider Geschichtendarbietungsarten bestanden, erweitert. Letzteres erlaubte es, Unterschiede in der sprachlichen Komplexität zwischen den Darbietungsarten vom Erzählerverhalten experimentell zu trennen. Den vier- bis sechsjährigen Kindergartenkindern (N = 60) wurden die Geschichten jeweils zweimal gemäß der jeweiligen Experimentalbedingung präsentiert. Studie 1 ergab, dass keine Unterschiede zwischen freiem Erzählen und Vorlesen hinsichtlich der kindlichen Aufmerksamkeit und des Wortlernens bestanden, wenn sich die Erzähler beider Bedingungen nicht hinsichtlich des Erzählerverhaltens im Sinne von Augenkontakt und Gestikulation unterschieden. Studie 2 zeigte hingegen, dass eine naturalistischere Operationalisierung des freien Erzählens mit mehr Augenkontakt und Gestikulation zu höherer Aufmerksamkeit, höherem Wortlernen und besserem Geschichtenverständnis führte. Die Ergebnisse aus beiden Studien legen zudem nahe, dass Unterschiede in der sprachlichen Komplexität keinen Einfluss auf die kurzfristigen Lerneffekte hatten. Eine Aussage über die Bedeutung des Augenkontaktes verglichen mit der Gestikulation im Allgemeinen oder mit verschiedenen Arten von Gesten zu treffen, erlauben die beiden Studien aufgrund ihres Designs jedoch nicht.

Die letzten beiden Studien der vorliegenden Dissertation (Studien 3 und 4) untersuchten den Einfluss des kognitiven Anspruchsniveaus (niedrig vs. hoch vs. „scaffolding“-artig aufsteigend von niedrig zu hoch) und der Platzierung von Fragen (innerhalb der Geschichte vs. nach der Geschichte) sowie mögliche Interaktionen mit den kognitiven Fähigkeiten der Kinder. Den vier- bis sechsjährigen Kindergartenkindern wurden Geschichten im Einzel- (Studie 3; $N = 86$) oder Kleingruppensetting (Studie 4; $N = 91$) jeweils dreimal auf die gleiche Weise dargeboten. Je nach Versuchsbedingung wurden den Kindern die Geschichten mit den unterschiedlichen Fragetypen oder ohne Fragen vorgelesen. In beiden Studien ermunterte der Erzähler die Kinder, über die Fragen nachzudenken und sie zu beantworten. In Studie 2 sollten die Kinder zudem die Beiträge der anderen Kinder aufgreifen und diskutieren. Zur Sicherstellung der internen Validität der Experimentalbedingungen durfte der Erzähler hingegen nur unterstützend tätig werden und sich nicht inhaltlich an der Diskussion beteiligen. Beide Studien ergaben, dass der Einbezug von Fragen hinsichtlich verschiedener Aspekte des Wortlernens einen positiven Einfluss hatte. Im Widerspruch zu unterschiedlichen Hypothesen zeigten sich jedoch weder Lernunterschiede in Abhängigkeit vom kognitiven Anspruchsniveau oder der Platzierung von Fragen, noch ergab sich eine Interaktion mit dem Wortschatz oder der Gedächtnisleistung der Kinder. Die Ergebnisse beider Studien legen daher nahe, dass Unterschiede im kognitiven Anspruchsniveau und bei der Platzierung von Fragen keinen Einfluss auf das kindliche Wortlernen haben – zumindest, wenn den Fragen und den Antworten der Kinder keine weiterführenden Erklärungen und Verbesserungen durch den Erzähler folgen. Es ist jedoch noch nicht geklärt, ob eine durch verschiedene Fragetypen ausgelöste differenzielle Interaktion zwischen Kind und Erzähler zu unterschiedlichen Lernzuwächsen führen könnte.

Zusammengefasst betonen die vier Studien der vorliegenden Dissertation die zentrale Rolle des Erzählers für den Erfolg von geschichtenbasierten Interventionen.

Der Erzähler muss nicht nur die Kinder in die Geschichte involvieren, sondern sollte auch die kindlichen Äußerungen korrigieren und weiterentwickeln, um Geschichten als besonders förderliche Lernumwelt zu gestalten. Die vorliegenden Studien erweitern somit das Wissen hinsichtlich der Gestaltung von Geschichten durch Fragen und durch unterschiedliche Darbietungsarten. Es bleibt jedoch zukünftiger Forschung vorbehalten, insbesondere die Bedeutung verschiedener Fragetypen für differenzielle Erzähler-Kind-Diskussionen sowie die Rolle des nonverbalen Erzählerverhaltens weiter zu untersuchen.

Abstract

Sharing stories has become increasingly popular as a means to foster young children's vocabulary development and to target early vocabulary gaps between disadvantaged children and their better-equipped peers. Although, in general, the beneficial effects of story interventions have been demonstrated (Marulis & Neuman, 2010, 2013), many factors possibly moderating those effects – including method of story delivery as well as questioning style – merit further examination (R. L. Walsh & Hodge, 2018).

The aim of the present doctoral thesis was to test predictions from different theories on methods of story delivery and questioning styles regarding their influence on children's vocabulary learning from listening to stories. Method of story delivery refers to the general way of how stories can be conveyed, with reading aloud and free-telling of stories (i.e., the narrator telling stories without reading from text) representing different approaches that are assumed to differ regarding narrator behavior and linguistic complexity. Questioning styles refer to different combinations of questions' cognitive demand level (low vs. high vs. scaffolding-like increasing from low to high) and/or placement (within the story vs. after the story) during story sessions.

In the present doctoral thesis, the first two studies (Studies 1 and 2) compared reading aloud and free-telling of stories as different methods of story delivery. Study 1 consisted of two experiments utilizing a within-subjects design with 3- to 6-year-old preschool children ($N_{\text{experiment1}}$ = 83; $N_{\text{experiment2}}$ = 48) listening to stories once either presented read aloud or freely told. Study 2 extended the first study by examining effects on story comprehension and additionally including audiotape versions of both story-delivery methods as experimental conditions, which allowed separating narrator

behavior and linguistic complexity. With the second study being conducted as a between-subjects design, 4- to 6-year-old preschool children (N = 60) heard each of the stories twice, but listened only to one type of story delivery. The results of Study 1 indicated that no differences between methods of story delivery regarding word learning and child engagement were observable when narrator behavior in terms of eye contact and gesticulation was similar. However in Study 2, when free-telling was operationalized in a more naturalistic way, marked by higher rates of eye contact and gesticulation, it resulted in better child engagement, greater vocabulary learning, and better story comprehension than reading aloud. In contrast, as indicated by both studies, differences in linguistic complexity had no short-term impact on learning and comprehension. The studies, however, could not isolate the influence of eye contact versus gesture usage and could not distinguish between different types of gestures.

The second set of studies (Studies 3 and 4) contrasted the effects of different types of question demand level (low vs. high vs. scaffolding-like increasing from low to high) and placement (within the story vs. after the story) and examined potential interactions with children's cognitive skills. In one-to-one reading sessions (Study 3; N = 86) or small-group reading sessions (Study 4; N = 91) 4- to 6-year-old preschool children heard stories three times marked by different types of question demand level and placement or simply read-aloud without questions. The adult narrators encouraged the children to reflect on and answer questions (Study 1) and to give feedback on other children's comments (Study 2), but in both studies, to ensure fidelity of the experimental conditions, the adult narrators did not provide corrective feedback or elaborate on the children's answers. Results on measures of different facets of word learning indicated that asking questions resulted in better vocabulary learning than simply reading the stories aloud. However, in contrast to proposed hypotheses and across both studies, different types of question demand level and placement did not exert differential effects and they did not interact with children's general vocabulary knowledge or memory skills. Thus, both studies suggest that those two types of questions features have no impact on children's vocabulary learning, if questions are not followed up by narrator feedback and elaborations. However, whether different types of question placement and demand level produce differential learning gains through adult-child discussion following different questioning styles has still to be determined.

Taken together, the four studies of the present doctoral thesis underline the central role that adults play for successful story sessions with young children not only for engaging children in the story but also for extending and for correcting their utterances. Although the presented studies extend existing knowledge about methods of story delivery and questioning styles during story sessions, further research needs to examine the impact of questioning styles on word learning through subsequent adult-child discussion and to gain a better understanding of the role of nonverbal narrator behavior during story delivery.

Table of Contents

1 Introduction

The ability to communicate plays a pivotal role in human societies. To take part in verbal as well as written conversations, the recognition of words, an understanding of their meanings and the ability to use them in a situation-appropriate way are crucial prerequisites. This is not only true for official or academic contexts, but also for day-to-day situations – for example, when we want to rent a flat or when we take a new acquaintance out on a date. Without words and the knowledge reflected in them, we are virtually helpless to master many aspects of our daily life. In other words, words are "tools for communication" that we need in a wide array of situations (Scott, Nagy, & Flinspach, 2008, p. 202). Accordingly, we can see that young children's communicative success increases gradually as their toolbox in terms of vocabulary size and depth grows.

The development of vocabulary knowledge is a continuous lifelong process resulting in vocabulary sizes of around 20,000 word-families in well-educated native English speakers (Nation, 2006). Studies estimate that between 800 und 1,000 new root words are learned per year until the end of elementary school (Biemiller & Slonim, 2001). Unfortunately, non-negligible differences emerge in early childhood, with at-risk children, such as those from low socioeconomic backgrounds, displaying considerably smaller vocabulary knowledge than their peers (Fernald, Marchman, & Weisleder, 2013; Hart & Risley, 1995, 2003). As these differences do not cease to exist or even become smaller with formal schooling (Biemiller & Slonim, 2001; Christian, Morrison, Frazier, & Massetti, 2000), particularly in the long run they cause severe academic disadvantages for children with a restricted vocabulary (Walker, Greenwood, Hart, & Carta, 1994).

Vocabulary knowledge at school entry is one of the most important predictors of academic success. For instance, Hemphill and Tivnan (2008) found that vocabulary measured in Grade 1 was the best predictor of reading comprehension in Grade 2 and 3 whereas the influence of early print-related skills and phonemic-awareness slowly decreased across grades. In terms of long-term outcomes, Cunningham and Stanovich (1997) showed that vocabulary assessed in Grade 1 explained over 30% of the variance in reading comprehension measured in Grade 11. Although sufficient word knowledge seems to be particularly important for the development of text comprehension in primary (e.g., Hemphill & Tivnan, 2008; Muter, Hulme, Stevenson & Snowling, 2004; Sénéchal, Ouellette, & Rodney, 2006; Storch & Whitehurst, 2002) and secondary school (e.g., Cromley & Azevedo, 2007; Cunningham & Stanovich, 1997), this is also true for other academic domains such as mathematics (e.g., Purpura, Hume, Sims, & Lonigan, 2011) or science (e.g., Taboada, 2012). In addition, there is also emerging evidence that early vocabulary is related to self-regulation, with children displaying a larger vocabulary showing fewer externalizing and internalizing problem behaviors at kindergarten entry (Morgan, Farkas, Hillemeier, Hammer, & Maczuga, 2015).

Although academic problems may not be observable at school entry, studies targeting poor readers show that reading difficulties typically emerge when demands of textbooks dramatically increase beyond students' vocabulary knowledge (Chall, Jacobs, & Baldwin, 1990), which seems to be the case starting from Grade 3 (Biemiller, 2003). In fact, "by grade three, the gap in reading skills becomes too large for many children to ‘catch up’” (Biemiller, 2003, p. 332). Consequently, early interventions are needed to target vocabulary gaps before at-risk children enter school and drop behind their better-equipped peers.

Over the last thirty years, story-based intervention programs have become increasingly popular to target this gap and foster kindergarten and preschool children’s vocabulary development (e.g., Dickinson & Smith, 1994; Hargrave & Sénéchal, 2000; Whitehurst et al., 1988, 1994). These interventions were predominantly stimulated by observational studies that examined interactional patterns between adults and children during book reading, pointing out the beneficial effects of those situations for children’s language development (e.g., Ninio, 1980; Ninio & Bruner, 1978; Snow & Goldfield, 1983; Wheeler, 1983).

Shared-reading situations represent an early form of a “ritualized dialogue” as they are characterized by a circular adult-child interaction, with the adult drawing attention to specific aspects of a story, asking questions, as well as giving corrective feedback and elaborating on the children’s utterances (Ninio & Bruner, 1978, p. 1). Although meta-analyses and reviews report an overall positive effect of frequency of book reading on children’s language and literacy outcomes (Bus, van IJzendoorn, & Pellegrini, 1995; Mol & Bus, 2011; Scarborough & Dobrich, 1994), it became soon evident that not only the mere frequency but also the quality of book-reading situations seemed to matter (Justice & Pullen, 2003; Fletcher & Reese, 2005).

The quality of shared book reading or, more generally, of story interventions is multifaceted and refers to the way the adult narrator handles the different aspects of the circular ritualized adult-child dialogue. Observational studies have typically examined the quality of book reading on a broad level, reporting vast reading-style differences between parents (e.g., Haden, Reese, & Fivush, 1996; Heath, 1982; Hindman, Connor, Jewkes, & Morrison, 2008; Ninio, 1980) or teachers (e.g., Dickinson & Keebler, 1989; Dickinson & Smith, 1994; Martinez & Teale, 1993; Smith & Dickinson, 1994). Although the different studies tend to emphasize different aspects of the adult-child dialogue, main findings indicate that beneficial reading styles were typically marked by higher amounts of adult input such as questions, comments, and explanations as well as better child involvement.

In line with these findings, most early intervention studies focused on demonstrating that interventions targeting these aspects could improve either parent-child book reading at home (e.g., Whitehurst et al., 1988) or teacher-child book reading in day-care, kindergarten, or preschool (e.g., Elley, 1989; Valdez-Menchaca & Whitehurst, 1992; Whitehurst et al., 1994). Although these shared-reading interventions effectively help to foster children's vocabulary development, the role of factors moderating those effects still need further clarification (for meta-analyses see, Flack, Field, & Horst, 2018; Marulis & Neuman, 2010, 2013; Mol, Bus, & de Jong, 2009; Mol, Bus, de Jong, & Smeets, 2008; Stahl & Fairbanks, 1986). This is especially true for the micro-level of those factors, namely not if various types of adult input, such as questions, comments, or explanations are generally beneficial, but how they should be designed in order to best foster children's vocabulary development (for recent reviews see, R. L. Walsh & Hodge, 2018; Wasik, Hindman, & Snell, 2016).

Accordingly, the central goal of this thesis was to extend our knowledge regarding those factors that are still underspecified and that may affect the impact of story interventions on young children's vocabulary development. As there is a wide array of different factors involved in adult-child dialogues, the current doctoral thesis focused on two particularly promising factors.

The first factor that was examined in two studies (Study 1 and 2) goes beyond the moderators typically discussed in shared book reading research and concerns a more general way of how stories can be conveyed. As most interventions make use of books, one method of story delivery that is relatively little explored is the narrator telling stories without reading from text – termed free-telling of stories. This method of story delivery offers additional opportunities for language learning because it can be applied to fictional and real-life experiences without requiring the availability of books or written stories (Curenton, 2006). Free-telling may be especially promising for adults from backgrounds in which it is not normal or natural to read storybooks or to interact with a child during shared book reading (Reese, 2012).

The second factor that was examined in two studies (Study 3 and 4) concerned the effects of questioning during shared book reading, which is an integral part of the dialogic cycle (Ninio & Bruner, 1978) and regarded as a key aspect for successful interactive book reading due to opportunity to actively involve the child (R.L. Walsh & Hodge, 2018). Although we know that the use of questions generally increases learning gains compared to just-reading conditions (Marulis & Neuman, 2010, 2013; Mol et al., 2008; Wasik et al., 2016), a recent literature review points out that despite the popularity of story-intervention programs and an increasing amount of research "there appear to be too few experimental studies to allow any generalisable comments about the optimal level of demand, placement of questions, frequency of questioning or group size” (R. L. Walsh & Hodge, 2018, p. 289-290).

In the present doctoral thesis, Chapter 2 gives an overview of the theoretical and empirical background concerning these two factors, namely the effects of different methods of story delivery (free-telling vs. reading aloud; Section 2.1) and of different types of question demand level and placement (Section 2.2) on children’s vocabulary development. In Chapter 3, the research questions of the four studies and their relation to previous research are presented. Chapters 4, 5, 6, and 7 comprise the empirical studies that were conducted to answer the research questions. Finally, in Chapter 8 the findings, limitations, and implications of the four studies are summarized and directions for future studies are discussed.

2 Theoretical and Empirical Background

This chapter gives an overview of the theoretical and empirical background of the four studies included in this doctoral thesis. The first section deals with two different methods of story delivery, namely reading aloud and free-telling of stories. After describing similarities and differences between both methods of story delivery, empirical evidence on the relation to children's vocabulary learning is reviewed. The second section approaches demand level and placement of questions during story sessions and their effects on children's vocabulary development.

2.1 The Effects of Method of Story Delivery

Stories can be conveyed in a number of different ways including book reading, telling stories without making use of books – termed free-telling of stories or free story telling –, audiobooks, TV serials, and movies. Although audiobooks, TV serials, and movies have become widely accessible and thus increasingly popular to foster children's language development, adults reading or telling stories to young children seem to offer better opportunities for high-quality parent-child interactions to occur (for reviews see Kostyrka-Allchorne, Cooper, & Simpson, 2017; Linebarger & Vaala, 2010).

In contrast to reading aloud, which has been used in various intervention programs over the last decades (e.g., Biemiller & Boote, 2006; Whitehurst et al., 1988, 1994), free-telling of stories has received relatively little attention in educational contexts so far (Isbell, Sobol, Lindauer, & Lowrance, 2004). Free-telling of stories or oral story telling comprises the verbal expression of both real-life and fictional experiences to other persons (Curenton, 2006). Therefore, it is also related to the concept of parental reminiscing of real-life past events (e.g., Fivush, Haden, & Reese, 2006; Reese, Leyva, Sparks, & Grolnick, 2010; Sparks & Reese, 2013), but it does not necessarily comprise a connection to narrators' or the child's own – often mutually shared – experiences.

In terms of an alternative method of story delivery, free-telling of (primarily fictional) stories offers additional opportunities for language learning (Collins, 1999; McCabe, 1997) and may be especially promising for adults from backgrounds in which it is not normal or natural to read books aloud or to interact with a child during shared book reading (Heath, 1982; Reese, 2012). However, before free-telling of stories should be included in broadly applied language interventions, research needs to determine: (a) the relative efficacy of free-telling approaches compared to more traditional read-aloud approaches concerning children's vocabulary development, and (b) mechanisms that may explain potential differences between both methods of story delivery.

2.1.1 Reading Aloud and Free-Telling of Stories as Differing Methods of Story Delivery

Reading aloud as well as free-telling of stories offer rich opportunities for extended adult-child interaction and thus for language learning to occur (Isbell et al., 2004). Although both convey information embedded within stories, they are also marked by substantial differences (Curenton, Craig, & Flanigan, 2008). These differences probably arise because, without reading from a text, the narrator is freed from focusing on the exact wording of the text and from physical constraints imposed by handling a book, allowing him/her better to focus on the story-telling process (Suggate, Lenhard, Neudecker, & Schneider, 2013).

First, due to the lack of written material, free-telling of stories may allow the narrator to use more eye contact (Myers, 1990). Higher amounts and longer durations of eye contact may not only create a more personalized experience during story telling (Raines & Isbell, 1994; Zeece, 1997) but also increase learning, comprehension, and performance in a wide array of tasks (e.g., K. Bloom, 1974; Holler et al., 2014; Monk & Gale, 2002). In particular, for young children eye contact may help to focus and maintain attention on the story.

Second, due to the lack of a book, free-telling of stories makes it easier for the narrator to employ gestures. Gestures such as representational gestures, which depict concrete actions (iconic gestures) or refer to abstract concepts (metaphoric gestures), or deictic gestures, in which a speaker points to an object or location in the physical environment, provide semantic information and may directly help to better understand certain aspects of the story (Goldin-Meadow, 1999). For these types of gestures, Hostetter (2011) reports in her recent meta-analysis a significant, moderate benefit for listeners' comprehension, which was even greater for children. In addition, other gestures such as beat gestures, which emphasize the prosody or structure of speech without providing semantic information, offer additional visual input and may guide attention to specific words or aspects within the story (McNeill, 1992). For instance, Krahmer and Swerts (2007) showed that, on the one hand, making beat gestures influenced the speakers' acoustic realization of the co-occuring speech (Experiment 1 and 2); on the other hand, seeing a beat gesture on a word increased listeners' perception of the prominence of that word (Experiment 3). Moreover, recent studies also demonstrated that beat gestures impact speech perception at the neural level, suggesting a common neural substrate for speech and gesture processing (e.g., Biau, Fernàndez, Holle, Avila, & Soto-Faraco, 2016; Hubbard, Wilson, Callan, & Dapretto, 2009). Therefore, irrespective of their intended communicative value and comparable to the role of eye contact, gestures in general may also help children to maintain their attention on the story and to focus on important story aspects.

Third, oral and written language differ also substantially regarding a wide array of linguistic surface characteristics. Written language is typically marked by more demanding syntactic, semantic, and lexical structures (e.g., Akinnaso, 1982; Hayes, 1988; Hayes & Ahrens, 1988; Montag, Jones, & Smith, 2015; Nation, 2006; O'Donnell, 1974; Woolbert, 1922). Nation (2006), for example, analyzed the number of words that are required for comprehension of written and spoken English and concluded that a vocabulary of 6,000 to 7,000 word-families is sufficient for spoken text, but that an 8,000 to 9,000 word-family vocabulary is needed for comprehension of written material. Moreover, Montag et al. (2015) recently found that the vocabulary contained in children's books was considerably more diverse than child-directed speech and adult-child conversations. This is important for child education because greater diversity and complexity of linguistic input have been shown to be associated with better language outcomes (e.g., Hart & Risley, 1995; Hoff, 2003; Huttenlocher, Waterfall, Vasilyeva, Vevea, & Hedges, 2010).

Finally, it has also been proposed that free-telling of stories offers better linguistic and behavioral flexibility to adapt to the audience, to react to signs of disinterest, and to re-gain listeners' attention if lost (Myers, 1990; Roney, 1996). However, particularly concerning flexible adaption of language complexity, the situation is more complicated than proposed. Although theoretically free-telling may provide better opportunities to adapt language to the audience, for example by rephrasing, and may thus enhance interest and comprehension (Roney, 1996), this proposed advantage depends heavily on the narrators' language skills (e.g., vocabulary knowledge, grammar, articulateness) as well as their overall storytelling competence. Moreover, the narrator needs to be familiar with the listeners and their cognitive skills in order to provide a tailored linguistic input.

In sum, reading aloud and free-telling of stories differ regarding several aspects of story delivery, with narrator behavior and flexibility favoring free-telling, but linguistic complexity supporting the use of books and other written material. These differences in turn may have an influence on children's learning from stories.

2.1.2 The Effects of Reading Aloud versus Free-Telling of Stories on Vocabulary Acquisition

Although free-telling of stories is strongly endorsed (e.g., Curenton & Craig, 2011) and a wide array of guidelines to conduct story-telling sessions have been published over the last decades (e.g., Barton, 1986; Grugeon & Gardner, 2012; Howe &

Johnson, 1992; Wright, 1995), so far only few studies have examined the effects of free-telling of fictional stories and compared them to book reading.

Observing a storyteller perform both types of story delivery, Myers (1990) reported that school children aged 7 to 11 tended to laugh more, ask more questions, and give more responses during free-telling of stories than during reading aloud. In addition, the children’s responses during story delivery indicated that the freely told stories tended to engage children better and were marked by less behavior indicating inattention (e.g., looking away from the narrator or fidgeting). However, the single narrator, who performed the readings and free-tellings, received no guidelines or restrictions, favored himself the free-telling condition, and asked considerably more questions during this condition. Therefore, it is unclear which aspects of free-telling of story delivery might be superior to reading aloud. For example, in Myers’ (1990) study it might be that the higher rate of questions, which are known to considerably increase learning gains in shared book reading (e.g., Ewers & Brownson, 1999; Sénéchal, 1997; B.A. Walsh & Blewitt, 2006), might be the cause of the observed differences between both types of story delivery.

Concerning cognitive learning gains, studies with elementary school children (Suggate et al., 2013; Trostle & Hicks, 1998) and second language learners (Uchiyama, 2011) have shown that free-telling of stories resulted in better vocabulary learning and story comprehension than reading aloud. However, as Trostle and Hicks (1998) as well as Uchiyama (2011) used a method of free-telling that involved dressing as the protagonist in the story and pantomimic acting, it is unclear whether differences were due to features of this specific type of storytelling and if they can be generalized to other less enriched forms of free-telling.

This problem was targeted by Suggate et al. (2013) who used one-to-one reading sessions in a sample of 37 elementary school children, with grade as a between-subjects factor (two vs. four) and story delivery as a within-subjects factor (independent reading vs. adult read-aloud vs. storytelling). For both second and fourth grade, Suggate et al. reported that children learned more target words in the storytelling condition than in the read-aloud condition. However, as Suggate et al.'s study did not include a pretest, it was impossible to identify actual learning gains and apart from the notion that the free-telling took more time and showed more variability in duration, no information on narrator and child behavior during story delivery were collected.

Finally, Isbell et al. (2004) who used a story-retelling task and a wordless-picture-book task, in which a story had to be told according to the pictures, reported mixed findings in their study with 38 3- to 5-year-old preschool children. Although descriptive analyses of children's utterances indicated higher gains regarding word fluency and diversity for reading aloud, there were little differences in mean length of utterances and mixed differences regarding aspects of story conventions and comprehension. However, as no significance tests were conducted and no effect sizes or information needed to calculate them were reported, it is difficult to evaluate Isbell et al.'s findings properly.

In sum, empirical evidence regarding the effects of reading aloud versus free-telling of stories on children's vocabulary development is still sparse. Although most observational (Myers, 1990) and experimental studies (Suggate et al., 2013; Trostle &

Hicks, 1998; Uchiyama, 2011) favor free-telling of stories, they typically targeted older populations and/or second language learners and most were marked by methodological problems. Additionally, the studies provide only limited insight into mechanisms as, for example, they did not connect narrator behavior or child engagement to children's vocabulary learning or story comprehension. Therefore, further research is clearly needed (a) to shed light on potential effects of reading aloud and free-telling in preschool age and (b) to link potential narrator- and child-based mechanisms to learning gains.

2.2 The Effects of Questions during Story Interventions

Besides other explicit elements such as comments and explanations, questions are widely used to enhance the effects of story interventions on children's vocabulary development. Questions, by their nature, offer especially rich opportunities for active child involvement, which is regarded as a key aspect of successful story interventions (Zucker, Justice, Piasta, & Kaderavek, 2010). Although numerous studies have demonstrated the beneficial influence of questions on learning gains compared to exclusively reading a book (e.g., Blewitt, Rump, Shealy, & Cook, 2009; Ewers & Brownson, 1999; Sénéchal, 1997; Sénéchal, Thomas, & Monker, 1995; B.A. Walsh & Blewitt, 2006), the importance of question characteristics has still to be determined (R.L. Walsh & Hodge, 2018). In particular, research needs to resolve: (a) how cognitively challenging or demanding questions should be (i.e., cognitive demand level of questions), and (b) when questions should be asked (i.e., placement of questions).

2.2.1 The Effects of Demand Level of Questions

Defining demand level of questions. The demand level of a question is defined by the cognitive demand that is imposed upon the child in order to provide an answer. Cognitive demand increases with the level of abstraction or the representational demand of the question (van Kleeck, Gillam, Hamilton, & McGrath, 1997). Although cognitive demand represents a continuum (Anderson et al., 2001; Sigel & McGillicuddy-Delisi, 1984; van Kleeck et al., 1997), empirical studies typically use only a dichotomous categorization. Depending on the specific study, the poles are termed low- versus high-demand questions (e.g., Blewitt et al., 2009; Reese & Cox, 1999; B. A. Walsh, Sánchez, & Burnham, 2016), literal versus inferential questions (e.g., van Kleeck, Vander Woude, & Hammett, 2006), perceptual versus conceptual questions (e.g., Justice, 2002), contextualized versus decontextualized talk (e.g., Hindman et al., 2008; Hindman, Wasik, & Erhart, 2012), or immediate versus non-immediate talk (e.g., De Temple, 2001; De Temple & Snow, 2003).

Although the terminology used across studies differs, studies typically use very similar conceptions to define cognitive demand and to separate low- from high-demand questions (R. L. Walsh & Hodge, 2018). Low-demand questions are marked by low levels of abstraction and they focus on descriptions or mere repetitions of concrete information presented in the story. High-demand questions, in contrast, are more abstract as they require the child to go beyond the information presented in the story and to connect story aspects to prior knowledge (for an overview of features of low- and high-demand questions used in different studies, see Table 1).

Table 1

Features of Low- and High-demand Questions Used in Different Studies

Study and terminology	Low demand	High demand
Blewitt et al. (2009, p. 295) *low vs. high demand questions*	• describing story and picture content	• making inferences about why an event happened or about characters' feelings • explaining the meaning of a word • making predictions • relating story elements to personal experience
De Temple & Snow (2003, p. 19, 21) *immediate vs. non-immediate talk*	• labeling objects	• making predictions • making connections to the child's past experiences, to other books, or the real world • drawing inferences • analyzing information • discussing the meaning of words • offering explanations
Hindman et al. (2012, p. 453) *contextualized vs. decontextualized talk*	• labeling and describing illustrations	• drawing inferences and conclusions • making predictions • summarizing or recalling what has happened
Justice (2002, p. 90–91) *perceptual vs. conceptual questions*	• describing concrete perceptual features of the text or illustrations	• making judgments • making predictions • explaining concepts depicted in the text or illustrations
Reese & Cox (1999, p. 21) *low vs. high demand questions*	• describing and labeling pictures • what-questions	• making inferences about story events • making predictions about story events • reason explanations • affective commentary
van Kleeck et al. (2006, p. 86) *literal vs. inferential questions*	• labeling objects and characters • describing objects and actions	• making inferences about attitudes, feelings, motives, etc. • identifying similarities and differences

		• making predictions • inferring the meaning of words • making connections between information
B. A. Walsh et al. (2016, p. 264) *low vs. high demand questions*	• naming details • identifying pictures	• making inferences • making predictions • hypothesizing • summarizing or explaining story events

These conceptualizations of cognitive demand also fit in the broader framework represented by Bloom's Taxonomy of Educational Objectives (B. S. Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956) or its revision (Anderson et al., 2001). In the revised taxonomy, the complexity of cognitive processes is represented by the following order: (1) remember, (2) understand, (3) apply, (4) analyze, (5) evaluate, and (6) create. Low-demand questions particularly tap into the first two categories whereas high-demand questions primarily focus on the categories 4, 5, and 6, which represent complex cognitive processes (see also Hindman et al., 2012).

Similarly, there is also a clear connection to Sigel's distancing theory, which assumes that cognitively challenging interaction is marked by a greater degree of cognitive distancing form the immediate environment (Sigel & McGillicuddy-DeLisi, 1984). In line with Sigel's distancing theory, low-demand questions can be answered by information provided in the text and thus need only little cognitive distancing. High-demand questions, in contrast, require the child to go beyond the immediate context of the story, representing a higher degree of cognitive distancing.

Question demand level and vocabulary acquisition. Taking up the low- vs. high-demand distinction, three major hypotheses on the relationship between demand level of questions and vocabulary learning from listening to stories have been proposed in shared book reading research.

The aptitude-treatment-interaction hypothesis. A first hypothesis appertains to the well-known aptitude-treatment-interaction framework (e.g., Connor et al., 2009) and assumes that the fit between children's skills and the cognitive demand of the parental input matters. Accordingly, Reese and Cox's (1999) aptitude-treatment-interaction hypothesis assumes that children of higher performance levels would profit more from high-demand questions, and less-skilled children would profit more from low-demand questions. Support for this hypothesis stems from several observational studies on parent-child interaction as well as from an experimental study conducted by Reese and Cox (1999).

Observational studies report that parents tend to adapt the cognitive demand of their talk during shared book reading to the age and cognitive development of their children (Pellegrini, Brody, & Sigel, 1985; Pellegrini, Perlmutter, Galda, & Brody, 1990; Sigel & McGillicuddy-DeLisi, 1984) and that this kind of adaption is positively related to children's language development (Dale, Crain-Thoreson, Notari-Syverson, & Cole, 1996; Pellegrini et al., 1985). Although these findings agree with the aptitude-treatment-interaction hypothesis, the studies did not exclusively focus on questions and findings are only correlational. Consequently, they provide only indirect evidence for the aptitude-treatment-interaction hypothesis.

In contrast, Reese and Cox's (1999) study, in which 48 4-year-old children heard approximately 30 stories presented in one of three reading styles across the span of six weeks, provides some causal evidence for the aptitude-treatment-interaction hypothesis. The reading styles were marked by either low-demand questions interspersed within the stories (describer style), or high-demand questions interspersed within the stories (comprehender style), or high-demand questions after the story presentation (performance-oriented style). Although the describer style resulted in the best vocabulary gains on a standardized test of general receptive vocabulary knowledge, this main effect of reading style was moderated through an interaction with children's vocabulary knowledge. Supporting the aptitude-treatment-interaction hypothesis, children with low vocabulary profited most from low-demand questions that were interspersed within the story, whereas children with better vocabulary gained most through high-demand questions placed after the story. Unfortunately, it is difficult to evaluate the robustness of the findings and the aptitude-treatment-interaction hypothesis because most subsequent studies did not test for aptitude-treatment interactions (Justice, 2002; B. A. Walsh et al., 2016) or did not find meaningful interactions between question demand level and children's language skills (Blewitt et al., 2009).

The scaffolding-like hypothesis. The second hypothesis on the effects of question demand level, proposed by Blewitt et al. (2009), draws on the scaffolding framework (Wood, Bruner, & Ross, 1976) and argues a scaffolding-like approach, consisting of a gradual transition from low- to high-demand questions, might be best suited to foster some aspects of children's novel word learning. In contrast to the aptitude-treatment-interaction hypothesis, Blewitt et al.'s scaffolding-like hypothesis does not assume a general superiority of a scaffolding-like procedure. Instead, Blewitt et al. separate between processes reflecting the initial acquisition of a novel word and subsequent processes leading to more elaborated word knowledge. They argue that cognitive demand level of questions does not affect the initial acquisition of a word-referent association. However, after this initial step of acquisition a scaffolding-like procedure of increasing demand level should especially foster deeper understanding of a word’s meaning.

With their line of argument, Blewitt et al. explicitly draw on the distinction between fast- and slow-mapping processes in novel word acquisition that was originally proposed by Carey (1978). Fast mapping refers to the first step in word acquisition, by which some of its phonological, syntactic, or semantic features enter the mental lexicon, building an initial word representation. A few or even only a single encounter with a word can suffice to acquire this kind of initial mental representations (Carey, 1978; Carey & Bartlett, 1978). Slow mapping, in contrast, represents a longer incremental process, leading to deeper, more elaborated, and better-structured word knowledge (Carey, 2010; Swingley, 2010).

Despite not separating between word learning processes, studies provide some observational evidence indirectly supporting the general idea of a scaffolding-like procedure for question demand level. First, adult’s input during shared book reading with their children is marked by a mixture of lower-level and higher-level talk (e.g., DeLoache & DeMendoza, 1987; Hammett, van Kleeck, & Huberty, 2003; van Kleeck et al., 1997). Second, during shared book reading, parents and children tend to match the cognitive demand of each other’s language input, with the adults typically raising the level of abstraction (e.g., Danis, Bernard, & Leproux, 2000; Tompkins, Zucker, Justice, & Binici, 2013; Zucker et al., 2010).

Direct empirical evidence for the scaffolding-like hypothesis is provided by Blewitt et al. (2009) who found no differences between low-demand, high-demand and scaffolding-like questions on a measure of receptive target-word learning that they used as an indicator of initial word acquisition. Furthermore, they reported that in a sample of

3-year-olds a scaffolding-like condition resulted in significantly higher learning gains than either the low- or the high-demand condition on a word definition task that they used as an indicator of deeper word knowledge. However, as Blewitt et al.'s study represents the only experimental work having examined a scaffolding-like procedure for question demand level during shared book reading, further research is needed to replicate and evaluate the scaffolding-like hypothesis.

The high-demand hypothesis. A third hypothesis, the high-demand hypothesis, also includes the distinction between initial and deeper word acquisition. In contrast to the scaffolding-like hypothesis, it proposes that high-demand questions might be especially beneficial to create and enrich deeper word knowledge (B. A. Walsh et al., 2016).

Thus, conceptually, the high-demand hypothesis is related to the classical notion of depth or levels of processing in memory research (Craik & Lockhart, 1972). According to the depth-of-processing framework, greater depth is achieved by a greater degree of cognitive analysis or elaboration and results in a more persistent memory trace (Craik & Lockhart, 1972). Studies have demonstrated this for qualitatively different types of encoding such as phonemic versus semantic encoding, with semantic encoding requiring deeper processing and leading to better recognition and recall than phonemic processing (Craik & Tulving, 1975). In addition, within specific domains such as the domain of semantic encoding, higher amounts of elaboration were shown to be associated with better memory performance (Craik & Tulving, 1975; for a review see Lockhart & Craik, 1990). However, as these studies targeted mainly rote learning of single items and were mostly conducted with older students, it is unclear if these findings from memory research can be transferred to shared book reading with preschool and kindergarten children.

The high-demand hypothesis does also fit in recent cognitive-learning theories and frameworks (e.g., Chi, 2009; Wittrock, 2010), which argue that questions become more effective as the level of cognitive processing that is needed to solve them increases. Although studies with university students generally support this idea (e.g., Cerdán, Vidal-Abarca, Martínez, Gilabert, & Gil, 2009; Jensen, McDaniel, Woodard, & Kummer, 2014; Roelle & Berthold, 2017), findings are not always consistent, being moderated, for example, through the type of criterion measure (e.g., Cerdán et al., 2009; Roelle, Roelle, & Berthold, 2019) or the instructional context (e.g., Roelle & Berthold, 2017). Most importantly, however, the results of these studies are difficult to transfer to the effects of question demand level during shared book reading with children, as none of the studies was actually conducted with young children nor did any of them focus on language learning.

Turning to shared book reading with children, results of multiple observational studies suggest that children benefit when their parents (e.g., Blake, Macdonald, Bayrami, Agosta, & Milian, 2006; Haden et al., 1996; Leseman & de Jong, 1998) or teachers (e.g., Dickinson & Porche, 2011; Dickinson & Smith, 1994; Hindman et al., 2008) use cognitively challenging talk. The studies, however, do typically not separate between initial and deeper word-knowledge acquisition. In addition, experimental evidence in favor of the high-demand hypothesis is sparse. So far, the only experimental study on word learning in shared book reading lending partial support for the high-demand hypothesis, was conducted by B. A. Walsh et al. (2016), reporting higher gains for expressive target words for the high-demand questions in a sample of 3- to 5-years-old dual language learners. However, since none of the experimental conditions was actually better than an exclusively reading condition, B. A. Walsh et al.'s results are not very convincing and should be interpreted with caution. Other experiments on question demand level did not find differences between low- and high-demand questions (Blewitt et al., 2009; Justice, 2002) or even showed that although high-demand question could be beneficial for highly skilled children, lower skilled children did learn more through low-demand questions (Reese & Cox, 1999).

Taken together, there seems to be theoretical and empirical support for the high-demand hypothesis from studies targeting memory functioning was well as learning from texts in older students. However, in shared-reading, the situation is much more complicated. Although theoretical considerations as well as observational studies tend to support the idea that cognitively challenging input seems to be especially beneficial for language development, experimental studies lend little support for that notion.

Finally, when summarizing the studies on cognitive demand level of questions, the emerging picture is inconclusive, favoring neither of the proposed hypotheses. In addition, integrating the existing findings is difficult as the studies did not only differ regarding important design features such as the number of reading sessions and sample characteristics such as the mean age or the age span, but there is also the problem that most studies did not use comprehensive designs that allowed to test the various hypotheses simultaneously. Consequently, more work is required using comprehensive studies to test the various positions within the same research design.

2.2.2 The Effects of Placement of Questions

Defining placement of questions. In contrast to demand level of questions, question placement during shared book reading has received relatively little attention so far. Theoretically, two suggestions dominate the literature, namely recommendations to embed questions within stories termed “interrupting” questioning style, or to place questions prior to or following the stories termed “non-interrupting” questioning style (R. L. Walsh & Hodge, 2018). Although naturalistic studies point out, that most narrators simultaneously use both types of placement, they often show a more or less pronounced gradual preference for questions placed either within or around the stories (e.g., Dickinson & Keebler, 1989; Dickinson & Smith, 1994; Martinez & Teale, 1993).

Question placement and vocabulary acquisition. Three different hypotheses regarding the effects of question placement on novel word learning can be derived from literature.

Interruption hypothesis. First, it has been proposed that an interrupting style, which is marked by embedding questions within stories, may offer the possibility to reflect on and clarify difficult aspects of the story such as unknown or difficult words as soon as they arise, and thus enhance comprehension and learning of novel words (Brabham & Lynch-Brown, 2002). Although the assumption of the interruption hypothesis theoretically makes sense, empirical findings are mixed at best.

For instance, Brabham and Lynch-Brown (2002) reported in their study with elementary school children (2^{nd} and 4^{th} grade) that an interrupting question placement was superior to a non-interrupting one. However, other studies do not (Blewitt et al., 2009; Dickinson & Smith, 1994; Gonzalez et al., 2014; Jimenez & Saylor, 2017; B. A. Walsh et al., 2016) or at least not fully (Reese & Cox, 1999) support the interrupting hypothesis. For example, Reese and Cox's (1999) results generally favored low-demand questions interspersed within the stories, but indicated an interaction with children's language skills. Children with good vocabulary profited most from high-demand questions asked after the story. Moreover, Blewitt et al. (2009) and B. A. Walsh et al. (2016) experimentally contrasted both types of question placement in a one-to-one respectively a small-group design with preschoolers, but found no differences between questions within or after the stories.

No-interruption hypothesis. Second, it was proposed that a non-interruptive questioning style might be beneficial because the story flow is not disrupted, which in turn might better sustain children's attention and motivation and therefore also their comprehension of the story (Strasser, Larraín, & Lissi, 2013). Partly in line with the non-interrupting hypothesis, an analysis of teacher talk in preschool book reading sessions indicated that only the duration dedicated to after-reading interaction was significantly related to expressive (but not receptive) vocabulary development (Gonzalez et al., 2014). In addition, Dickinson and Smith (1994; also see Dickinson & Tabors, 2001) conducted an observational study in 25 kindergarten classrooms and found that classrooms marked by a non-interrupting reading style with primarily high-demand questions woven around the stories were associated with better vocabulary development than classrooms characterized by an interrupting reading style with primarily low-demand questions. Both styles were, however, not different from a third style that consisted of an interrupting style with primarily high-demand questions. Finally, as already mentioned, most studies did not find any differences between both types of question placement (Blewitt et al., 2009; B. A. Walsh et al., 2016) or indicated an interaction with question demand level (Dickinson & Smith, 1994), children's cognitive skills (Jimenez & Saylor, 2017), or even a three-way interaction comprising children's cognitive skills, question demand level and placement (Reese & Cox, 1999).

Aptitude-treatment-interaction hypothesis. Third, it was also proposed that neither an interrupting nor a non-interrupting questioning style might be the single best style. In terms of an aptitude-treatment interaction, question placement effects may depend on cognitive characteristics of the listeners such as their language (Reese & Cox, 1999) or memory skills (Jimenez & Saylor, 2017). Based on the observation that adults typically use a high number of low-demand questions during shared book reading with younger children, Reese and Cox (1999) assumed that not only low-demand questions but also an interrupting style should be especially beneficial for younger and lower-skilled children. In contrast, Jimenez and Saylor (2017) argued that asking questions within stories increases cognitive load and therefore should lead to a performance decrease in lower-ability children. Although both research groups reported evidence for their respective hypotheses (Jimenez & Saylor, 2017; Reese & Cox, 1999), the only other experimental study on questions during shared book reading found no relationship between children's general vocabulary and question placement or demand level (Blewitt et al., 2009).

2.2.3 Interactions between Question Demand Level and Placement

In addition to hypotheses on question demand level and placement, research has also examined potential interaction effects between both question features. Particularly, combinations of high-demand questions placed after the stories and low-demand questions during the stories have been proposed.

For instance, Dickinson and Smith (1994) reported that a reading style consisting of high-demand questions after the stories was positively associated with preschool children's vocabulary development. Reese and Cox's (1999) results, however, indicate that this demand-level-placement combination might only be suited for high-ability children, whereas others might profit more from an interrupting style with low-demand questions.

This finding does not only concur with observational studies showing that parents tend to adapt their input to age and cognitive development of their children (Pellegrini et al., 1985, 1990; Sigel & McGillicuddy-DeLisi, 1984), but it does also help to explain the seemingly contradictory findings reported for question placement (Jimenez & Saylor, 2017; Reese & Cox, 1999). The effects of question placement may depend on question demand level and the addressed population. Low-ability children might simply need information provided from low-demanding input in order to understand the story and to learn novel words. In contrast, if input is too challenging or – in other words – if it imposes too much cognitive load, it might not help but even impair understanding and learning (Jimenez & Saylor, 2017). Supporting this, Strasser et al. (2013) found that elaborations of target words had differential effects depending on children's language skills and that only if elaborations increased target-word learning, they also boosted story comprehension.

In sum, regarding the high number of proposed hypotheses in conjunction with the numerically and at times methodologically limited research that has been conducted so far, further studies are clearly needed to shed light on effects of question demand level and placement. Importantly, however, these studies should always report sample characteristics and examine the interaction of question features with children's cognitive skills. If not, different results may simply reflect the cognitive level of a particular sample.

3 The Present Research

The studies that are reported in this doctoral thesis build on previous research on the effects of different types of story delivery (Study 1 and 2) and of question demand level and placement (Study 3 and Study 4). By systematically examining these factors, the present research aimed not only at establishing a connection between story-intervention features and word learning, but also at shedding more light on potential underlying mechanisms.

Regarding the effects of different methods of story delivery, research has discussed similarities and differences between reading aloud and free-telling of stories (see Section 2.1). In particular, narrator behavior, language complexity, and child engagement are assumed to differ between both methods of story delivery (e.g., Isbell et al., 2004; Suggate et al., 2013). Although research has shown that more eye contact (e.g., Holler et al., 2014; Monk & Gale, 2002) and more gesture usage (e.g., Hostetter, 2011; Krahmer & Swerts, 2007) are associated with increased learning, none of the previous studies comparing reading aloud versus free-telling of stories has linked potential differences in narrator behavior to child engagement and actual learning from the stories. Moreover, as most studies (e.g., Suggate et al., 2013; Trostle & Hicks, 1998; Uchiyama, 2011) were conducted with school children, it is unknown if their findings that favored free-telling regarding story comprehension and word learning can be transferred to preschool children.

Study 1 comprised two experiments and was designed to target these open questions. We used a procedure similar to the study conducted by Suggate et al. (2013). In each experiment, a sample of 3- to 6-year-old children heard stories freely told and read aloud. We examined differences in narrator behavior, child engagement, and their relation to word learning during story presentation. The study is presented in detail in Chapter 4.

Study 2 was designed to further extend previous experimental studies (including Study 1). Extensions to previous research concerned particularly two aspects. First, by including a measure of word learning as well as a measure of story comprehension, Study 2 expanded the list of dependent variables and allowed to relate narrator behavior and child engagement to different facets of learning. Second, instead of two conditions, the study included four story-presentation conditions: (a) *in vivo* free-telling, assumed to be marked by large amounts of eye contact and narrator gesticulation and less complex language, (b) *in vivo* read-aloud, marked by more complex language, but less gesticulation and eye contact, (c) an audiotaped version of free-telling, and (d) an audiotaped version of reading aloud, both without narrator gesticulation or eye contact. By doing so, it was possible to separate effects of narrator behavior and effects of language complexity and to examine their differential relation to word learning and story comprehension. The study is presented in detail in Chapter 5.

Regarding question demand level and placement (see Section 2.2), three different hypotheses were proposed for both demand level (aptitude-treatment-interaction hypothesis, scaffolding-like hypothesis, high-demand hypothesis) and placement (interruption hypothesis, no-interruption hypothesis, aptitude-treatment-interaction hypothesis). Additionally, previous literature indicated that it might not be sufficient to examine each of the factors alone but that meaningful interactions comprising those question features and children's cognitive skills are likely to occur.

Although all of the proposed hypotheses are theoretically well grounded, results from empirical studies are conflicting. Moreover, the scaffolding-like hypothesis for question demand level has only been examined by the same study that had originally proposed it (Blewitt et al., 2009). Consequently, further studies that, on the one hand, comprise comprehensive designs of question demand level, question placement, and children's cognitive skills and that, on the other hand, include measures of initial and deeper word learning were needed to test predictions made by the different hypotheses.

Study 3 was designed to closely mirror Blewitt et al.'s (2009) study that had introduced the scaffolding-like condition for question demand level and emphasized the differentiation between initial shallow and subsequent deeper learning processes during novel word acquisition. The major aim of Study 3 was to test predictions by different hypotheses on question demand level and placement within a single comprehensive design. In concord with most studies in which the narrator read to children individually (e.g., Blewitt et al., 2009; Justice, 2002; Reese & Cox, 1999), children heard the stories in one-to-one settings. This setting controls for confounding variables and allows best to establish links between question features and learning gains. Finally, to measure longer-term effects, Study 1 included a delayed posttest in addition to an immediate posttest. The study is presented in detail in Chapter 6.

Study 4 built on and systematically extended Study 3. In the discussion of Study 3, a separation between effects of question features caused by individual cognitive processes and differential discussion following questions were developed. Following this differentiation, we proposed that for young children question demand level and placement may have little differential effect without differential discussion following the questions. To examine this suggestion, we retained the comprehensive design and the differentiation between measures of initial shallow and subsequent deeper learning processes during novel word acquisition, but switched from one-to-one book readings to a small-group design. The study is presented in detail in Chapter 7.

4 Study 1: Incidental Vocabulary Acquisition from Listening to Stories: A Comparison between Read-aloud and Free Storytelling Approaches

A version of this chapter was published in:

Lenhart, J., Lenhard, W., Vaahtoranta, E., & Suggate, S. (2018). Incidental vocabulary acquisition from listening to stories: A comparison between read-aloud and free storytelling approaches. *Educational Psychology, 38*, 596–616. doi:10.1080/01443410.2017.1363377

Abstract

Shared book reading is a well-established intervention to foster vocabulary development. Factors influencing its effectiveness are, however, less well studied, particularly with regards to story-delivery. We contrasted a read-aloud with a free storytelling approach and tested effects on vocabulary learning. In the first study, 83 preschoolers aged three to six were told six stories in a randomized, single-blind and counterbalanced design. Stories were either read aloud or told freely and included rare target-words. Measures of target-word acquisition, receptive vocabulary, phonological working memory, and speech comprehension were administered. There was a small to moderate learning gain (d = 0.37), but no effect of story-delivery. In a second study, 24 of the youngest and 24 of the oldest participants were tested again, using the same procedure but with stories designed to be more intrinsically motivating and age-appropriate. Results indicated negligible vocabulary gains (d = 0.08) and no effect of story-delivery, except for small differences in child behaviour during storytelling.

Keywords: language development, incidental learning, storytelling, read-aloud, vocabulary acquisition

Vocabulary knowledge plays a pivotal role for success in modern societies and is a clear prerequisite for everyday communication. Vocabulary is often deemed to be crucial for the development of future reading skills (Sénéchal, Ouellette, & Rodney, 2006) and is an important predictor of reading comprehension in adolescence (Cromley & Azevedo, 2007). Although there is unanimous agreement that vocabulary should be fostered as early as possible (Biemiller, 2003), factors influencing its development in intervention settings are less well studied – particularly with regard to story-delivery. Accordingly, the current study presents the findings from two randomized experiments conducted on preschoolers looking at the effect of story-delivery (Study 1) and additionally story-content (Study 2) on vocabulary acquisition.

Early Vocabulary Development

Vocabulary growth is a continuous, lifelong process resulting in a dictionary size of around 20,000 word-families in well-educated native speakers (Nation, 2006). Particularly during the first two decades of life, vocabulary size and depth increase dramatically. For instance, Biemiller and Slonim (2001) estimate that between 800 und 1000 new root words are learnt per year until the end of elementary school. This growth rate is, however, by no means homogenous and children's vocabulary sizes differ vastly during the first years of life with socially disadvantaged children exhibiting considerably smaller lexicons than their peers (e.g., Biemiller & Slonim, 2001; Hart & Risley, 1995). These differences in vocabulary do not reduce during elementary school (Biemiller & Slonim, 2001), with formal schooling seeming to have little compensatory effect (Christian, Morrison, Frazier, & Massetti, 2000).

Vocabulary growth is thought to result from a number of different influences, including maternal and parental factors (Baydar, Brooks-Gunn, & Furstenberg, 1993; Nelson, Welsh, Trup, & Greenberg, 2011; Pancsofar, & Vernon-Feagans, 2006), genetic influences (Kovas, Hayiou-Thomas, Oliver, Dale, Bishop, & Plomin, 2005; Hayiou-Thomas, Harlaar, Dale, & Plomin, 2006), and specific features of oral language environments (Downer & Pianta, 2006; Rowe, 2008). In terms of oral language environments, peer and parent-child interactions appear to play important roles (e.g., Ramírez-Esparza, García-Sierra, & Kuhl, 2014), alongside more specific factors such as sharing stories, shared-reading, and specific vocabulary teaching (Mol & Bus, 2011; Swanborn & de Glopper, 1999). Focusing on specific language interactions intended to foster vocabulary development, there are different approaches regarding how much direct teaching versus incidental learning should occur (Mol & Bus, 2011; Nation, 2006; Roney, 1996; Zucker, Solari, Landry, & Swank, 2013).

Indeed, considering the vast number of words that are learnt during childhood and adolescence (Biemiller & Slonim, 2001; Nation, 2006), it could be argued that direct teaching of vocabulary can only account for a small number of these words, such that most of them must be learned incidentally (i.e., without direct instruction). This line of thought is supported by observations showing little direct vocabulary instruction during the first school years (e.g., Durkin, 1978-1979; Jenkins & Dixon, 1983) and by experimental studies providing evidence that vocabulary knowledge can be gained incidentally through reading (Nagy, Herman, & Anderson, 1985; for a meta-analysis see Swanborn & de Glopper, 1999) and listening (e.g., Gampe, Liebal, & Tomasello, 2012; Robbins & Ehri, 1994). Although the importance of reading for vocabulary development seems to increase during the school years (Swanborn & de Glopper, 1999), overhearing of conversations (e.g., Akhtar, 2005; Gampe et al., 2012), and listening to stories (e.g., Elley, 1989; Robbins & Ehri, 1994) are considered to be among the main driving factors during childhood lexical growth. Storytelling or book reading settings seem to offer especially rich opportunities for word learning to occur (Ninio & Bruner, 1978).

Listening to Stories as a Vocabulary Intervention

In contrast to foreign language learning in school, which relies heavily on direct teaching methods, interventions targeting vocabulary knowledge in young children are usually indirect in nature. Typically, these interventions are embedded within storytelling contexts (e.g., Ard & Beverly, 2004; Blewitt, Rump, Shealy, & Cook, 2009; Whitehurst et al., 1988).

Story-based interventions which are purely implicit and rely solely on incidental vocabulary acquisition typically find small to moderate gains (e.g., Coyne, McCoach, & Kapp, 2007; Elley, 1989). These gains, however, can be enhanced by repeated readings (Beck & McKeown, 2007) or by including explicit features (Elley, 1989; for a meta-analysis of older studies, see Stahl & Fairbanks, 1986; for more recent studies, see Marulis & Neuman, 2010). Yet, there is considerable variation depending on the specific features of the interventions (Stahl & Fairbanks, 1986), with factors such as type and placement of questions (e.g., Blewitt et al., 2009; Walsh, Sánchez, & Burnham, 2016), or the influence of explanations and comments (e.g., Ard & Beverly, 2004; Justice, 2002) playing roles. Given that virtually all story-based interventions make use of books or other written material, little is, however, known about the impact of story-delivery, namely, whether the story is read aloud or told freely.

Read-Aloud and Free Storytelling Approaches

Theoretically, read-aloud and free storytelling approaches have some basic characteristics in common. They impart knowledge in the context of a story, require adequate listening comprehension to learn new vocabulary and to deepen existing semantic word knowledge, and offer the opportunity for interaction with the narrator (Isbell, Sobol, Lindauer, & Lowrance, 2004). Yet, they also differ to some extent. The most obvious difference is that free storytelling is a non-textual, oral language experience, while reading aloud is text-based. As a logical consequence, reading aloud reproduces exactly the same text each time, whereas in a storytelling event the verbal shape of the story is never the same and created each time anew (Roney, 1996). Undoubtedly, being able to tailor the story according to the interests of children as well as to their cognitive and verbal competencies creates the opportunity to ensure comprehension and to enhance participation (Roney, 1996). Another advantage may be that without having to concentrate on the exact wording of a book it may be easier to maintain eye contact, employ gestures, and use voice modulation to capture the listeners' attention (Myers, 1990). Yet, written material may also have its benefits. For instance, studies have shown that written contexts usually provide a richer vocabulary than spoken language (Nation, 2006) and that children's books contain more unique words than child-directed speech or conversations (Montag, Jones, & Smith, 2015). Considering the fact that child and caregiver vocabulary are strongly related to socioeconomic status (Hart & Risley, 1995), this discrepancy may be especially pronounced in families displaying a low socioeconomic status. In addition, the use of written material in contrast to free storytelling offers the children the possibility to get acquainted with letters and may thus facilitate learning to read and to write (e.g., Reese & Cox, 1999). In line with this, Mol and Bus (2011) reported in a recent meta-analysis small to moderate relationships (r = .18 and r = .29) between measures of print exposure (e.g., the frequency of shared book reading) and basic reading skills in preschool and kindergarten children.

Despite the pervasiveness of both activities to foster a broad range of (pre-) literacy skills, little empirical research has been conducted to contrast their effectiveness. Myers (1990), for instance, observed elementary school children's interest and attentiveness during free storytelling and read-aloud sessions. She reported higher rates of laughs, questions, child-initiated responses, and responses to questions during free storytelling, while reading-aloud was marked by higher amounts of behaviours that indicated inattentiveness (e.g., looking away or yawning). In line with this, Trostle and Hicks (1998) found that elementary school children performed better on vocabulary and comprehension measures when they heard a story told freely than when it was simply read aloud. Yet, both studies suffer methodological limitations because their storytelling conditions differed in various ways from the read-aloud condition which might per se explain the results in favour of the storytelling condition. In Myers' (1990) study the storyteller asked considerably more questions in the storytelling condition than in the read-aloud condition (e.g., 30 vs. 2 per story in the younger age group), which is by itself a major determinant of learning from stories (Blewitt et al., 2009). In addition, the storyteller was a professional who was much more comfortable with the storytelling condition. In Trostle and Hicks' (1998) study the storytelling condition seemed by itself to be much more appealing as it involved, for example, the storyteller dressing up as the protagonist as well as lots of pantomimic acting, whereas the read-aloud condition did not contain such potentially attention capturing features.

Another study that compared the effects of both story conditions in preschoolers stems from Isbell et al. (2004). In their study, story reading produced higher gains in fluency (i.e., total number of words) and vocabulary diversity during story retelling. These gains, however, are difficult to interpret as no statistical significance or effect sizes were reported. In addition, the results regarding mean length of utterances as well as the use of formal story conventions, such as a beginning and ending, were mixed indicating no general superiority of any story condition.

In a more recent study with elementary school children Suggate, Lenhard, Neudecker, and Schneider (2013) targeted these methodological problems by designing the conditions to be as closely matched as possible and controlling for structural features found to affect vocabulary acquisition (e.g., story duration, number of words, frequency of occurrence of target items). Using a counterbalanced 2 × 3 mixed design with grade as a between-subjects variable (two vs. four) and story condition as a within-subjects factor (independent reading vs. adult read-aloud vs. free storytelling), they reported the highest gains for children in the free storytelling condition followed by the children in the adult read-aloud condition. However, given the small sample size (n = 37) coupled with not having included a pre-test, it is technically possible that post-storytelling gains were due to Type I error. Additionally, Suggate et al. (2013) did not include measures of story-delivery, thus it is not known to what extent the two conditions differed in terms of interactivity and story narration. Finally, they included Grade 2 and 4 samples, whereas a question in the current study arises as to whether free storytelling is beneficial for preschool vocabulary growth, at a time when children perhaps need this most (Biemiller, 2003).

As none of the studies mentioned in this section provided visual exposure to the text during the story conditions, differences between the conditions cannot be explained by print exposure. In sum, there is some evidence favouring the free storytelling condition regarding its interest to children and also in terms of potential language learning. However, taken as a whole previous research has been marked by methodological drawbacks (missing statistics, no pre-test, sample size) or has targeted elementary school children. Consequently, it remains unclear whether there are any differential effects of story-delivery on vocabulary learning in the important preschool age bracket (Biemiller, 2003).

The Current Study

There is clear evidence that children experience a vast and rapid vocabulary growth in the years preceding school entry. As these differences do not disappear during elementary school, early interventions are clearly needed (Biemiller, 2003). Research indicates that children are able to benefit from environmental linguistic input, to a large extent without explicit teaching of rules, which suggests the importance of incidental learning mechanisms (e.g., Akhtar, 2005; Elley, 1989; Gampe et al., 2012; Robbins & Ehri, 1994). Despite the extensive use of reading and storytelling activities and interventions to foster language development, little empirical research has been conducted to contrast the effectiveness of the underlying method of story-delivery, with respect to whether the story is told freely or read aloud.

On the one hand, free story telling does not require narrators to focus on written material, so that these can therefore direct more attention to the storytelling process as happens in dyadic interactions (e.g., Reese, 2013). Specifically, storytellers can use more eye contact, gestures, and voice modulation, thereby enhancing attention, motivation and potential learning (Isbell et al., 2004; Myers, 1990; Roney, 1996). On the other hand, books typically provide more complex syntactic and lexical structures than free speech and are in general less dependent on the adult reader's language skills (Nation, 2006; Montag et al., 2015). At first glance free storytelling seems to produce better results than simple read-alouds (e.g., Myers, 1990; Suggate et al., 2013; Trostle & Hicks, 1998). However, when taking methodologically problematic (Myers, 1990; Trostle & Hicks, 1998) and mixed results (Isbell et al., 2004) into account, there remains only sparse evidence in favour of free storytelling.

The current study was conducted to extend previous work (e.g., Isbell et al., 2004; Myers, 1990; Suggate et al., 2013; Trostle & Hicks, 1998) to a preschool sample – a time when children's vocabulary is growing rapidly (Biemiller, 2003) and influenced by interventions (Downer & Pianta, 2006; Mol & Bus, 2011; Ramírez-Esparza et al., 2014; Zucker et al., 2013). Second, the current study sought to address key methodological drawbacks of previous studies, by (a) including a pretest and a follow-up, (b) using a large sample size, and (c) controlling for structural differences between the story-delivery conditions (e.g., number of target-word appearance, story duration, narrator's speech rate, narrator's questions and comments). The latter is especially important, as the number of questions and explanations given by the narrator play a pivotal role in learning new words from book reading (e.g., Blewitt et al., 2009). To this end, four research questions were examined.

First, we wanted to test the hypothesis that preschoolers can develop their vocabularies through listening to stories, in the absence of explicit word explanations such as word definitions or questions, as predicted by incidental accounts of word learning (e.g., Swanborn & de Glopper, 1999). In addition to an immediate post-test, we included a two to three-week follow-up to examine the stability of potential learning gains. Third, we sought to determine whether the finding that free storytelling resulted in greater vocabulary gains than read-aloud in elementary school transferred to preschool. To do so, we controlled for relevant structural features, such as story duration, number of target-word appearance, narrator's speaking rate, and questions or explanations provided by the narrator. Fourth, we examined behavioural features of both the narrators (i.e., frequency of eye contact, gestures, voice modulation) and children (i.e., attentiveness, active engagement) that might be more pronounced in free storytelling and also investigated if these were related to novel word learning. Fifth, we sought to explore whether children of all age groups and language abilities profited from the intervention. Potential interaction effects for age and language as a function of story-delivery were of particular interest, because of recommendations that language interventions should be especially target disadvantaged children and as early as possible (Biemiller, 2003).

Study 1

Based upon existing research and theoretical considerations, we formulated three concrete hypotheses. First, as relevant literature typically reports small- to medium-sized gains for incidental learning from listening to stories (e.g., Coyne et al., 2007; Elley, 1989), we argued that both methods of story-delivery, namely read-aloud and free storytelling, would result in small- to medium-sized effects (around $d = 0.2$ to $d = 0.5$, see Cohen, 1992). Second, we assumed storytelling to be related to more eye contact, gestures and voice modulation on the narrators' side and to more attentiveness on the children's side, given that it appears that interactiveness improves language learning (e.g., Reese, 2013). Third, in line with hypothesis two, we expected that storytelling would result in larger vocabulary gains than story reading (Suggate et al., 2013). Finally, as existing results are mixed regarding the relationship between novel word learning and prior language abilities (e.g., Blewitt et al., 2009; Justice, Meier, & Walpole, 2005; Reese & Cox, 1999; Suggate et al., 2013), we refrained from formulating a specific hypothesis. Thus, the examination of these correlations should be regarded as exploratory and interpreted accordingly.

Methods

Participants. Three kindergartens were contacted and asked to distribute letters of participation to the children's parents. For their participation in the study, parents were offered feedback regarding their children's language abilities. In total, the parents of 88 children provided written consent for their child's participation. The participants lived in a middle-sized city (around 130,000 inhabitants) in Germany. From this sample, five cases were excluded (one missed the first measurement point, two had insufficient language abilities, another due to experimenter error, and one switched kindergarten midway through the study). Of the remaining 83 children 47% were female and had a mean age of 57.11 months ($SD = 12.88$).

All but three children were born in Germany, although 16% had one parent and 32% both parents born in a country other than Germany, and 46% of the children spoke a second language at home. All children had at least one parent who left school with a formal educational qualification, whereby the sample distribution of mothers' and fathers' highest levels of education mirrored closely the German population's educational levels (Statistisches Bundesamt, 2015).

Measures. All materials, including the parental questionnaire, the different language tests, as well as the stories, were in German. To augment the response rate of the parental questionnaire, a Russian, Turkish, Romanian, or English translation was distributed alongside the German version, as necessary.

Demographics. Parents completed a questionnaire regarding their and their children's country of birth, languages spoken at home and the parents' highest educational qualification. Response rate was 99%.

Receptive vocabulary. To assess children's receptive vocabulary, a German adaption of the Peabody Picture Vocabulary Test IV (PPVT-IV; A. Lenhard, Lenhard, Segerer, & Suggate, 2015) was used. The 228 stimuli of this test consist of sets of four pictures combined with a spoken word. Children had to point the correct picture matching the word. Split-half reliability (odd-even split of administered items) was .95 in our sample.

Speech comprehension. The subtest Sentence Comprehension (Verstehen von Sätzen; VS) of the Speech Development Test (Sprachentwicklungstest für drei- bis fünfjährige Kinder) SETK 3-5 (Grimm, 2010) was used to assess speech comprehension. Here, the children had either to choose the correct picture from several alternatives or they had to manipulate a given set of objects. Different versions were used for younger (till 3;11 years) and older children (4+ years). Cronbach's α was .78 for the older and .91 for the younger children.

Phonological working memory. The subtests Phonological Working Memory for Non-Words (Phonologisches Arbeitsgedächtnis für Nichtwörter; PGN) and Memory for Sentences (Satzgedächtnis; SG) of the SETK 3-5 (Grimm, 2010) assessed phonological working memory. The subtest PGN required the child to repeat non-words, while whole sentences had to be repeated in the subtest SG. Different versions of the PGN were used for younger (till 3;11 years) and older children (4+ years), whereas SG was only applied to the older group. Cronbach's α was .76 (PGN) and .90 (SG) for the older children and .77 (PGN) for the younger children.

Stimuli. Six of the nine fictional stories, which were used in a previous study targeting elementary school children (Suggate et al., 2013), were selected according to their assumed interest to preschoolers. To ensure suitability they were slightly modified so that the readability index LIX (Björnsson, 1968) of the stories ranged between 23 and 28, with a mean of 26, denoting very easy texts of comparable difficulty (using the calculation program of W. Lenhard & Lenhard, 2014). Each of the stories consisted of 212 to 225 words and contained two different target-words, which we incorporated three times into central parts of the storylines. Due to their shortness, all stories featured a linear plot and the target words were relevant but not essential for understanding the general meaning of the episodes. The target-words were all concrete nouns and selected to have a low frequency of occurrence in everyday language as determined by the occurrence in the lexical database Deutscher Wortschatz (1998-2015). Accordingly, target-words were expected to be largely unfamiliar to the participating children. Examples of these target-words are *Klampfe* (an old German word for guitar), *Kardätsche* (an old German word for comb), and *Remise* (outdated word for coach house). None of the target-words was defined explicitly, but their meaning could be inferred form contextual information. The *Kardätsche*, for example, was used by a little boy to comb the fur of his dog after bathing it. Although it was not necessary to exactly identify the meaning of *Kardätsche*, it represents an important aspect of the event.

Target-word acquisition. To assess target-word acquisition, a test of receptive vocabulary similar to the PPVT-IV was constructed. The experimenter presented each of the twelve target-words orally and the child had to indicate from one of four pictures the one that best matched the word. With a chance performance through random guessing of three, the maximum score was 12 and the minimum score zero. As the words were selected to be unfamiliar to the participating children and chance performance was to be expected in the pretest and was assumed to pertain also to some extent in the posttests, reliability scores were not interpretable.

Observational measures of narrators' and children's behaviour. A second experimenter rated narrators' behaviour (eye contact, gestures, voice modulation, and speech rate) and children's behaviour (attentiveness and active engagement) on a five point Likert-scale (e.g., for attentiveness: 1 = "very attentive" to 5 = "very inattentive"; for active engagement: 1 = "no question or comment" to 5 = "more than five questions or comments"). Inattentiveness was indicated by children moving about in their chairs, yawning, or looking away. Children's active engagement was tallied by the number of questions or comments which were made by the child during storytelling or reading. To allow for subsequent analyses, all ratings were aggregated over the three stories that were freely told and the three stories that were read aloud per child, resulting in scales with a minimum score of three and a maximum score of 15. To check for reliability of the second experimenter's judgment, a third experimenter was assigned to make notes in approximately 15% of the cases. Interrater-reliability was assessed using a one-way random, single-measures intraclass correlation (ICC) (McGraw & Wong, 1996). The mean ICC, calculated according to Olkin and Pratt (1958), was .83, indicating excellent agreement (Cicchetti, 1994). The ICCs of the respective scales ranged from fair (narrator's voice modulation during free storytelling: .51) to excellent (child's questions or comments during free storytelling: 1.00). To control for differences in structural features of the story conditions, the second experimenter also made notes of the number of appearances of each target-word in the story. In addition, the duration of the stories was recorded in seconds.

Procedure. The study took place in separate rooms in the kindergartens and was conducted by trained student research assistants and the first author. Each child was tested individually at two different points for a duration of approximately 40 minutes per session. At the first assessment point the pretest of the target-words was administered first. Next two stories were presented in one block, with each block containing one of each of the story modalities with the order randomly assigned for each pair. Then the first half of the PPVT-IV was conducted. After that, the next two stories, once again one was freely told and one read aloud, were presented. Before telling the last block of stories, another language test was administered as a filler. To reduce recency effects the second half of the PPVT-IV was conducted before the immediate posttest of the target-words was finally administered. Two to three weeks later, the second measurement point took place. Following the delayed posttest of the target-words several language tests were administered.

The story conditions.

Read-aloud condition. In the read-aloud condition one of the experimenters read the stories to the children at a normal pace and speaking with a clear voice (approximately 150-200 words per minute). No visual exposure to the text was provided to the children. The experimenters made no mention of the existence of target-words or that any words were to be remembered. In addition, target-words were not emphasized – they were spoken aloud in the same manner as other words in the story. To avoid artificial differences between the read-aloud and the free storytelling condition, narrators were instructed to behave as naturally as possible, as “if they told or read a story to a younger sibling.” In addition, they were not allowed to ask questions or to give any explanations during reading or storytelling. If a child made a comment or enquiry during the presentation of the story, narrators were instructed to give positive feedback (e.g., “That is an interesting comment/question.”) but not to engage in a discussion. By doing so, we wanted to avoid artificial effects of those variables, which represent important learning factors but are per se unrelated to method of story-delivery.

Free storytelling condition. In the free storytelling condition one of the experimenters retold the gist of the stories, again at a normal pace (approximately 150-200 words per minute) and in a clear voice. Comparable to the read-aloud condition, target-words were not emphasized. The second experimenter who was not telling the story, but making notes regarding the narrator’s and the child’s behaviour, also recorded the number of times that the target-words were mentioned, to ensure that the storyteller mentioned the word exactly three times, to match the read-aloud condition. After the third mention of the target-words, the second researcher was instructed to provide a discrete hand signal to the narrator to indicate that the word should not be mentioned again. All narrators were again instructed to refrain from asking questions or answering to children's enquiries about aspects of the stories in order to keep the experimental conditions as comparable as possible.

Design. As the preschool teachers, parents and their children believed that the purpose of the study was to measure children's language abilities, neither of them were aware of the real purpose of the study, making it single-blind. The design of the study was 2 x 3, with story modality (read-aloud vs. free storytelling) and time (pretest, immediate posttest, delayed posttest) being within-subjects factors. To avoid order and story effects both the order of the stories as well as their assignment to the story conditions were randomized.

Because a small proportion of the children missed or did not want to take part in some of the language ability tests (depending on the test between 3-23%), missing values were imputed using multiple imputation (Lüdtke, Robitzsch, Trautwein, & Köller, 2007). As different language subtests of the Speech Development Test were used for younger (to 3;11 years) and older children (4+ years), multiple imputation was applied to both age groups separately.

Results

Validity check of the story conditions. To ensure internal validity, both story conditions were compared regarding structural features, namely story duration, number of target-word appearances and narrators' speaking rate, which might unduly influence the results. There were no differences regarding total duration in seconds, $t(77) = 0.00$, $p = .99$ (read-aloud: $M = 266.86$, $SD = 29.48$; free storytelling: $M = 267.90$, $SD =$

34.04), observer ratings of narrators' speaking rate, $Z = 1.31$, $p = .12$ (read-aloud: M = 8.38, SD = 1.46; free storytelling: M = 8.51, SD = 1.33), or number of target-word appearances, $Z = 0.70$, $p = .49$ (read-aloud: M = 17.65, SD = 0.86; free storytelling: M = 17.53, SD = 1.04).

Effects of story modality on target-word acquisition. A 2 x 3 ANOVA was conducted on the target-word acquisition scores with story modality (read-aloud vs. free storytelling) and time (pretest, immediate posttest, delayed posttest) as within-subjects factors (see Table 1). As Mauchly's Test of Sphericity was neither significant for time ($p = .36$) nor for the interaction between time and story modality ($p = .37$), sphericity was assumed.

Table 1

Performance on Target-Word Acquisition as a Function of Time and Story Modality

	Pretest	Immediate posttest	Delayed posttest
Story modality	*M (SD)*	*M (SD)*	*M (SD)*
Read-aloud	1.58 (1.11)	1.83 (1.11)	1.61 (1.12)
Storytelling	1.49 (0.86)	1.75 (1.02)	1.61 (1.05)

Note. N = 83. The maximum possible number of words correct is six each for read-aloud and free storytelling. Chance level is 1.5 per story condition.

The main effect of time, $F(2, 164) = 3.59$, $p = .03$, $\eta_p^2 = .04$, was significant, showing a small to medium sized effect of vocabulary acquisition. In contrast, neither the main effect of story modality, $F(1, 82) = 0.16$, $p = .69$, $\eta_p^2 = .002$, nor the interaction effect, $F(2, 164) = 0.17$, $p = .84$, $\eta_p^2 = .002$, reached significance.

For the main effect of time, pairwise comparisons indicated a significant increase from pretest to immediate posttest for target-words, $t(82) = 2.49$, $p = .01$, $d = 0.37$ (d calculated according to Dunlap, Cortina, Vaslow, & Burke, 1996, p. 171, formula 3), and a marginally significant decrease from immediate posttest to delayed posttest, $t(82) = -1.76$, $p = .08$, $d = 0.24$. Pretest and delayed posttest scores did not differ significantly, $t(82) = 0.88$, $p = .38$, $d = 0.11$.

Narrators' and children's behaviour as a function of story modality. In contrast to predictions from relevant literature, read-aloud and free storytelling did not differ regarding observer ratings of narrators' frequency of eye contact, $Z = 0.36$, $p = .72$ (read-aloud: $M = 11.04$, $SD = 2.03$; free storytelling: $M = 11.07$, $SD = 1.76$), gestures, $Z = 1.24$, $p = .21$ (read-aloud: $M = 5.28$, $SD = 2.12$; free storytelling: $M = 4.82$, $SD = 1.83$), voice modulation, $Z = 0.31$, $p = .76$ (read-aloud: $M = 12.76$, $SD = 1.40$; free storytelling: $M = 12.79$, $SD = 1.45$), children's attentiveness, $Z = 1.26$, $p = .21$ (read-aloud: $M = 5.73$, $SD = 2.66$; free storytelling: $M = 5.50$, $SD = 2.44$), or the number of children's questions and comments, $Z = 0.57$, $p = .57$ (read-aloud: $M = 3.66$ $SD = 1.72$; free storytelling: $M = 3.67$, $SD = 1.68$).

Relationship between target-word acquisition, age and language skills. An inspection of Table 2 shows that neither age nor any of the language skills were significantly correlated with raw gain scores of target-words. This was true for both conditions. As expected, language skills showed moderate to high intercorrelations and were significantly related to age.

Table 2

Correlation Coefficients between Target-Word Acquisition, Age and Language skills

	1	2	3	4	5	6	7	8	9
1 Change scores storytelling (target-words) [a]	-	.27*	.13	.14	.31	-.06	.20	.05	-.07
2 Change scores read-aloud (target-words) [a]		-	-.07	.00	-.02	.07	-.15	.01	-.14
3 Age (in months)			-	.66**	.40	.37**	.29	.41**	.42**
4 Receptive vocabulary				-	.82**	.68**	.56**	.24	.64**
5 Speech comprehension (to 3;11 years)					-		.78**		
6 Speech comprehension (4;0+ years)						-		.29*	.60**
7 Phonological memory for non-words (to 3;11 years)							-		
8 Phonological memory for non-words (4;0+ years)								-	.38**
9 Phonological memory for sentences (4;0+ years)									-

Note: Sixty-one children completed the version of the Speech Development Test for older children (4;0+ years), 22 children the version for younger children (to 3;11 years).

* $p < .05$, ** $p < .01$ (two-tailed).

[a] Gains from pretest to immediate posttest.

Discussion

Results of the first study indicated that vocabulary was learned incidentally through listening to stories. On the one hand, the effect was moderate (*d* = 0.37; roughly 0.5 words in total), and additionally the gain was not stable and reduced at the delayed posttest two and a half weeks later. The mean gain of 6% (of the maximum gain possible) from pretest to immediate posttest was similar or somewhat smaller than the gains reported in comparable studies examining incidental learning (e.g., Coyne et al., 2007; Robbins & Ehri, 1994). On the other hand, given the low intensity of the intervention (stories were presented only once including each target word 3 times without further explanation or interaction), the results support the idea that children can catch word meanings “on the fly”, inferring meaning without interaction or reinforcement. Studies reporting higher gains usually rely on more frequent incidental exposures (e.g., Sénéchal, 1997) or add explicit techniques, such as explanations or questions (e.g., Blewitt et al., 2009). Given the mean scores of the pretest of target-words, which did not exceed the chance level, and the slightly higher scores at the immediate posttest, ceiling effects of potential learning gains can be ruled out as a limiting factor.

In contrast to other studies either supporting a Matthew-effect, namely that children with more advanced language abilities show higher word learning gains from listening to stories, (e.g., Blewitt et al., 2009; Shany & Biemiller, 2010; Suggate et al., 2013) or the reverse (e.g., Elley, 1989; Justice et al., 2005), none of the language skills was positively or negatively related to gains through incidental learning. Possibly, the somewhat small overall gains might be the reason for non-existent relations with language skills via the restricted variance attenuating the correlation coefficients. We found, however, a negative relationship between pretest and gain scores (storytelling: $r = -.48$, $p < .01$; read-aloud: $r = -.54$, $p < .01$) and a positive one between posttest and gain scores (storytelling: $r = .67$, $p < .01$; read-aloud: $r = .55$, $p < .01$). Considering that none of the examined language abilities correlated with change scores, it is highly plausible that regression towards the mean plays a pivotal role here. Consequently, these correlations should not be interpreted as indicating a Matthew-effect.

Somewhat surprising and in contrast to previous studies (Myers, 1990; Suggate et al., 2013; Trostle & Hicks, 1998) we found no evidence of a differential effect of story modality. In both conditions narrators' as well as children's behaviour in terms of eye contact, voice modulation, gestures, attentiveness, and active engagement was similar and target-word acquisition through incidental learning from storytelling and read-aloud was comparable in size. This might, however, also be a consequence of the small overall gains.

Interestingly, Elley's studies (1989) that found mean gains of around 15% reported that one of the stories in the second study deviated from that percentage and showed substantially lower gains, namely 4%. Elley's interpretation of the finding was that the story might not have been appropriate for children of this age and thus might not have elicited enough motivation to listen attentively to the story ("lack of involvement", Elley, 1989, p. 185). As the stories used in the current study were originally developed for elementary school children and almost no questions and comments concerning the story were made by the children (78% of the children did not make any interruption during reading aloud, the respective value for free storytelling was also 78%), one cannot exclude the possibility that stories were not inherently motivating for the children, thereby compromising vocabulary gains.

Study 2

To test the possibility that the use of potentially inadequate stories had produced both the small gains and occluded a differential effect of story modality in the first study, a set of six new children's stories were selected from different children's books, which were written by children's book writers for this age span. In line with the argument of the first study, we hypothesized that there is a differential effect of story modality in preschool age using appropriate children's books. Comparable to the first study, we also checked if children's language skills were related to gains. In addition, we checked again for any differences in narrators' or children's behaviour during free storytelling or read-aloud. Due to the approaching end of the preschool year it was not possible to include the whole sample of the first study into the second study. With regards to the possibility of an age related onset of differential effects of method of story-delivery, we focused on the youngest and the oldest children of the original sample and included those, for whom we were able to obtain their parents' consent to participate. As we were primarily interested in short term word acquisition, we only included a posttest of target word acquisition and no follow-up.

Method

Participants. We selected 24 younger (42% female; age: M = 46.33 months, SD = 4.04) and 24 older children (38% female; age: M = 72.83 months, SD = 4.79) out of the original sample of the first study. The age groups did not differ regarding gender, p = .39 (exact test), their parents' migration backgrounds, p = .25, language spoken at home, p = .10, maternal level of education, p = .88, and paternal level of education, p = .91.

Measures. As the participants were selected from the original sample of the first study, all measures – except the new test of target-word acquisition – were drawn from the first study. To test target-word acquisition, a test identical to the target-word test in the first study – but using other words and pictures – was designed.

Stimuli. Six stories were selected from different contemporary children's books, which contained appropriately short stories designed for children between 3 and 6 years of age. The stories were slightly modified. To retain the original story structures, an effort was made to change the stories as little as possible. When feasible, two words per story were replaced with two less well-known alternatives (as determined by the frequency of occurrence in the lexical database Deutscher Wortschatz, 1998-2015), which constituted the new target-words. As we changed as little as possible within the stories, there was greater variability regarding story length (366 to 439 words) and readability index LIX (Björnsson, 1968) (24 to 38) and as indicated by higher LIX values (mean LIX = 30) syntax was slightly more difficult than in the first study. In line with the first study, each target-word appeared three times in the stories.

Procedure. The procedure was similar to that of the first measurement point of the first study. Instead of the PPVT-IV parts of the old test of target-word acquisition were intersected between the stories. The story conditions and general procedure were identical to those used in the first study.

Design. The design was 2 x 2 x 2 mixed, with age group being the between-subjects factor (younger children vs. older children) and story modality (read-aloud vs. free storytelling) and time (pretest, immediate posttest) the within-subjects factors. To avoid order and story effects, both the order of the stories as well as their assignment to the story conditions were randomized within both age groups. Similarly to Study 1, missing data of the language ability tests were imputed (depending on the test between 6-15%). As only 15 children completed the version for younger children of the Speech Development test, multiple imputation could only be applied to the version for older children comprising 33 children.

Results

Validity check of the story conditions. There were no differences regarding narrators' speaking rate, $Z = 0.00$, $p = 1.00$ (read-aloud: $M = 8.38$, $SD = 1.28$; free storytelling: $M = 8.36$, $SD = 1.34$), number of target word appearances, $Z = 1.62$, $p = 0.10$ (read-aloud: $M = 17.50$, $SD = 0.78$; free storytelling: $M = 17.74$, $SD = 0.74$), and story duration in seconds, $t(44) = -1.95$, $p = .06$ (read-aloud: $M = 490.11$, $SD = 43.62$; free storytelling: $M = 510.13$, $SD = 51.60$).

Effects of story modality and age on target-word acquisition. In Table 3 scores on the target-word acquisition test as a function of age group, time and story modality are reported. A 2 x 2 x 2 mixed ANOVA on target-word acquisition was conducted, with age group (younger vs. older children) being the between-subjects factor, time (pretest, immediate posttest) and story modality (read-aloud vs. free storytelling) the within-subjects factors.

Table 3

Performance on Target-Word Acquisition as a Function of Age Group, Time and Story Modality

	Pretest		Immediate posttest		Difference scores	
	Read-aloud	Storytelling	Read-aloud	Storytelling	Read-aloud	Storytelling
Age group	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*
Younger group (*n* = 24)	1.54 (1.06)	1.33 (0.87)	1.58 (1.02)	1.29 (1.08)	0.04 (1.20)	-0.04 (1.49)
Older group (*n* = 24)	1.67 (1.37)	2.04 (1.43)	1.71 (1.23)	2.29 (1.08)	0.04 (0.91)	0.25 (1.26)
n = 48	1.60 (1.22)	1.69 (1.22)	1.65 (1.12)	1.79 (1.18)	0.05 (1.05)	0.10 (1.37)

Note: The maximum possible number of words correct is six each for read-aloud and free storytelling. Chance level is 1.5 per story condition.

Neither the main effect of time, $F(1, 46) = 0.37$, $p = .55$, $\eta_p^2 = .008$, nor the interaction effect Time x Method of story-delivery, $F(1, 46) = 0.06$, $p = .81$, $\eta_p^2 = .001$, were significant, indicating the absence of significant learning gain and giving no indication of differential learning effects as a function of story-delivery.

The main effect of age group was significant, $F(1, 46) = 4.33$, $p = .04$, $\eta_p^2 = .09$, with the group of older children correctly identifying more target-words. However, this effect was modified by an interaction between story modality and age group, $F(1, 46) = 5.41$, $p = .02$, $\eta_p^2 = .11$. Further examination indicated that in the younger age group target-word scores were descriptively higher in the read-aloud condition, $t(23) = 1.27$, $p = .22$, $d = 0.13$, whereas in the older group the opposite was true, $t(23) = -1.97$, $p = .06$, $d = 0.20$, but only on a mere descriptive level. As the three-way interaction as well as none of the other effects were significant, this does, however, not represent differential learning effects as a function of story-delivery.

Narrators' and children's behaviour as a function of story modality. Comparing both story conditions, there were no differences in narrators' frequency of gestures, $Z =.30$, $p = .76$ (read-aloud: $M = 6.72$, $SD = 2.49$; free storytelling: $M = 6.60$, $SD = 2.85$), voice modulation, $Z =.54$, $p = .59$ (read-aloud: $M = 13.17$, $SD = 1.42$; free storytelling: $M = 13.15$, $SD = 1.37$), and eye contact, $Z =.33$, $p = .74$ (read-aloud: $M = 10.73$, $SD = 2.28$; free storytelling: $M = 10.57$, $SD = 2.39$), or children's attentiveness, $Z =.40$, $p = .69$ (read-aloud: $M = 5.94$, $SD = 3.18$; free storytelling: $M = 6.04$, $SD = 2.94$). Slightly more children's questions and comments were, however, observed in the free storytelling condition, $Z = 1.99$, $p = .05$ (read-aloud: $M = 4.00$, $SD = 2.36$; free storytelling: $M = 4.22$, $SD = 2.44$).

Relationship between target-word acquisition, age and language skills. As can be seen in Table 4, none of the language skills was significantly related to target-word acquisition ($p > .05$). As in the first study, interrelations between language skills were moderate to high in size.

Table 4

Correlation Coefficients between Target-Word Acquisition, Age and Language skills

	1	2	3	4	5	6	7	8	9
1 Change scores storytelling (target-words) [a]	-	-.09	.10	.07[b]	-.05 [c]	.32	-.13 [d]	.26	.16
2 Change scores read-aloud (target-words) [a]		-	.05	-.15[b]	.09 [c]	-.15	.11 [d]	-.11	-.34
3 Age (in months)			-	.58**[b]	.59*[c]	.39*	.46[d]	.51**	.46*
4 Receptive vocabulary				-	.85** [c]	.65**	.65*[d]	.22	.63**
5 Speech comprehension (to 3;11 years)					-		.79** [d]		
6 Speech comprehension (4;0+ years)						-		.35	.74**
7 Phonological memory for non-words (to 3;11 years)							-		
8 Phonological memory for non-words (4;0+ years)								-	.55**
9 Phonological memory for sentences (4;0+ years)									-

Note: Thirty-three children completed the version of the Speech Development Test for older children (4;0+ years), 15 children the version for younger children (to 3;11 years). Because of the small sample size multiple imputation could not be applied to the version for younger children.

* p < .05, ** p < .01 (two-tailed).

[a] Gains from pretest to immediate posttest. [b] Correlations between age, change scores and receptive vocabulary were calculated for test version of the Speech Development Test separately, then they were Fisher-Z transformed and averaged according to sample sizes. [c] n = 14. [d] n = 13.

Discussion

Despite using age-appropriate and more interesting stories in the second study, no significant incidental gain through listening to stories was found (*d* = 0.08; computed after Dunlap et al., 1996; roughly 0.15 words in total). In light of the findings from previous studies (e.g., Elley, 1989; Robbins & Ehri, 1994) as well as the first study, this was unexpected, especially since we did not observe retest effects in the children, who were already familiar with the storytelling and testing procedure. To exclude the possibility that our reduction of the sample to *n* = 48 had caused the differing results between both studies, we re-ran our analysis of the first study with the reduced sample of the second study. The learning gain from pretest to immediate posttest Study 1 was 0.57 words in total (*d* = 0.35; computed after Dunlap et al., 1996) and therefore comparable to that of the original sample of roughly 0.5 words (*d* = 0.37).

Considering our selection criteria as well as the LIX values of the stories (Björnsson, 1968), which indicate a low text difficulty (mean LIX was 30), it is unlikely that the stories were too demanding. However, after the insertion of the target-words they might have contained too many words which are unknown in this age. Hsueh-Chao and Nation (2000), for example, recommend a maximum of 2-3% unknown words if adequate story comprehension is to be achieved. Given the fact that these stories were chosen from children's books appropriate for this age and that results were similar for both the younger and the older children, this explanation seems unlikely. In addition, observers' ratings of children's attentiveness as well as their active engagement were rather similar to those of the first study.

Consistent with the first study, there was no evidence of a differential effect of story modality. Both read-aloud as well as free storytelling did not produce different target-word acquisition rates. Comparable to the results of the first study change scores were only related to pretest (storytelling: r = -.59; read-aloud: r = -.52) posttest scores (storytelling: r = .55; read-aloud: r = .37), indicating regression effects. The interpretation of this result is difficult because there was no significant gain at all. Therefore, it is still possible, that free storytelling and read-aloud might under different circumstances differ regarding their impact on incidental vocabulary acquisition from listening.

Interestingly, in contrast to the first study, we found significant differences between the story modalities regarding higher numbers of interruptions during free storytelling, but not regarding other facets of narrators' and children's behaviour. The higher number of interruptions is most likely the main cause of the marginally significant longer story duration in the free storytelling condition. As in the first study, children's interruptions represented in the main part comments about certain words in the stories (e.g., 'Our dog's name is Max!'), but did – with only one exception – not refer to any of the target-words.

General Discussion

Summarizing the results of both studies, we were able to demonstrate vocabulary gains through incidental learning in children of age 3 to 6, but the effects were small (a pooled d = .25 across both studies). Accordingly, children can in principal acquire vocabulary simply by listening, without further interaction or explanation. Regarding the other research questions and as a consequence of the overall small effects, we failed to substantiate differential effect of story modality and we could not verify a link between language predictors and learning gains. Therefore, further explanations as to the lack of vocabulary gains are considered next.

Firstly, there simply might not be any differential effect of story modality for preschoolers. Given the fact that Suggate et al. (2013) found differential effects in second and fourth graders, these differences might be age-related or they might just not appear until formal schooling has taken place. As Suggate et al.'s youngest group consisted of second graders, the onset of this process would have to occur within the first two years of formal schooling. Considering the fact that children not only learn to read and to write during this period, but also to sit quietly and to pay attention to prolonged instructions from the teacher, they may then first acquire the competencies to profit from different types of oral presentation.

Secondly, the study design with six different stories being told only once in one session may have constituted too great a cognitive overload for incidental learning to occur in a meaningful way. Six different short stories employing different characters, story lines and settings might possibly have contained too much incoherent information to map the unknown target-words to certain meanings. Yet, considering the facts that at least in the first study incidental learning actually did occur and that in a previous study (Suggate et al., 2013) a similar procedure (with nine stories) was successfully applied to second-grade elementary school children, who were only around a year and a half older than our oldest children, it is very unlikely that cognitive overload represents the only explanation of our results.

Thirdly, and somewhat intertwined with the second possibility, the operationalization of the storytelling and adult read-aloud condition may have been problematic. In both studies the stories were very short. Mean reading or telling time in the first study was around a minute and a half and in the second study less than three minutes per story. More importantly, as indicated by low LIX values (first study: mean LIX = 26; second study: mean LIX = 30) syntax was easy and the story structures were rather simple. As the narrators read or told each story multiple times per day and knew them almost by heart, this might have resulted in very similar phrasings. In addition, the narrators had to switch multiple times between both story conditions in one session, which might also have led to overly similar types of story presentation. This conclusion is tempting, as the narrators' behaviour did not differ between the conditions whereas it could be expected that a genuine free-telling would fundamentally differ to a shared-reading. In contrast to the cross-condition homogeneity in narrator behaviour, Suggate et al. (2013) used a very similar design and did actually find differences in favour of the storytelling condition. However, as in their study the free storytelling condition was explicitly designed to enable more eye contact and interactivity (Suggate et al., 2013, p. 559), the narrators might have consciously or unconsciously paid attention to differ in terms of behaviour (i.e., eye contact, gestures, and voice modulation). Unfortunately, these variables were not measured in Suggate et al.'s study and we simply do not know whether this could be a viable explanation. To enhance ecological validity in our study, the narrators were instructed to act as naturally as possible, as "if they were telling or reading a story to a younger sibling". But since the students conducting the study were no practiced storytellers and were not informed a priori about our hypotheses concerning assumed differences between the storytelling conditions and hypothesized effects of eye contact, voice modulation, and gestures, this might also have led to very similar read-aloud and free storytelling behaviours. Clearly, a fundamental challenge is to retain the authenticity of the respective conditions, while simultaneously matching the stories for relevant structural features such as story duration or narrator's speech rate.

Regarding the ambiguity of our results, further studies targeting both age groups, preschool and elementary school children, are needed. In order to disentangle the aforementioned lines of argument, it would also be worthwhile to revise the employed study design by using fewer stories and assessing whether the stories were appropriate for the target age group by checking the children's understanding of the story content. In addition, adding one or two repetitions of each story over the span of one or two weeks might increase rates of incidental learning and therefore maximize the possibility to detect potential differential effects of story modality (Stahl & Fairbanks, 1986). It also would be worthwhile to actively manipulate the extent of narrator's eye-contact, gestures and voice modulation to relate them to children's attentiveness and active engagement. Yet, it is also clear that one must abstain from artificially enhancing one of the story conditions (here most likely the free storytelling condition) by adding higher numbers of questions (e.g., Myers, 1990) or dressing as a character in the story (e.g., Trostle & Hicks, 1998). By doing so, differences may well be found, but they would be due to these specific didactical techniques as opposed to simply telling versus reading.

Limitations

Results yielded inconsistent finding across both studies. Whereas the moderate gains from Study 1 are to some extent consistent with findings from previous studies (e.g., Coyne et al., 2007; Robbins & Ehri, 1994), the results of the second study were unexpected because no significant incidental gains from listening to stories were found despite using professionally written short stories. As discussed, we do not think that this was caused by a reduction of the sample size in the second study or an inappropriate story selection. However, we cannot totally discount the latter possibility because we did not check the children’s understanding of the stories and we did not use an objective rating scale to measure whether the children liked the stories or not.

Interestingly, preschoolers also seemed not to profit from being acquainted with the testing procedure and from enhanced story material. Otherwise there should have been greater effects for target-word acquisition in the second study. Additionally, pretesting the target-words did not influence learning outcomes. Indeed, Biemiller and Boote (2006), who examined pretesting target-words as a mean to enhance incidental word learning from listening to stories, demonstrated that pretest performance did not play any significant role in their preschool and elementary school sample. Based on several pilot studies, Elley (1989) came to a similar conclusion. Thus, it seems that younger children have limited ability to use this kind of information to enhance their word learning from listening to stories. It would be worthwhile to explore whether this represents a characteristic specific to younger children or if adolescents and adults are affected in a similar way.

Summary

Both, reading stories aloud and telling stories freely to children at the age of 3 to 6 years may result in vocabulary gains. Children acquire a small number of words even without interacting with the narrator and they do so irrespective of the presentation condition (stories read vs. freely told). These gains, however, seem to be somewhat limited using only one session and thus it would be unwise to rely solely on incidental vocabulary acquisition in educational settings. Regarding the widespread use of storytelling and book reading behaviour, further research targeting differences in narrators’ and children’s behaviour as well as differential learning effects of both conditions is clearly needed.

References

Akhtar, N. (2005). The robustness of learning through overhearing. *Developmental Science, 8*(2), 199–209. doi:10.1111/j.1467-7687.2005.00406.x

Ard, L. M., & Beverly, B. L. (2004). Preschool word learning during joint book reading: Effect of adult questions and comments. *Communication Disorders Quarterly*, *26*(1), 17–28. doi:10.1177/15257401040260010101

Baydar, N., Brooks-Gunn, J., & Furstenberg, F. F. (1993). Early warning signs of functional illiteracy: Predictors in childhood and adolescence. *Child Development, 64*, 815–829. doi: 10.1111/j.1467-8624.1993.tb02945.x

Beck, I. L., & McKeown, M. G. (2007). Increasing young low-income children's oral vocabulary repertoires through rich and focused instruction. *The Elementary School Journal*, *107*(3), 251–271. doi:10.1086/511706

Biemiller, A. (2003). Vocabulary: Needed if more children are to read well. *Reading Psychology, 24*(3-4), 323–335. doi:10.1080/02702710390227297

Biemiller, A., & Boote, C. (2006). An effective method for building meaning vocabulary in primary grades. *Journal of Educational Psychology*, *98*(1), 44–62. doi:10.1037/0022-0663.98.1.44

Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of Educational Psychology*, *93*(3), 498–520. doi:10.1037/0022-0663.93.3.498

Björnsson, C.-H. (1968). *Lesbarkeit durch Lix: Tech. Rep. No. 6* [Readability through LIX]. Stockholm, Sweden.

Blewitt, P., Rump, K. M., Shealy, S. E., & Cook, S. A. (2009). Shared book reading: When and how questions affect young children's word learning. *Journal of Educational Psychology, 101*(2), 294–304. doi:10.1037/a0013844

Christian, K., Morrison, F. J., Frazier, J. A., & Massetti, G. (2000). Specificity in the nature and timing of cognitive growth in kindergarten and first grade. *Journal of Cognition and Development*, *1*(4), 429–448. doi:10.1207/S15327647JCD0104_04

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. doi:10.1037/1040-3590.6.4.284

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Coyne, M. D., McCoach, D. B., & Kapp, S. (2007). Vocabulary intervention for kindergarten students: Comparing extended instruction to embedded instruction and incidental exposure. *Learning Disability Quarterly*, *30*(2), 74–88. doi:10.2307/30035543

Cromley, J. G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology, 99*(2), 311–325. doi:10.1037/0022-0663.99.2.311

Deutscher Wortschatz (1998-2015) [German thesaurus (1998-2015)]: Universität Leipzig. Retrieved from http://corpora.uni-leipzig.de/de

Downer, J. T., & Pianta, R. C. (2006). Academic and cognitive functioning in first grade: Associations with earlier home and child care predictors and with concurrent home and classroom experiences. *School Psychology Review, 35*, 11–30.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*(2), 170–177. doi:10.1037/1082-989X.1.2.170

Durkin, D. (1978-1979). What classroom observations reveal about reading comprehension instruction. *Reading Research Quarterly*, *14*(4), 481–533.

Elley, W. B. (1989). Vocabulary acquisition from listening to stories. *Reading Research Quarterly*, *24*(2), 174–187. doi:10.2307/747863

Gampe, A., Liebal, K., & Tomasello, M. (2012). Eighteen-month-olds learn novel words through overhearing. *First Language*, *32*(3), 385–397. doi:10.1177/0142723711433584

Grimm, H. (2010). *Sprachentwicklungstest für drei- bis fünfjährige Kinder (SETK 3-5)* [Speech Development Test for three to five year old children] (2. überarb. Aufl.). Göttingen: Hogrefe.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: P.H. Brookes.

Hayiou-Thomas, M. E., Harlaar, N., Dale, P., & Plomin, R. (2006). Genetic and environmental mediation of the prediction from preschool language and nonverbal ability to 7-year reading. *Journal of Research in Reading, 29,* 50–74. doi: 10.1111/j.1467-9817.2006.00292.x

Hsueh-Chao, M. H., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*(1), 403–430.

Isbell, R., Sobol, J., Lindauer, L., & Lowrance, A. (2004). The effects of storytelling and story reading on the oral language complexity and story comprehension of young children. *Early Childhood Education Journal*, *32*(3), 157–163. doi:10.1023/B:ECEJ.0000048967.94189.a3

Jenkins, J. R., & Dixon, R. (1983). Vocabulary learning. *Contemporary Educational Psychology*, *8*(3), 237–260. doi:10.1016/0361-476X(83)90016-4

Justice, L. M. (2002). Word exposure conditions and preschoolers' novel word learning during shared storybook reading. *Reading Psychology*, *23*(2), 87–106. doi:10.1080/027027102760351016

Justice, L. M., Meier, J., & Walpole, S. (2005). Learning new words from storybooks. *Language, Speech, and Hearing Services in Schools*, *36*(1), 17–32. doi:10.1044/0161-1461(2005/003)

Kovas, Y., Hayiou-Thomas, M. E., Oliver, B., Dale, P. S., Bishop, D. V., & Plomin, R. (2005). Genetic influences in different aspects of language development: The etiology of language skills in 4.5-year-old twins. *Child Development*, *76*(3), 632–651. doi: 10.1111/j.1467-8624.2005.00868.x

Lenhard, A., Lenhard, W., Segerer, R., & Suggate, S. (2015). *Peabody Picture Vocabulary Test – Revision IV (German Adaption)*. Frankfurt am Main: Pearson Assessment.

Lenhard, W. & Lenhard, A. (2014). Berechnung des Lesbarkeitsindex LIX nach Björnsson [Calculation of Björnsson's readability index LIX] [Software]. Bibergau: Psychometrica. Retrieved from http://www.psychometrica.de/lix.html

Lüdtke, O., Robitzsch, A., Trautwein, U., & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung [Handling of missing data in psychological research: Problems and solutions.]. *Psychologische Rundschau*, *58*(2), 103–117. doi:10.1026/0033-3042.58.2.103

Marulis, L. M., & Neuman, S. B. (2010). The effects of vocabulary intervention on young children's word learning: A Meta-Analysis. *Review of Educational Research*, *80*(3), 300–335. doi:10.3102/0034654310377087

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. doi:10.1037/1082-989X.1.1.30

Mol, S. E., & Bus, A. G. (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin, 137*(2), 267–296. doi:10.1037/a0021890

Montag, J. L., Jones, M. N., & Smith, L. B. (2015). The words children hear: Picture books and the statistics for language learning. *Psychological Science*, *26*(9), 1489–1496. doi:10.1177/0956797615594361

Myers, P. (1990). Stories from print. *Language Arts*, *67*(8), 824–831.

Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly, 20*(2), 233. doi:10.2307/747758

Nation, I. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*(1), 59–82. doi:10.3138/cmlr.63.1.59

Nelson, K. E., Welsh, J. A., Trup, E. M. V., & Greenberg, M. T. (2011). Language delays of impoverished preschool children in relation to early academic and emotion recognition skills. *First Language*, *31,* 164–194. doi: 10.1177/0142723710391887

Ninio, A., & Bruner, J. (1978). The achievement and antecedents of labelling. *Journal of Child Language*, *5*(1), 1–15. doi:10.1017/S0305000900001896

Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *The Annals of Mathematical Statistics*, *29*(1), 201–211. doi:10.1214/aoms/1177706717

Pancsofar, N., & Vernon-Feagans, L. (2006). Mother and father language input to young children: Contributions to later language development. *Journal of Applied Developmental Psychology*, *27*(6), 571–587. doi: 10.1016/j.appdev.2006.08.003

Ramírez-Esparza, N., García-Sierra, A., & Kuhl, P. K. (2014). Look who's talking: Speech style and social context in language input to infants are linked to concurrent and future speech development. *Developmental Science*, *17*(6), 880–891. doi: 10.1111/desc.12172

Reese, E. (2013). *Tell me a story: Sharing stories to enrich your child's world*. New York: Oxford University Press.

Reese, E., & Cox, A. (1999). Quality of adult book reading affects children's emergent literacy. *Developmental Psychology*, *35*(1), 20–28. doi:10.1037/0012-1649.35.1.20

Robbins, C., & Ehri, L. C. (1994). Reading storybooks to kindergartners helps them learn new vocabulary words. *Journal of Educational Psychology*, *86*(1), 54–64. doi:10.1037/0022-0663.86.1.54

Roney, R. C. (1996). Storytelling in the classroom: Some theoretical thoughts. *Storytelling World*, *9*, 7–9.

Rowe, M. L. (2008). Child-directed speech: Relation to socioeconomic status, knowledge of child development and child vocabulary skill. *Journal of Child Language*, *35*, 185–205. doi: 10.1017/S0305000907008343

Sénéchal, M. (1997). The differential effect of storybook reading on preschoolers' acquisition of expressive and receptive vocabulary. *Journal of Child Language, 24*(1), 123–138. doi:10.1017/S0305000996003005

Sénéchal, M., Ouellette, G., & Rodney, D. (2006). The misunderstood giant: On the predictive role of early vocabulary to future reading. In S. B. Neuman & D. K. Dickinson (Eds.), *Handbook of early literacy research* (Vol. 2, pp. 173–182). New York: Guilford Press.

Shany, M., & Biemiller, A. (2010). Individual differences in reading comprehension gains from assisted reading practice: Preexisting conditions, vocabulary acquisition, and amounts of practice. *Reading and Writing*, *23*(9), 1071–1083. doi:10.1007/s11145-009-9196-4

Stahl, S. A., & Fairbanks, M. M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research*, *56*(1), 72–110. doi:10.3102/00346543056001072

Statistisches Bundesamt. (2015). Bildungsstand der Bevölkerung 2015 [Educational level of the population 2015]. Retrieved from https://www.destatis.de/DE/Publikationen/Thematisch/BildungForschungKultur/Bildungsstand/BildungsstandBevoelkerung5210002157004.pdf?__blob=publicationFile

Suggate, S. P., Lenhard, W., Neudecker, E., & Schneider, W. (2013). Incidental vocabulary acquisition from stories: Second and fourth graders learn more from listening than reading. *First Language*, *33*(6), 551–571. doi:10.1177/0142723713503144

Swanborn, M., & de Glopper, K. (1999). Incidental word learning while reading: A meta-analysis. *Review of Educational Research*, *69*(3), 261–285. doi:10.3102/00346543069003261

Trostle, S., & Hicks, S. J. (1998). The effects of storytelling versus story reading on comprehension and vocabulary knowledge of British primary school children. *Reading Improvement*, *35*(3), 127–136.

Walsh, B. A., Sánchez, C., & Burnham, M. M. (2016). Shared storybook reading in Head Start: Impact of questioning styles on the vocabulary of Hispanic dual language learners. *Early Childhood Educational Journal*, *44*(3), 263–273. doi:10.1007/s10643-015-0708-3

Whitehurst, G. J., Falco, F. L., Lonigan, C. J., Fischel, J. E., DeBaryshe, B. D., Valdez-Menchaca, M. C., & Caulfield, M. (1988). Accelerating language development through picture book reading. *Developmental Psychology*, *24*, 552–559. doi:10.1037/0012-1649.24.4.552

Zucker, T. A., Solari, E. J., Landry, S. H., & Swank, P. R. (2013). Effects of a brief tiered language intervention for prekindergartners at risk. *Early Education and Development*, *24*, 366–392. doi: 10.1080/10409289.2012.664763

5 Study 2: More than Words: Narrator Engagement during Storytelling Increases Children’s Word Learning, Story Comprehension, and On-Task Behavior

A version of this chapter was submitted as:

Lenhart, J., Lenhard W., Vaahtoranta, E., & Suggate, S. (submitted). More than words: Narrator engagement during storytelling increases children’s word learning, story comprehension, and on-task behavior. *Manuscript submitted for publication.*

Draft version 15.03.2019. This paper has not (yet) been published and is not therefore the authoritative document of record. Please do not copy or cite without author’s permission.

Abstract

Reading stories to children fosters their language development. An approach rarely investigated is narrators telling stories without reading from text (i.e., free-telling of stories). Free-telling may differ from more commonly employed read-aloud approaches in terms of language complexity and the opportunity to regulate the storytelling process via attention-guiding behavior, such as eye contact and gesticulation. By experimentally separating the influences of language complexity and attention-guiding behavior, the current study tried to shed light on the effect of story-delivery method (free-telling vs. read-aloud) and its underlying mechanisms on novel word acquisition, story comprehension, and children's on-task behavior. In a 4 x 2 mixed-design, with story presentation (*in vivo* read-aloud vs. *in vivo* free-telling vs. audiotaped read-aloud vs. audiotaped free-telling) as a between-subjects factor and time (pretest vs. posttest) as a within-subjects factor, a sample of 60 four- to six-year-old children listened to four short stories in one of the four conditions twice. Target-word learning from pre- to posttest as well as story comprehension were measured. Additionally, in the *in vivo* conditions storyteller and child behavior was coded. Although learning occurred across conditions, *in vivo* free-telling resulted in the largest gains in receptive target-vocabulary and greater story comprehension. In addition, children were less restless and more attentive.

Keywords: language development; free-telling; storytelling; read-aloud; vocabulary acquisition

Introduction

Vocabulary knowledge is the bedrock of day-to-day communication and is crucial for the development of reading and text comprehension skills (Cromley & Azevedo, 2007; Cunningham & Stanovich, 1997; Sénéchal, Ouellette, & Rodney, 2006; Storch & Whitehurst, 2002). Unfortunately, a tremendous individual variation in vocabulary knowledge arises during early childhood (Fernald, Marchman, & Weisleder, 2013; Hart & Risley, 1995) with severe academic disadvantages for those children with a restricted vocabulary (Dollinger, Matyja, & Huber, 2008; Purpura, Hume, Sims, & Lonigan, 2011). These individual differences tend to persist across the school years (Biemiller & Slonim, 2001; Christian, Morrison, Frazier, & Massetti, 2000) if not targeted in a timely manner (Biemiller, 2003).

Despite the widespread use of story-based interventions to address these discrepancies in vocabulary development (e.g., Hargrave & Sénéchal, 2000; Whitehurst, Arnold, Epstein, Angell, Smith, & Fischel, 1994), research examining the importance of specific factors is still incomplete and empirical results are often inconclusive (Wasik, Hindman, & Snell, 2016). In particular, one approach little explored is the effect of the narrator telling stories without reading from text (i.e., free-telling). Free-telling may differ from more commonly employed read-aloud approaches in terms of language complexity (e.g., sentence length, word length, vocabulary breadth) and the opportunity to regulate the storytelling process via attention-guiding behavior (e.g., eye contact, gesticulation). By experimentally separating the influences of language complexity and attention-guiding behavior, the current study tried to shed light on the effect of story-delivery method (free-telling vs. read-aloud) and its underlying mechanisms, namely language complexity and attention-guiding behavior, on novel word acquisition.

Fostering Vocabulary Development Through Stories

Among other influences, overhearing conversations (Akhtar, 2005; Gampe, Liebal, & Tomasello, 2012) and listening to stories (Elley, 1989; Robbins & Ehri, 1994) are key situations for vocabulary development in early and later childhood. In particular, story-based contexts seem to offer rich and interactive opportunities for implicit and explicit language learning to occur because they represent highly ritualized dialogue, in which the narrator guides attention, asks questions, provides explanations and gives feedback (e.g., Dickinson & Smith, 1994; Haden, Reese, & Fivush, 1996; Ninio & Bruner, 1978).

Additional evidence for the effectiveness of story-based interventions stems from experimental studies (for meta-analyses see Marulis & Neuman, 2010, 2013; Mol, Bus, & de Jong, 2009; Mol, Bus, de Jong, & Smeets, 2008; Stahl & Fairbanks, 1986). Typically, these show that studies on incidental vocabulary acquisition (i.e., without direct teaching of words) mainly find small to moderate gains (e.g., Coyne, McCoach, & Kapp, 2007; Elley, 1989) and that the effectiveness of interventions can be enhanced by including explicit elements, such as explanations, questions, or other supporting activities (e.g., Biemiller & Boote, 2006; Silverman, 2007). However, given that these story-based interventions traditionally make use of books, one approach little explored is the effect of the narrator telling stories without reading from text (i.e., free-telling of stories).

Read-Aloud and Free-Telling of Stories

Read-aloud and free-telling of stories provide vocabulary knowledge in the context of a story and offer the opportunity for interaction between listener and narrator (Isbell, Sobol, Lindauer, & Lowrance, 2004). Yet, they also differ regarding central aspects of the storytelling experience, such as language complexity or the opportunity to guide the listener's attention and to tailor delivery to the audience (Roney, 1996).

The use of attention-guiding behavior may be more pronounced in free-telling of stories as it liberates the narrator from the wording and the physical constraints imposed by handling a book. Consequently, it may be easier for the narrator to focus on the interactive storytelling process, by maintaining eye contact with the listeners, and employing gestures to capture attention (Myers, 1990). Furthermore, free-telling also offers more flexibility to react to signs of disinterest in the audience and to re-gain listeners' attention if lost, for example by gesticulation or eye contact (Roney, 1996).

Oral and written language differ also substantially regarding a wide array of surface characteristics. Written language is typically marked by more demanding syntactic and semantic structures, such as a preference for subordinating conjunctions instead of coordinating conjunctions, the use of subjunctive, higher frequency of passive voice, and less repetitions and breaks in the sentence structure (e.g., Akinnaso, 1982; O'Donnell, 1974; Woolbert, 1922). In addition, oral language typically consists of less complex and less diverse lexical structures than written texts (e.g., Hayes, 1988; Hayes & Ahrens, 1988; Montag, Jones, & Smith, 2015; Nation, 2006). For instance, Nation (2006) analyzed the amount of words that are required for comprehension of written and spoken English and concluded that a vocabulary of 6,000 to 7,000 word-families is sufficient for spoken text, but that an 8,000 to 9,000 word-family vocabulary is needed for comprehension of written material. Moreover, Montag et al. (2015) recently underlined the importance of books for children's lexical development because they found that the vocabulary contained in children's picture books was considerably more diverse than child-directed speech and conversations.

The situation concerning adequacy of language complexity is more complicated. Here, free-telling may provide flexibility to adapt language to the audience, for example by rephrasing, and may thus enhance interest and comprehension (Roney, 1996). Consequently, one may be tempted to argue that, in terms of an aptitude-treatment approach (e.g., Connor et al., 2009) free-telling of stories may also provide an opportunity to tailor the wording according to the cognitive or more particularly the language skills of the listeners. Yet, this possible advantage may depend heavily on the language skills (e.g., vocabulary, grammar) and the general storytelling competence of the narrator as well as on sufficient familiarity with the listeners and their skills.

Finally, a clear disadvantage of free-telling is that it does not offer the possibility to acquire early print knowledge due to the absence of text. In shared-reading situations, in contrast, the use of books provides an opportunity for children to become acquainted with letters (e.g., Reese & Cox, 1999) and may thus facilitate learning to read and write (Mol & Bus, 2011).

Empirical Evidence on Free-telling versus Reading-Aloud

Despite the widespread use of both free-telling and shared-reading activities to foster a broad range of language and pre-literacy skills, only a few studies have contrasted their effectiveness so far. Naturalistic studies seem to favor free-telling approaches for vocabulary acquisition as well as children's behavior during storytelling. Thus, in an observational study, Myers (1990) found that children tended to laugh more, ask more questions, and give more responses during free storytelling. Reading-aloud sessions, in contrast, were marked by higher rates of behaviors indicating lack of interest and boredom (e.g., looking away or yawning). Trostle and Hicks (1998) reported higher vocabulary and comprehension scores for free-telling than for reading-aloud in elementary school. In both studies, however, the free storytelling conditions were designed to be much more appealing, hence may not be comparable. For example, Myers' (1990) free-telling condition was more interactive as the storyteller asked considerably more questions, which is by itself a major determinant of learning from stories (Blewitt, Rump, Shealy, & Cook, 2009). In Trostle and Hicks' (1998) study, free-telling was enriched by pantomimic acting and the storyteller being costumed.

In contrast to these studies, experimental studies provide a more inconclusive picture. For instance, Suggate, Lenhard, Neudecker, and Schneider (2013) used one-to-one reading sessions in a sample of elementary school children, with grade as a between-subjects variable (two vs. four) and story-delivery as a within-subjects factor (independent reading vs. adult read-aloud vs. free storytelling). They found the highest gains for both second and fourth grade children in the free storytelling condition followed by the adult read-aloud condition. However, not having included a pretest, it was impossible to measure actual learning gains and it is possible that through counter-balancing story conditions, item material and sequence of presentation, post-storytelling differences were confounded with pre-existing differences. More importantly, apart from audio recordings that indicated that the free-telling took more time and showed more variability in duration, Suggate et al. (2013) did not include direct observational measures of narrator and child behavior. Therefore, it is unclear what differences in delivery existed and whether these related to differences in child behavior or vocabulary gains.

More recently, Lenhart, Lenhard, Vaahtoranta, and Suggate (2018) tried to replicate and extend Suggate et al.'s (2013) finding that free-telling was superior to read-aloud in a sample of three- to six-year-old children. In two studies, they found no significant differences in word learning or in narrator or child behavior. As they reported only modest vocabulary gains in Study 1 and statistically non-significant growth in Study 2, it is, however, unclear whether group differences were masked by floor effects in the measures.

Using group-administered story conditions in contrast to one-to-one sessions, Vaahtoranta, Suggate, Jachmann, Lenhart, & Lenhard (2018) contrasted story delivery (read-aloud vs. free-telling) and two types of storytelling style (explicit vs. elaborative), but also found no significant differences in word learning between freely-told and read-aloud story-delivery conditions. Although there was no main effect of story delivery on child behavior, which comprised ratings of attentiveness and restlessness, children in the explicit storytelling style conditions were less restless when stories were freely told. Additionally, there was a difference in narrator behavior, with storytellers using more voice variation when telling stories than when reading them aloud. In contrast, speaking rate or use of gestures did not differ between the conditions.

A general problem associated with such highly-controlled experimental studies, is a lack of variability in the different conditions. Because narrators memorized the stories for the free-telling condition and thus already knew the stories by heart, they may have had little need to actually read the books in the read-aloud condition, thus obscuring differences that define both delivery-methods in real-world settings. Consequently, as noted by the authors in the studies from Lenhart et al. (2018) or Vaahtoranta et al. (2018), conditions may have differed little in terms of delivery, which in turn may have resulted in the lack of differences in child behavior and word learning.

Finally, it is possible that there is an interaction between children's language skills and type of storytelling. Given that oral language is typically less complex than written texts (Montag et al., 2015; Nation, 2006), less proficient children might profit more from a linguistically less demanding free-telling condition and children that are more proficient could profit more from a linguistically more demanding read-aloud condition. Somewhat inconsistent with this assumption, Suggate et al. (2013) found that free-telling of stories was more effective than read-aloud both in second and in fourth grade and reported a substantial correlation of $r = .46$ between general receptive vocabulary and target-word knowledge at the posttest across both oral modalities. However, as none of the previous studies has examined (e.g., Isbell et al., 2004; Suggate et al., 2013; Vaahtoranta et al., 2018) or found (Lenhart et al., 2018) an interaction between type of storytelling and children's language skills, further research is clearly needed.

The Current Study

Despite the extensive use of story-based activities and interventions to foster language development, little is known regarding the effect of story-delivery, with respect to whether the story is told freely or read aloud. One compelling argument is that free-telling of stories does not require narrators to hold a book in their hands and to focus on following the text, such that they can direct more attention to the storytelling process as it is common in conversations (e.g., Reese, 2013). Consequently, storytellers should be able to employ more eye contact and gesticulation, thereby enhancing children's attention, motivation and their learning gains at the same time (e.g., Myers, 1990; Roney, 1996). However, free-telling as a form of oral language may offer less complex and less diverse lexical and grammatical input (e.g., Montag et al., 2015), but it is not clear whether differences in language complexity may affect the outcome of a single story-telling situation.

The current study was conducted to extend previous research concerning the effect of different types of story-telling, namely free-telling vs. reading aloud, on children's engagement (i.e., restlessness and attentiveness), story comprehension, and vocabulary learning outcomes. In particular, it focused on isolating the influence of narrators' attention-guiding behavior from potential effects of language complexity that have been confounded in previous naturalistic and experimental studies. To do so, we used four story-telling conditions (a) *in vivo* free-telling, assumed to be marked by large amounts of attention-guiding behavior and less complex language, (b) *in vivo* read-aloud, marked by more complex language, but less gesticulation and eye contact, (c) an audiotaped version of free-telling and (d) an audiotaped version of reading aloud. Although the latter conditions comprise no narrator and therefore no attention-guiding behavior, they differ concerning language complexity comparable to their respective *in vivo* conditions.

With regard to our main research questions, namely (a) whether free-telling of stories was superior to reading-aloud and (b) whether this superiority might be explained by the narrator being able to display more attention-guiding behavior, we examined the following hypotheses:

First, as proposed by theoretical accounts and indicated by observational studies (e.g., Myers, 1990; Roney, 1996) we tested whether *in vivo* free-telling of stories better engaged children through more attention-guiding behavior provided by the narrator, expecting to find a more engaged child behavior (i.e., less restlessness and more attentiveness) during *in vivo* free-telling than during *in vivo* read-aloud (Hypothesis 1).

Second, due to greater story-engagement, we predicted the *in vivo* free-telling condition to result in higher vocabulary learning gains (Hypothesis 2a) and better story comprehension (Hypothesis 2b) than the read-aloud conditions and its own audiotaped version. In contrast, we assumed that language complexity had no general effect on vocabulary learning and story comprehension, expecting to find no differences between audiotaped read-aloud and audiotaped free-telling regarding vocabulary learning (Hypothesis 3a) as well as story comprehension (Hypothesis 3b).

Finally, we wanted to examine in an explorative analysis whether there was an interaction between children's cognitive skills and story presentation. Given that, oral language is typically less complex than written texts (Montag et al., 2015; Nation, 2006), it is possible, that in terms of an aptitude-treatment interaction (e.g., Reese & Cox, 1999), less proficient children profit more from a linguistically less-demanding free-telling condition. In contrast, children that are more proficient may profit more from a linguistically more demanding read-aloud condition. The same may also apply for attention-guiding behavior, which may be especially helpful for children displaying low cognitive skills. Consequently, there may be an interaction between children's language skills and type of presentation regarding target-word learning and story comprehension.

Methods

Participants

Children were recruited from three kindergartens in a middle-sized city (approximately 130,000 inhabitants) in Germany. Parents of 61 children provided written consent for their child's participation. One child missed the entire study due to routine absence, so that the final sample consisted of 60 children. Of these children 45% were female and had a mean age of 64.85 months (*SD* = 8.59; min = 50; max = 80). All children except one were born in Germany, although 17% had one parent and 13% both parents born in a country other than Germany, and 25% of the children spoke a second language at home. All children had at least one parent who left school with a formal educational qualification, with 72% of the mothers and 75% of the fathers having received a high-school diploma or a university degree. Consequently, the educational level of our sample was noticeably higher than that of the German population (Statistisches Bundesamt, 2017).

Design and Procedure

Design. To separate differential effects of attention-guiding behavior and language complexity on novel word learning, we used a 4 x 2 mixed-design, with story-telling conditions (*in vivo* read-aloud vs. *in vivo* free-telling vs. audiotaped read-aloud vs. audiotaped free-telling) being a between-subjects factor and time (pretest vs. posttest) a within-subjects factor. Story comprehension was only assessed at the posttest. Story-telling condition was a between-subjects factor so that children were not required to participate in all conditions, listening to a challenging number of 16 different stories. The children were randomly assigned to one of the four between-subject conditions.

Procedure. The study spanned approximately across two weeks and comprised three one-on-one sessions per child, which took place in separate rooms in children's kindergartens and were conducted by trained student research assistants. In the first session, pretests of target words and of language skills were administered. In the second session, four stories were presented directly one after another in a randomized order for the first time to each child. In the final session, that took place around four days later (M = 3.69, SD = 3.72), the same four stories were presented for the second time, followed by an immediate posttest of receptive target-word acquisition and story comprehension.

Consequently, each child heard a total of four stories twice, all in one story-telling condition.

During the testing and the story presentation, only the research assistant and the child were present in the room. One of four trained research assistants (three female, one male) presented the four stories to the child, with the same research assistant conducting the story presentation sessions and the immediate posttest for a specific child. Each research assistant was scheduled to provide story presentations in all four conditions (*in vivo* read-aloud, *in-vivo* free-telling, audiotaped read-aloud, audiotaped free-telling). Due to one child missing the story presentation phase, random child absence, and scheduling problems in the kindergartens, we could not perfectly balance the research assistant assignment to the story presentation conditions. Consequently, the four research assistants finally acted as narrators for 13, 14, 16 and 17 children, with one of the four research assistants not providing *in vivo* free-telling. Although there was no significant relationship between story presentation condition and research assistant/narrator, $\chi^2(9) = 8.15$, $p = .519$, Cramer's V = .21, we included additional analyses for target-word learning and story comprehension in which we controlled for differences in narrator assignment.

Finally, in the *in vivo* conditions, during one of the two story presentation sessions an independent observer rated narrator and child behavior. The observers were also trained research assistants and blind to the goal of the study. Moreover, they did not participate as narrators in the study.

The Story Conditions

***In vivo* conditions.** In the *in vivo* read-aloud condition, one of the four research assistants read the stories to the children at a normal pace and speaking with a clear voice (approximately 150-200 words per minute). The experimenters made no mention of the existence of target-words or that any words were to be remembered. There was no emphasis on the target-words – they were spoken aloud in the same manner as other words in the story. In the *in vivo* free-telling condition, one of the experimenters retold the gist of the stories, again at a normal pace (approximately 150-200 words per minute) and in a clear voice. As with the read-aloud condition, target-words were not emphasized. Most importantly, to obtain comparability with the other conditions, narrators were instructed to use the four target-words per story exactly once.

To avoid unintended artificial similarities in narrator story-delivery of the free-telling and read-aloud conditions, we refrained from extensive training and memorization of the stories and instructed the narrators to read or tell the stories in a natural manner. Finally, as established in previous studies, questions or comments asked by the narrator have a beneficial effect on word learning from stories (e.g., Biemiller & Boote, 2006; Blewitt et al., 2009; Elley, 1989), we wanted to avoid confounding effects of story presentation in the current study and those caused by differing use of questions or comments. Consequently, narrators were not allowed to ask questions or to give any explanations during reading or free-telling. If a child commented or enquired during the presentation of the story, narrators were instructed to give positive feedback (e.g., “That is an interesting comment/question.”) but not to engage in a discussion.

Audiotaped conditions. In the audiotaped condition of each story-delivery method, the experimenter presented the stories in a randomized order to the children. To do so, all four narrators had recorded a read-aloud and two free-telling versions of each story. We used two different audiotape versions for each story in the free-telling condition to match the *in vivo* condition as closely as possible. One version was assigned randomly to the first story session, the other to the second story session.

Stimuli

Stories. Four fictional stories that were suitable for children aged between four and six years were selected from commercial storybooks and enriched with target-words. The stories were of comparable length (383 to 399 words) and contained four different target-words, which we incorporated once each into central parts of the storyline. To include the target-words, we either slightly adjusted the stories by including an additional sentence or – where feasible – simply replaced a word in the story with a rarer synonym. To ensure comparability between conditions, no pictures were shown during read-aloud or free-telling.

Target words. We used only four target-words per story, to ensure that narrators were able to tell the stories freely, using each target-word exactly once. The 16 target-words comprised only concrete nouns. Consistent with research examining word learning from stories and to avoid ceiling effects (e.g., Blewitt et al., 2009; Lenhart et al., 2018, Vaahtoranta et al., 2018), target-words were selected to have a low frequency of occurrence in children's books as determined by the frequency in the lexical database ChildLex (Schroeder, Würzner, Heister, Geyken, & Kliegl, 2015). Accordingly, target-words were expected to be largely unfamiliar to the participating children. Examples of these target-words are *Kodex* (an old German word for book; English: *codex*) and *Remise* (outdated word for coach house or depot). None of the target-words was defined explicitly, but their meaning could be inferred form contextual information. The *Kodex*, for example, was used by a little wizard to read and learn new magic spells. Thus, although it was not necessary to identify the exact meaning of *Kodex* to comprehend the story line, the children could easily guess its meaning.

Measures

All materials, including the parental questionnaire, the different language tests, as well as the stories, were presented in German.

Demographics. All parents completed a questionnaire regarding their own and their children's country of birth, languages spoken at home and their highest educational qualification.

Language covariates.

General receptive vocabulary. To explore a potential interaction between children's general vocabulary knowledge and storytelling conditions, we included a test for general vocabulary. A German adaption of the Peabody Picture Vocabulary Test-4 (PPVT-4; A. Lenhard, Lenhard, Segerer, & Suggate, 2015) was used to assess children's receptive vocabulary knowledge. Children had to point out the correct picture from four alternatives matching an orally presented word. Split-half reliability (odd-even split of administered items) was $r = .92$.

Phonological working memory. Phonological working memory was included as a potential control variable, because of its role for the processing of novel sound patterns and novel word learning (Baddeley, 2003).The subtest Phonological Working Memory for Non-Words of the SETK 3-5 (Grimm, 2010) assessed phonological working memory. The subtest required the child to repeat non-words. The scale consistency amounted to $\alpha = .77$.

Dependent variables.

Receptive target vocabulary. To assess target-word acquisition as a dependent variable, a test of word recognition similar to the PPVT-4 was constructed. The experimenter presented each of the 16 words orally and the child had to indicate from one of four pictures the one that best matched the word. With a chance performance through random guessing of four, the maximum score was 16 and the minimum score zero. Cronbach's α was .50 at the pretest and .64 at the immediate posttest. As the words were selected to be unfamiliar to the participating children, we expected performance to be influenced by random guessing at the pretest and, albeit to a somewhat lesser degree, at the posttest. Consequently, we expected low reliability scores at pretest, with an increase at posttest as a by-product of word learning. To avoid frustration caused by largely unfamiliar target-words and to verify comprehension of instructions, we also included four well-known control-words within the test.

Story comprehension. To measure story comprehension a set of three open-ended questions per story (i.e., 12 questions in total) were constructed. The questions focused on different aspects of the stories. Answers of the children were audio recorded and transcribed. Correct answers received two points, partially correct answers one point and wrong answers zero points. All answers were coded independently by two raters. Inter-rater reliability was assessed using a two-way mixed, single-measures intra-class-correlation (ICC; c.f. McGraw & Wong, 1996). ICC for consistency was .97, indicating excellent agreement (Cicchetti, 1994). Finally, deviating ratings were discussed until agreement for each case was achieved. Cronbach's α was .84.

Child and narrator behavior during in vivo conditions. In the *in vivo* conditions, narrator behavior (eye contact, gesticulation, and voice modulation) as well as child behavior (motoric restlessness, attentiveness, and active engagement) were rated by an observer during one storytelling session per child. The first author randomly chose one of the two story presentation sessions per child to be observed. The observers rated each of the six variables on a five-point Likert scale per story, resulting in a total of four ratings for each variable per child (the observation protocol is included in the electronic supplement material).

Unfortunately, parents did not provide their consent to record video tapes during the sessions, confining us to simultaneous ratings of child and narrator behavior. As we were not allowed to make video tapes of the sessions and we assumed that the presence of three adults (i.e., one narrator and two observers) would be intimidating and distracting and therefore may have an impact on children's learning, we used results from a small accompanying study with 14 children who were not included in the present study to check the reliability of the observational judgments. Interrater-reliability was assessed using a two-way random, single-measures intra-class-correlation (ICC; c.f. McGraw & Wong, 1996). The ICCs for narrator gesture usage (.88) and eye contact (.86) were excellent (Cicchetti, 1994). ICC for children's motoric restlessness (.64) and children's attentiveness (.49) were good or fair, but those for narrators' voice modulation (.26) and active child engagement (-.11) were poor. Due to poor interrater-agreement, the latter variables were not included in the analyses.

Language characteristics.

Language content. We checked implementation fidelity of free-telling conditions by examining similarity in language content between the original written material and the freely told stories. Here, we expected a considerable degree of overlap, but no perfect match between written and freely told stories. To evaluate semantic content across conditions, we employed latent semantic analysis (LSA), which returns a coefficient $cos\ \alpha$ that ranges between zero and one, with a high coefficient indicating greater semantic overlap (W. Lenhard, Baier, Endlich, Schneider, & Hoffmann, 2013).

Language surface complexity. We also checked implementation fidelity of the conditions concerning language surface complexity (e.g., word length, sentence length) between the original written material and the freely told stories. Here, we expected freely told stories of the audiotaped and the *in vivo* condition to be of similar complexity, but to be less complex than the original written material.

To examine differences in language surface complexity, we transcribed all stimuli of the audiotape free-telling condition (32 audiotapes; 8 per narrator) and two randomly selected *in vivo* free-telling sessions per narrator (24 recordings; due to schedule problems and child absence, one of the four narrators did not provide *in vivo* free-tellings). The transcripts were then analyzed with the text analysis software *Ratisbon tool for text analysis* (*Regensburger Analysetool für Texte*; RATTE; version 1.6.1; Wild & Pissarek, 2016). For indicators of surface complexity, we used two different readability indices. The readability index LIX (Björnsson, 1968) is calculated based on the mean sentence length and the frequency of long words within a text, returning a coefficient usually ranging between 15 and 80, with lower values depicting easier texts. The FLESCH index (Flesch, 1948) is calculated based on the mean word and sentence length, ranging between 0 and 100, with higher scores denoting easier language.

Data Preparation and Analysis

Data preparation and analysis were conducted using IBM SPSS 23 and R (version 3.5.2; R Core Team, 2018). To examine word learning from listening to stories and story comprehension as a function of story presentation, we used mixed-effects modeling. Following recommendations from psycholinguistic research (Baayen, Davidson, & Bates, 2008; Jaeger, 2008; Quené & van den Bergh, 2008), we modeled participants and target-word/comprehension items as Subject X Item crossed random effects. Prior to analyses, categorical independent variables were effect coded and continuous variables were standardized (Cohen, Cohen, West, & Aiken, 2003). As the target-word recognition task resulted in binary outcomes, namely 0 or 1 point per word, we used the *glmer*-function of the R package *lme4* (version 1.1-21; Bates, Mächler, Bolker, & Walker, 2015). For the analysis of story comprehension scores that were scored on an ordinal scale, we used the R package *ordinal* (version 2018.8-25; Christensen, 2018). Contrasts comparing individual groups were computed using the R package *emmeans* (version 1.3.3; Lenth, 2019). Effect sizes for regressions coefficients within mixed-model analyses are reported as *Odds Ratio* (*OR*), those for individual group comparisons as Cohen's *d*.

As our goal was confirmatory hypothesis testing, we included all random slopes for all fixed effects that varied within subjects or items in a first step, thereby specifying a maximal random structure (Barr, 2013; Barr, Levy, Scheepers, & Tily, 2013). However, due to the high number of parameters to be estimated, these models are often overparametrized or fail to converge (Bates, Kliegl, Vasishth, & Baayen, 2015), so that additional model building strategies are necessary (e.g., Bates, Kliegl, et al., 2015; Hoffman & Rovine, 2007; Linck & Cunnings, 2015). Therefore, in our analyses, we followed guidelines provided by Bates, Kliegl, et al. (2015). In a first step, we used principal component analysis of the random effect structure (provided in the *RePsychLing* package; Baayen, Bates, Kliegl, & Vasishth, 2015) to identify the number of variance components supported by the data (not provided for models fitted with the *ordinal* package). In a second step, we compared the goodness of fit of nested models with likelihood ratio tests and AIC-values, starting with dropping the highest order interaction term. The resulting model was considered the optimal linear-mixed-model for the data (Bates, Kliegl, et al., 2015).

Missing data. The amount of missing data was small for language covariates (phonological working memory: missing = 5%), the target-word recognition test (missing = 3% at the posttest), and the story comprehension test (missing = 5%). As mixed-models are robust against small proportions of missing data (Quené & van den Bergh, 2008), we did not impute missing data.

Results

Preliminary Analyses

Equivalence of experimental groups. The descriptive statistics for the four groups are displayed in Table 1. As intended by the randomized assignment to the experimental conditions, the four story-presentation groups did not differ regarding gender composition, $\chi^2(3) = 3.49$, $p = .338$, age, $F(3, 56) = 1.55$, $p = .213$, phonological working memory, $F(3, 53) = 1.16$, $p = .335$, general receptive vocabulary, $F(3, 56) = 1.19$, $p = .320$, or knowledge of control words, $F(3, 56) = 0.34$, $p = .798$, and target words at the pretest, $F(3, 56) = 1.42$, $p = .247$.

In addition, children were able to correctly answer the four well-known control words at both measurement points, that were interspersed within the target-word recognition task, (pretest: $M = 3.95$, $SD = 0.22$, $Min = 3$, $Max = 4$; posttest: $M = 3.98$, $SD = 0.13$, $Min = 3$, $Max = 4$), indicating that all children understood the testing procedure.

Table 1

Descriptive Statistics for the Experimental Groups in the Final Sample

	In vivo **Read-aloud (*n* = 16)** *M* (*SD*)	***In vivo*** **Free-telling (*n* = 14)** *M* (*SD*)	**Audiotaped Read-aloud (*n* = 15)** *M* (*SD*)	**Audiotaped Free-telling (*n* = 15)** *M* (*SD*)
Demographics				
Gender (female)	31%	64%	47%	40%
Age (months)	63.19 (7.59) [51, 77]	61.86 (8.61) [53, 80]	67.60 (6.52) [56, 76]	66.67 (10.67) [50, 80]
Language skills				
General receptive vocabulary	105.88 (31.70)	104.43 (28.69)	122.93 (21.90)	109.00 (35.52)
Phonological working memory	11.29 (3.56)[a]	12.77 (3.49)[b]	13.53 (3.58)	13.07 (3.01)
Control words				
Word recognition pretest	3.94 (0.25)	3.93 (0.27)	4.00 (0.00)	3.93 (0.26)
Word recognition posttest	3.94 (0.25)	4.00 (0.00)[b]	4.00 (0.00)	4.00 (0.00)[a]
Target words				
Word recognition pretest	8.25 (2.11)	8.07 (3.00)	9.67 (2.64)	8.13 (2.07)
Word recognition posttest	8.38 (3.22)	10.38 (2.33)[b]	10.20 (2.68)	8.79 (2.69)[a]
Story comprehension test	12.67 (6.08)[c]	16.31 (4.53)[b]	15.13 (5.28)	14.58 (5.69)[a]
Narrator behavior				

Eye contact	1.62 (0.88)[c]	4.92 (0.16)[b]	-	-
Gesture usage	1.27 (0.62)[c]	3.75 (0.96)[b]	-	-
Child behavior				
Restlessness[d]	3.85 (0.86)[c]	4.50 (0.52)[b]	-	-
Attention	3.78 (0.82)[c]	4.68 (0.66)[b]	-	-

Note: For language skills, raw scores are reported. Target-word recognition test: maximum score = 16, chance performance = 4. Control-word recognition tests: maximum score = 4, chance performance = 1. Story comprehension test: maximum score = 24. Narrator and child behavior: maximum score = 5.

[a] *n* = 14. [b] *n* = 13. [c] *n* = 15. [d] Restlessness is coded inversely, with higher values representing lower restlessness.

Implementation fidelity of the experimental conditions. To check for implementation fidelity of experimental conditions, we analyzed language content and surface characteristics as well as narrator behavior during *in vivo* story presentation.

Language content. Using a latent semantic analysis, we examined semantic agreement of *in vivo* and audiotaped free-telling among one another and with original texts that were employed in read-aloud conditions. The latent semantic analysis coefficient indicated that the audiotaped free-telling ($cos\ \alpha$ = .71) and the *in vivo* free-telling versions ($cos\ \alpha$ = .63) were similar in content to the written stories and that both coincided to a high degree ($cos\ \alpha$ = .72). In addition, an examination of the recordings of the *in vivo* free-telling condition indicated that in 98% of the cases (431 out of 440) target-words were mentioned only once per story, as planned. However, four times a target-word was mentioned twice, two times it was used thrice, one time it was mentioned four times, and in two instances a narrator forgot to include a target-word.

Language surface characteristics. To check comparability between audiotaped free-telling and *in vivo* free-telling, we compared both conditions regarding surface language complexity using the software program RATTE. Both conditions were highly similar regarding LIX (audiotaped free-telling: M = 33.80, SD = 3.62; *in vivo* free-telling: M = 34.56, SD = 4.34; difference: d = 0.19) and FLESCH (audiotaped free-telling: M = 74.60, SD = 4.80; *in vivo* free-telling: M = 75.67, SD = 5.63; difference: d = 0.19), denoting easy or fairly easy texts. Additionally, as expected both conditions provided less complex language input than the original texts (LIX: M = 36.25, SD = 4.78; d = 0.53; FLESCH: M = 71.17, SD = 6.46; difference: d = 0.74). Yet, despite medium-sized effects between free-telling and read-aloud versions, the original stories still denoted easy (LIX) or fairly easy (FLESCH) texts.

Comparably to Suggate et al.'s (2013) study, the *in vivo* free-telling and *in vivo* read-aloud condition differed regarding duration of story presentation, $t(17.538) = 3.45$, $p = .003$, $d = 1.32$. *In vivo* free-tellings of the stories were on average 20 seconds longer than *in vivo* read-alouds ($M_{\text{free-telling}} = 190.93$ seconds, $SD_{\text{free-telling}} = 20.52$; $M_{\text{read-aloud}} = 170.41$ seconds, $SD_{\text{read-aloud}} = 9.24$). As expected, *in vivo* free-tellings showed also considerably more variation regarding story duration. However, mean story duration was neither in the *in vivo* read-aloud nor in the *in vivo* free-telling condition related to target-word gains, $r_{\text{read-aloud}}(16) = .02$, $p = .954$, $r_{\text{free-telling}}(13) = .20$, $p = .514$.

Narrator behavior during in vivo story presentation. In accordance with the finding that free-telling offers more opportunity for narrators to use gestures and to make eye contact with the child, our results indicated that the narrators used indeed greater amounts of eye contact, $t\ (15.03) = 14.28$, $p < .001$, $d = 5.05$, and gesticulation, $t(26) = 8.27$, $p < .001$, $d = 3.14$, during *in vivo* free-telling than during *in vivo* read-aloud. Taken together, the analyses of language content and surface characteristics as well as of narrator behavior indicate high implementation fidelity of the story presentation conditions.

Effects of Story-Delivery on Child Behavior during *in vivo* Story-Telling

In line with our first hypothesis, which assumed that children were more engaged during free-telling of stories, our results indicated that during free-telling of stories children were less restless, $t(27) = 2.43$, $p = .022$, $d = 0.90$, and paid more attention to the stories, $t(27) = 3.23$, $p = .003$, $d = 1.20$ (see Table 1 for the descriptive values).

Effects of Story-Delivery on Target-Word Acquisition

To test the influence of story presentation on target-word acquisition (Hypotheses 2a and 3a), we calculated a mixed-model with time and story presentation and their interaction as fixed effects and subject and item as crossed random effects. As the assignment of narrators to the four story-presentation conditions could not be perfectly counter-balanced (see methods section), we also included the narrators as a control variable. Analysis of the random structure indicated that a model including only random intercepts for subject and item was appropriate.

As shown in Model 1 (see Table 2), there was a significant main effect of time (*OR* = 1.15), indicating target-word learning from pre- to posttest. This main effect was modified by a significant interaction with story presentation. In contrast to the other presentation groups, which did not differ from average learning, *in vivo* free-telling led to significantly higher target-word learning gains (*OR* = 1.24). Pairwise one-tailed contrasts indicated that, in line with Hypothesis 2a, *in vivo* free-telling was better than *in vivo* read-aloud, *Estimate* = 0.69, *SE* = 0.31, $p = .015$, $d = 0.37$, audiotaped free-telling, *Estimate* = 0.52, *SE* = 0.32, $p = .050$, $d = 0.29$, as well as audiotaped read-aloud, *Estimate* = 0.52, *SE* = 0.32, $p = .051$, $d = 0.29$. In concord with Hypothesis 2b, audiotaped read-aloud, and audiotaped free-telling did not differ regarding target-word learning ($p = .998$), excluding the possibility that general differences in language complexity may have acted as the underlying mechanism. Finally, *in vivo* read-aloud did not differ from audiotaped read-aloud ($p = .620$) and audiotaped free-telling ($p = .621$), indicating that the mere presence of a narrator did not affect target-word learning.

Table 2

Effects of Story-Delivery on Target-Word Acquisition Controlling for Narrator (Model 1), and for General Vocabulary (Model 2)

	Model 1			Model 2		
	OR (95%-CI)	*SE*	*p*	*OR (95%-CI)*	*SE*	*p*
Fixed Effects						
(Intercept)	1.42 (0.75 – 2.71)	0.33	.283	1.88 (0.77 – 4.56)	0.45	.165
Time (posttest)	1.15 (1.04 – 1.28)	0.05	**.009**	1.14 (1.02 – 1.27)	0.06	**.020**
Group (*in vivo* free-telling)	0.94 (0.66 – 1.34)	0.18	.740	1.08 (0.87 – 1.35)	0.11	.492
Group (audiotaped read-aloud)	1.40 (1.00 – 1.94)	0.17	**.047**	1.08 (0.87 – 1.33)	0.11	.505
Group (audiotaped free-telling)	0.84 (0.60 – 1.18)	0.17	.317	0.89 (0.71 – 1.10)	0.11	.265
Narrator (A)	0.83 (0.60 – 1.14)	0.16	.243	0.87 (0.71 – 1.06)	0.10	.179
Narrator (B)	0.65 (0.45 – 0.93)	0.19	**.020**	0.69 (0.55 – 0.87)	0.12	**.002**

Narrator (C)	1.67 (1.19 – 2.33)	0.17	**.003**	1.33 (1.06 – 1.66)	0.11	**.012**
General Vocabulary Knowledge				2.31 (1.54 – 3.47)	0.21	**<.001**
Time X Group (*in vivo* free-telling)	1.24 (1.02 – 1.50)	0.10	**.029**	1.22 (1.01 – 1.48)	0.10	**.040**
Time X Group (audiotaped read-aloud)	0.95 (0.79 – 1.15)	0.09	.624	0.97 (0.80 – 1.17)	0.10	.764
Time X Group (audiotaped free-telling)	0.95 (0.79 – 1.15)	0.09	.620	0.94 (0.78 – 1.13)	0.10	.496
Random Effects						
Intercept variance, $_{\text{Subject}}$	0.39			0.04		
Intercept variance, $_{\text{Item}}$	1.57			3.12		
General Vocabulary Knowledge variance, $_{\text{Item}}$				0.56		
N_{Subject}	60			60		
N_{Item}	16			16		
Observations	1886			1886		
AIC	2157.7			2041.0		
logLik	-1065.9			-1004.5		

Note: Models were calculated using the *glmer* function of the *lme4* package (link function = logit). Categorical variables were effect coded and continuous variables were standardized.

As we were also interested in the influence of general vocabulary knowledge on target-word learning and potential interactions with story presentation, we calculated a second model. In a first step we added children's general vocabulary knowledge (the standardized PPVT score) to Model 1, which led to a significant model improvement, $\chi^2(1) = 64.55$, $p < .001$. In the subsequent steps, neither the additional inclusion of the interaction between time and PPVT score, $X^2(1) = 0.49$, $p = .486$, nor the full model comprising a three-way-interaction between time, PPVT score, and story presentation, $\chi^2(7) = 7.82$, $p = .349$ led to any further improvement, indicating that there were no differential learning effects as a function of general vocabulary knowledge. Analysis of the random structure indicated that a model including a random intercept for subject and item as well as a random item slope for general vocabulary knowledge were appropriate. As shown in Model 2 (see Table 2), general vocabulary knowledge was positively related to target-word knowledge (*OR* = 2.31) and its inclusion reduced the variance between the children, but it did not change the pattern of results (see Figure 1).

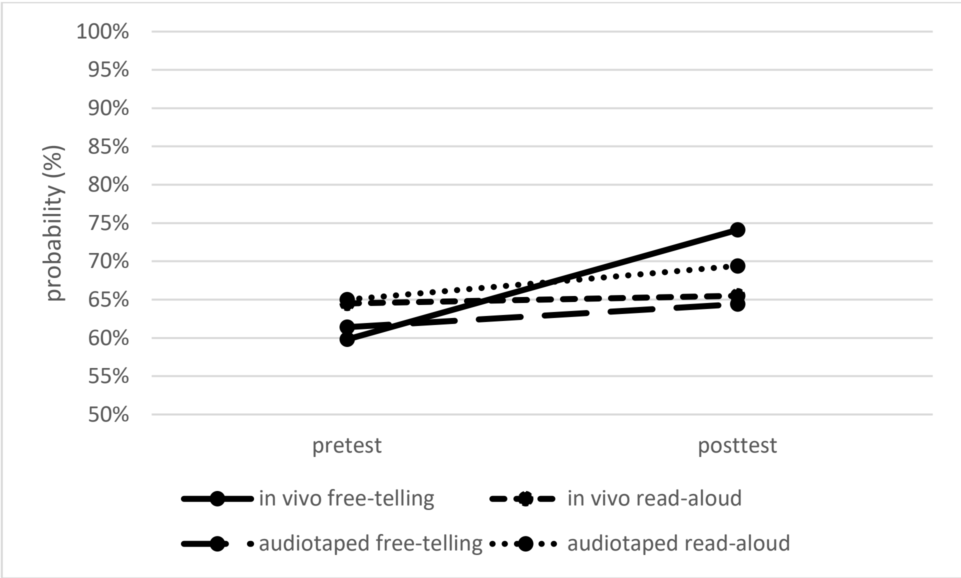

Figure 1. Target-word acquisition as a function of story presentation controlling for narrator influences and children's general vocabulary (Model 2).

Effects of Story-Delivery on Story Comprehension

To test the influence of story presentation on story comprehension (Hypotheses 2b and 3b), we calculated a mixed-model with time and story presentation and their interaction as fixed effects and subject and item as crossed random effects. Again, we also included the narrators as a control variable. Analysis of the random structure indicated that a model including only random intercepts for subject and item was appropriate.

Table 3

Effects of Story-Delivery on Story Comprehension Controlling for Narrator (Model 3), and for General Vocabulary Knowledge (Model 4)

	Model 3			Model 4		
	OR (95%-CI)	*SE*	*p*	*OR (95%-CI)*	*SE*	*p*
Fixed Effects						
(Intercept: 0|1)	0.23 (0.13 – 0.42)	0.30	**<.001**	0.24 (0.14 – 0.42)	0.28	**<.001**
(Intercept: 1|2)	1.22 (0.68 – 2.19)	0.30	.499	1.31 (0.77 – 2.22)	0.27	.326
Group (*in vivo* free-telling)	1.34 (0.71 – 2.53)	0.32	.362	1.60 (1.04 – 2.48)	0.22	**.034**
Group (audiotaped read-aloud)	1.16 (0.65 – 2.07)	0.29	.609	0.80 (0.53 – 1.19)	0.21	.274
Group (audiotaped free-telling)	0.93 (0.51 – 1.69)	0.31	.805	1.06 (0.70 – 1.61)	0.21	.770
Narrator (A)	0.71 (0.40 – 1.26)	0.29	.243	0.83 (0.56 – 1.22)	0.20	.342

Narrator (B)	0.52 (0.27 – 1.00)	0.33	**.050**	0.57 (0.36 – 0.89)	0.23	**.013**
Narrator (C)	2.49 (1.37 – 4.53)	0.31	**.003**	1.41 (0.92 – 2.18)	0.22	.117
General Vocabulary Knowledge				3.43 (2.32 – 5.05)	0.20	**<.001**
Random Effects						
Intercept variance$_{\text{, Subject}}$	1.31			0.40		
Intercept variance$_{\text{, Item}}$	0.67			0.68		
General Vocabulary Knowledge variance, $_{\text{Item}}$				0.18		
N_{Subject}	57			57		
N_{Item}	12			12		
Observations	684			684		
AIC	1276.96			1227.31		
logLik	-628.48			-600.66		

Note: Models were calculated using the *clmm* function of the *ordinal* package (link function = logit; threshold = flexible). Categorical variables were effect coded and continuous variables were standardized.

As shown in Model 3 (see Table 3), there was no significant main effect of experimental condition, indicating that the children could equally well answer comprehension questions pertaining to the four stories. Pairwise contrasts between the experimental conditions confirmed that, in line with Hypothesis 3b, audiotaped free-telling and audiotaped read-aloud did not differ (p = .643). Although there was trend in favor of *in vivo* free-telling, the effect postulated in Hypothesis 3a failed to reach significance. *In vivo* read-aloud (p = .099, d = .37) or both audiotaped conditions (free-telling: p = .239, d = .20; read-aloud: p = .388, d = .08; one-tailed contrasts) did not significantly differ from *in vivo* free-telling.

As we were also interested in the influence of general vocabulary knowledge on story comprehension and potential interactions with story presentation, we calculated a second model. In a first step we added the PPVT score to Model 3, which led to a significant model improvement, $\chi^2(1) = 47.03$, $p < .001$. The additional inclusion of the interaction between experimental condition and PPVT score did not lead to any further improvement, $\chi^2(3) = 2.42$, $p = .490$, indicating that story presentation had no differential effects as a function of general vocabulary knowledge. Analysis of the random structure indicated that a model including a random intercept for subject and item as well as a random item slope for general vocabulary knowledge were appropriate.

As shown in Model 4 (see Table 3), general vocabulary was positively related to story comprehension (*OR* = 3.43) and its inclusion reduced the variance between the children and the standard errors in the model. In addition, controlling for differences in general vocabulary knowledge, resulted also in significantly higher story comprehension scores in the *in vivo* free-telling condition (*OR* = 1.60). Pairwise one-tailed contrasts indicated that now and in line with Hypothesis 3a, *in vivo* free-telling was better than *in vivo* read-aloud, *Estimate* = 0.78, *SE* = 0.35, p = .013, d = 0.43, and audiotaped read-aloud, *Estimate* = 0.70, *SE* = 0.35, p = .024, d = 0.38, but not than audiotaped free-telling, *Estimate* = 0.41, *SE* = 0.36, p = .126, d = 0.23 (see Figure 2). Again, supporting Hypothesis 3b, audiotaped read-aloud and audiotaped free-telling did not differ (p = .399), excluding the possibility that general differences in language complexity may have acted as the underlying mechanism. Furthermore, in vivo read-aloud did also not differ from both audiotaped versions (p = .797 and p = .270), indicating that the mere presence of a narrator had no impact.

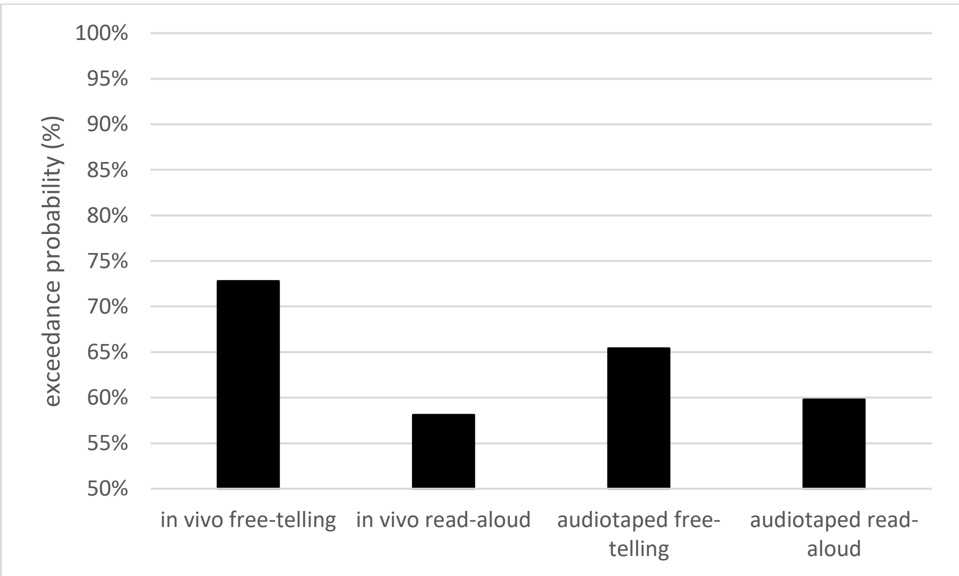

Figure 2. Story comprehension as a function of story presentation controlling for narrator influences and children's general vocabulary (Model 4).

Discussion

The current study investigated differential effects of two story-delivery methods, namely free-telling and read-aloud, on novel word learning, story comprehension, and children's on-task behavior during listening to stories, while trying to disentangle effects of language complexity and narrator's attention-guiding behavior. To do so, alongside *in vivo* presentations of read-aloud and free-telling audiotaped versions of both story-delivery conditions were included as experimental conditions. In line with our hypotheses, we found that free-telling conducted by a narrator *in vivo* resulted in the largest target-word gains and – after controlling for the effect of general vocabulary knowledge – also led to better story comprehension.

Importantly, the beneficial effect of *in vivo* free-telling cannot be attributed to a simple effect of narrator presence, because *in vivo* read-aloud was not better than its audiotaped version. Moreover, as audiotaped free-telling did not result in higher gains than audiotaped read-aloud, we can also exclude the possibility that it was simply due to the lesser language complexity that is typical for spoken language (Montag et al., 2015; Nation, 2006). This is also supported by our check of implementation fidelity showing that *in vivo* free-telling and audiotaped free-telling were largely similar, such that the audiotaped version represented an appropriate language control condition for its respective *in vivo* condition.

Moreover, we could also replicate Suggate et al.'s (2013) finding that *in vivo* free-telling of stories took more time than *in vivo* read-aloud, with free-telling showing considerably more variability in duration than read-aloud. Also in line with Suggate et al., we found that story duration was not significantly related to target-word learning gains in any of the conditions.

Finally, we could also demonstrate that, in line with Hypothesis 1, narrator and child behavior during *in vivo* storytelling was affected by story-delivery method, with narrators showing more gesticulation and eye contact and children being less restless and more attentive during free-telling. Conceptually, it could be these features of free-telling making it inherently more interesting and engaging for preschool children, which leads in turn to greater word learning. To test this idea, we conducted post-hoc analyses: Target-word learning gains were significantly related to narrators' using greater amounts of gesticulation (r = .45) and eye contact, (r = .43) as well as on a descriptive level to children being more attentive (r = .31) and less restless (r = .20). Findings indicated almost the same pattern for story comprehension. Narrator's gesticulation (r = .39), children's attentiveness (r = .48) and restlessness (r = .48) were significantly correlated with comprehension scores, and on a descriptive level there was also a relation to eye contact (r = .31).

Given that narrators were not instructed to use more attention-guiding behavior and were blind to the hypotheses of the study, it does not appear that differences in narrator or child behavior were an artefact of our instructions but instead arose out of natural interactions appropriate to free-telling and sharing books. Consequently, our study lends support to results of earlier, less controlled studies finding advantages for free-telling on children's behavior and engagement (e.g., Myers, 1990; Trostle & Hicks, 1998). Moreover, our study complements studies conducted on older children (Suggate et al., 2013) as well as studies, in which read-aloud and free-telling conditions may – due to strict experimental control – simply not have differed enough in terms of narrator behavior and consequently also in learning gains (Lenhart et al., 2018; Vaahtoranta et al., 2018). These findings thus extend previous work (e.g., Myers, 1990; Trostle &

Hicks, 1998) by showing that once differences between the free-telling and shared-reading conditions were implemented in a rigorous experimental design significant differences for word-learning and story comprehension were demonstrated. It might be that in a more naturalistic setting, in which narrators were allowed to use comments and questions to stimulate narrator-child interaction, for example as in Myers' (1990) study, differences in favor of the *in vivo* free-telling conditions would have been even larger.

Limitations and Directions for Future Research

Despite the clear experimental design, the study has limitations. The first is that we measured only recognition of target words, but we did not explore effects of story-delivery method on deeper word processing (Hoffman, Teale, & Paciga, 2014). Secondly, the current study did not include videotape control conditions. Instead, we used audiotaped control conditions to control for differences in language complexity between read-aloud and free-telling approaches. Although including the audiotaped conditions allowed us to directly test our research question by separating language exposure from narrators' attention-guiding behavior (i.e., eye contact and gesture usage), our results do not permit to separate effects of mere visual input from effects of narrator's reactive use of gestures as a way to regain a child's attention. To do so, future research should also include video control condition of free-telling and read-aloud.

Although the overall sample size was similar to, or even larger than in, other studies targeting differences between free-telling and read-aloud (e.g., Isbell et al., 2004; Suggate et al., 2013; Trostle & Hicks, 1998; Vaahtoranta et al., 2018), a third limitation is that our cell sizes with around 15 children per condition were somewhat small. This is due to the operationalization of story presentation as a between-subjects factor, which has the main advantage that children are not required to participate in all conditions and are therefore not over-challenged by listening to large number of different stories. In our study, children heard four stories twice, whereas in a pure within-subjects design they would have to listen to 16 stories twice. As studies favoring free-telling found large to medium sized effects (Trostle & Hicks, 1998: story comprehension: $f = .45$; vocabulary: $f = .48$; Suggate et al., 2013: vocabulary $f = .23$ in second grade), a power analysis with G*Power ($\alpha = .05$, $1\text{-}\beta = .80$; version 3.1.9.2; Faul, Erdfelder, Lang, & Buchner, 2007) indicated that a sample size of 60 children was sufficient to detect differences in target-word learning gains and story comprehension. However, the smaller effect sizes found in our study as well as the large standard errors observed for story comprehension suggest to use larger cell sizes in future studies.

Finally, our study examined only short-term effects of a very brief experimental intervention. Therefore, our conclusions are somewhat limited and need replication and extension to more intensive intervention studies. Particularly, in the long run written material may be preferred for a number of reasons. First, books may be more suitable to provide diverse, high-quality language input for kindergarten teachers and parents with limited time at their hands. Written contexts usually provide a richer vocabulary than spoken language (Nation, 2006) and children's books contain many unique words which are not present in normal parent-child communication (Montag et al., 2015). Second, in sharp contrast to read-aloud, free-telling may depend heavily on the language skills (e.g., vocabulary, grammar) of the narrator. Due to the relatively high and homogenous academic education in many reaches of the general population, this might be a lesser problem for interventions conducted by professionals, such as trained researchers (as in our current study) or preschool teachers. In contrast, for interventions targeting language in the home environment, parents' language levels may favor adhering to shared-reading over free-telling practices in many instances. Thus, given that parents of low socioeconomic status typically provide less language input in terms of quantity and diversity (Hart & Risley, 1995), this might be especially problematic for children for whom interventions are needed most (Biemiller, 2003). Consequently, in the long run, written material may be more suitable to provide diverse, high-quality language input in those cases. However, oral communication is a more naturalistic setting and parents, who might have reading and spelling problems themselves, probably find it easier to communicate with their child directly. In any case, helping parents to develop communicative competencies and bonding with their children, either through joint book reading or free-telling of stories, is always a good choice (Reese, 2013).

Future research needs to tackle these limitations by including a wider array of measures to capture the complexity of novel word acquisition and by using larger sample sizes as well as longer-term as well as more naturalistic intervention studies, in which the effects of both story-delivery methods can be compared in the long run. Particularly, longer termed studies may better portray a possible trade-off between losses in language diversity and the opportunity to enhance children's on-task behavior through better narrator engagement.

Conclusion

Taken together, it seems that free-telling of stories may be an equally effective – or sometimes even more effective – method to foster children's vocabulary, story comprehension, and participation in stories. Hence, free-telling should be added to the expanding repertoire of methods available to address the difficult task of fostering children's language development. However, since the state of the research is based on short-term experimental findings, and due to the practical considerations outlined in the previous section, future long-term studies on different methods of story-delivery are clearly needed.

References

Akhtar, N. (2005). The robustness of learning through overhearing. *Developmental science*, *8*, 199–209. https://doi.org/10.1111/j.1467-7687.2005.00406.x

Akinnaso, F. N. (1982). On the differences between spoken and written language. *Language and Speech*, *25*, 97–125. https://doi.org/10.1177/002383098202500201

Baayen, H., Bates, D., Kliegl, R., & Vasishth, S. (2015). RePsychLing: Data sets from psychology and linguistics experiments. R package version 0.0.4. Retrieved from https://github.com/dmbates/RePsychLing

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders*, *36*, 189–208. https://doi.org/10.1016/S0021-9924(03)00019-4

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4:328*. https://doi.org/10.3389/fpsyg.2013.00328

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. Retrieved from https://arxiv.org/abs/1506.04967v1

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Biemiller, A. (2003). Vocabulary: Needed if more children are to read well. *Reading Psychology*, *24*, 323–335. https://doi.org/10.1080/02702710390227297

Biemiller, A., & Boote, C. (2006). An effective method for building meaning vocabulary in primary grades. *Journal of Educational Psychology*, *98*, 44–62. https://doi.org/10.1037/0022-0663.98.1.44

Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of Educational Psychology*, *93*, 498–520. https://doi.org/10.1037/0022-0663.93.3.498

Björnsson, C.-H. (1968). *Lesbarkeit durch LIX*. (Tech. Rep. No. 6) [Readability through LIX]. Stockholm, Sweden: Pedagogiskt Centrum.

Blewitt, P., Rump, K. M., Shealy, S. E., & Cook, S. A. (2009). Shared book reading: When and how questions affect young children's word learning. *Journal of Educational Psychology*, *101*, 294–304. https://doi.org/10.1037/a0013844

Christensen, R. H. B. (2018). Ordinal - regression models for ordinal data. R package version 2018.8-25. Retrieved from https://www.cran.r-project.org/package=ordinal/.

Christian, K., Morrison, F. J., Frazier, J. A., & Massetti, G. (2000). Specificity in the nature and timing of cognitive growth in kindergarten and first grade. *Journal of Cognition and Development*, *1*, 429–448. https://doi.org/10.1207/S15327647JCD0104_04

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Connor, C. M., Piasta, S. B., Fishman, B., Glasney, S., Schatschneider, C., Crowe, E.,. . . Morrison, F. J. (2009). Individualizing student instruction precisely: Effects of child x instruction interactions on first graders' literacy development. *Child Development*, *80,* 77–100. https://doi.org/10.1111/j.1467-8624.2008.01247.x

Coyne, M. D., McCoach, D. B., & Kapp, S. (2007). Vocabulary intervention for kindergarten students: Comparing extended instruction to embedded instruction and incidental exposure. *Learning Disability Quarterly*, *30,* 74–88. https://doi.org/10.2307/30035543

Cromley, J. G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology*, *99*, 311–325. https://doi.org/10.1037/0022-0663.99.2.311

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology*, *33,* 934–945. https://doi.org/10.1037/0012-1649.33.6.934

Dickinson, D. K., & Smith, M. W. (1994). Long-term effects of preschool teachers' book readings on low-income children's vocabulary and story comprehension. *Reading Research Quarterly*, *29*, 104–122. https://doi.org/10.2307/747807

Dollinger, S. J., Matyja, A. M., & Huber, J. L. (2008). Which factors best account for academic success: Those which college students can control or those they cannot? *Journal of Research in Personality*, *42,* 872–885. https://doi.org/10.1016/j.jrp.2007.11.007

Elley, W. B. (1989). Vocabulary acquisition from listening to stories. *Reading Research Quarterly*, *24*, 174–187. https://doi.org/10.2307/747863

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*, 175–191. https://doi.org/10.3758/BF03193146

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science*, *16*, 234–248. https://doi.org/10.1111/desc.12019

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*, 221–233.

Gampe, A., Liebal, K., & Tomasello, M. (2012). Eighteen-month-olds learn novel words through overhearing. *First Language*, *32*, 385–397. https://doi.org/10.1177/0142723711433584

Grimm, H. (2010). *Sprachentwicklungstest für drei- bis fünfjährige Kinder (SETK 3–5)* [Speech development test for three- to five-year-old children] (2nd ed.). Göttingen, Germany: Hogrefe.

Haden, C. A., Reese, E., & Fivush, R. (1996). Mothers' extratextual comments during storybook reading: Stylistic differences over time and across texts. *Discourse Processes, 21,* 135–169. https://doi.org/10.1080/01638539609544953

Hargrave, A. C., & Sénéchal, M. (2000). A book reading intervention with preschool children who have limited vocabularies: The benefits of regular reading and dialogic reading. *Early Childhood Research Quarterly*, *15*, 75–90. https://doi.org/10.1016/S0885-2006(99)00038-1

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Paul H. Brookes.

Hayes, D. P. (1988). Speaking and writing: Distinct patterns of word choice. *Journal of memory and language*, *27,* 572–585. https://doi.org/10.1016/0749-596X(88)90027-7

Hayes, D. P., & Ahrens, M. G. (1988). Vocabulary simplification for children: A special case of 'motherese'? *Journal of Child Language, 15,* 395–410. https://doi.org/10.1017/S0305000900012411

Hoffman, J. L., Teale, W. H., & Paciga, K. A. (2014). Assessing vocabulary learning in early childhood. *Journal of Early Childhood Literacy*, *14*, 459–481. https://doi.org/10.1177/1468798413501184

Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods*, *39*, 101–117. https://doi.org/10.3758/BF03192848

Isbell, R., Sobol, J., Lindauer, L., & Lowrance, A. (2004). The effects of storytelling and story reading on the oral language complexity and story comprehension of young children. *Early Childhood Education Journal*, *32*, 157–163. https://doi.org/10.1023/B:ECEJ.0000048967.94189.a3

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Lenhard, A., Lenhard, W., Segerer, R., & Suggate, S. (2015). *PPVT-4: Peabody Picture Vocabulary Test - 4 (German Adaption)*. Frankfurt am Main, Germany: Pearson Assessment.

Lenhard, W., Baier, H., Endlich, D., Schneider, W., & Hoffmann, J. (2013). Rethinking strategy instruction: Direct reading strategy instruction versus computer-based guided practice. *Journal of Research in Reading*, *36*, 223–240. https://doi.org/10.1111/j.1467-9817.2011.01505.x

Lenhart, J., Lenhard, W., Vaahtoranta, E., & Suggate, S. (2018). Incidental vocabulary acquisition from listening to stories: A comparison between read-aloud and free storytelling approaches. *Educational Psychology*, *38*, 596–616. https://doi.org/10.1080/01443410.2017.1363377

Lenth, R. V. (2019). Emmeans: estimated marginal means, aka least-squares means. R package version 1.3.3. Retrieved from https://CRAN.R-project.org/package=emmeans

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning*, *65*, 185–207. https://doi.org/10.1111/lang.12117

Marulis, L. M., & Neuman, S. B. (2010). The effects of vocabulary intervention on young children's word learning: A meta-analysis. *Review of Educational Research, 80,* 300–335. https://doi.org/10.3102/0034654310377087

Marulis, L. M., & Neuman, S. B. (2013). How vocabulary interventions affect young children at risk: A meta-analytic review. *Journal of Research on Educational Effectiveness*, *6*, 223–262. https://doi.org/10.1080/19345747.2012.755591

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Mol, S. E., & Bus, A. G. (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin*, *137*, 267–296. https://doi.org/10.1037/a0021890

Mol, S. E., Bus, A. G., & de Jong, M. T. (2009). Interactive book reading in early education: A tool to stimulate print knowledge as well as oral language. *Review of Educational Research*, *79,* 979–1007. https://doi.org/10.3102/0034654309332561

Mol, S. E., Bus, A. G., de Jong, M. T., & Smeets, D. J. H. (2008). Added value of dialogic parent–child book readings: A meta-analysis. *Early Education and Development*, *19,* 7–26. https://doi.org/10.1080/10409280701838603

Montag, J. L., Jones, M. N., & Smith, L. B. (2015). The words children hear: Picture books and the statistics for language learning. *Psychological Science*, *26*, 1489–1496. https://doi.org/10.1177/0956797615594361

Myers, P. (1990). Stories from print. *Language Arts*, *67*, 824–831.

Nation, I. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*, 59–82. https://doi.org/10.3138/cmlr.63.1.59

Ninio, A., & Bruner, J. (1978). The achievement and antecedents of labelling. *Journal of Child Language*, *5,* 1–15. https://doi.org/10.1017/S0305000900001896

O'Donnell, R. C. (1974). Syntactic differences between speech and writing. *American Speech*, *49,* 102–110. https://doi.org/10.2307/3087922

Purpura, D. J., Hume, L. E., Sims, D. M., & Lonigan, C. J. (2011). Early literacy and early numeracy: The value of including early literacy skills in the prediction of numeracy development. *Journal of Experimental Child Psychology*, *110,* 647–658. https://doi.org/10.1016/j.jecp.2011.07.004

Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language, 59,* 413–425. https://doi.org/10.1016/j.jml.2008.02.002

R Core Team. (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Reese, E. (2013). *Tell me a story: Sharing stories to enrich your child's world*. New York, NY: Oxford University Press.

Reese, E., & Cox, A. (1999). Quality of adult book reading affects children's emergent literacy. *Developmental Psychology*, *35,* 20–28. https://doi.org/10.1037/0012-1649.35.1.20

Robbins, C., & Ehri, L. C. (1994). Reading storybooks to kindergartners helps them learn new vocabulary words. *Journal of Educational Psychology*, *86*, 54–64. https://doi.org/10.1037/0022-0663.86.1.54

Roney, R. C. (1996). Storytelling in the classroom: Some theoretical thoughts. *Storytelling World*, *9*, 7–9.

Schroeder, S., Würzner, K.-M., Heister, J., Geyken, A., & Kliegl, R. (2015). childLex: A lexical database of German read by children. *Behavior Research Methods, 47*, 1085–1094. https://doi.org/10.3758/s13428-014-0528-1

Sénéchal, M., Ouellette, G., & Rodney, D. (2006). The misunderstood giant: On the predictive role of early vocabulary to future reading. In D. K. Dickinson & S. B. Neuman (Eds.), *Handbook of early literacy research* (Vol. 2, pp. 173–182). New York, NY: The Guilford Press.

Silverman, R. (2007). A comparison of three methods of vocabulary instruction during read-alouds in kindergarten. *The Elementary School Journal*, *108*, 97–113. https://doi.org/10.1086/525549

Stahl, S. A., & Fairbanks, M. M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research*, *56*, 72–110. https://doi.org/10.3102/00346543056001072

Statistisches Bundesamt. (2017). *Bildungsstand der Bevölkerung 2016* [Educational level of the population 2016]. Retrieved from https://www.destatis.de/DE/Publikationen/Thematisch/BildungForschungKultur/Bildungsstand/BildungsstandBevoelkerung5210002167004.pdf?__blob=publicationFile. Accessed 16 January 2018.

Storch, S. A., & Whitehurst, G. J. (2002). Oral language and code-related precursors to reading: Evidence from a longitudinal structural model. *Developmental Psychology, 38*, 934–947. https://doi.org/10.1037/0012-1649.38.6.934

Suggate, S. P., Lenhard, W., Neudecker, E., & Schneider, W. (2013). Incidental vocabulary acquisition from stories: Second and fourth graders learn more from listening than reading. *First Language*, *33*, 551–571. https://doi.org/10.1177/0142723713503144

Trostle, S., & Hicks, S. J. (1998). The effects of storytelling versus story reading on comprehension and vocabulary knowledge of British primary school children. *Reading Improvement*, *35,* 127–136.

Vaahtoranta, E., Suggate, S., Jachmann, C., Lenhart, J., & Lenhard, W. (2018). Can explaining less be more? Enhancing vocabulary through explicit versus elaborative storytelling. *First Language*, *38,* 198–217. https://doi.org/10.1177/0142723717737452

Wasik, B. A., Hindman, A. H., & Snell, E. K. (2016). Book reading and vocabulary development: A systematic review. *Early Childhood Research Quarterly*, *37,* 39–57. https://doi.org/10.1016/j.ecresq.2016.04.003

Whitehurst, G. J., Arnold, D. S., Epstein, J. N., Angell, A. L., Smith, M., & Fischel, J. M. (1994). A picture book reading intervention in day care and home for children from low-income families. *Developmental Psychology, 30*, 679–689. https://doi.org/10.1037/0012-1649.30.5.679

Wild, J., & Pissarek, M. (2016). RATTE: Regensburger Analysetool für Texte. Dokumentation [RATTE: Ratisbon tool for text analysis. Documentation]. Retrieved from http://www.uni-regensburg.de/sprache-literatur-kultur/germanistik-did/medien/ratte_dokumentation.pdf. Accessed 16 January 2018.

Woolbert, C. H. (1922). Speaking and writing—A study of differences. *Quarterly Journal of Speech*, *8*, 271–285. https://doi.org/10.1080/00335632209379390

6 Study 3: The Effects of Questions during Shared-Reading: Do Demand-Level and Placement Really Matter?

A version of this chapter was published in:

Lenhart, J., Lenhard, W., Vaahtoranta, E., & Suggate, S. (2019). The effects of questions during shared-reading: Do demand-level and placement really matter? *Early Childhood Research Quarterly, 47*, 49–61. doi:10.1016/j.ecresq.2018.10.006

Abstract

Shared-reading fosters vocabulary development, although research has yielded mixed results regarding the effects of both demand-level (i.e., level of abstraction) and question placement on word learning. Different hypotheses drawing on broader theoretical frameworks have been proposed to explain individual findings. To test predictions made by these hypotheses, we read short stories to a sample of four- to six-year-old children (N = 86) in one-to-one reading sessions. We conducted a 2 x 3 mixed experiment with question placement (within the story vs. after the story) as within-subjects and demand-level (low vs. high vs. scaffolding-like by increasing from low to high) as a between-subjects factor. As additional controls, we utilized: (a) a control group in a just-reading condition without questions, and (b) control-words that were never accompanied by questions. Measures included receptive and expressive target- and control-vocabulary at the pre-and post-test along with general vocabulary and phonological working memory. Results indicate that question conditions were associated with higher gains for target-words at immediate and delayed post-test, but not for control-words. Contrary to proposed hypotheses, question placement or demand-level did not exert significant effects and they did not interact with language skills. However, children with greater general vocabulary showed most learning gains across conditions.

Keywords: questioning style; shared-reading; word learning; read-aloud; vocabulary acquisition; vocabulary

Introduction

Vocabulary knowledge plays a pivotal role in human cognitive development, is the central indicator for crystalized intelligence (Kail & Pellegrino, 1985), and an important predictor for text comprehension and academic success (Biemiller, 2006; Cromley & Azevedo, 2007; Sénéchal, Ouellette, & Rodney, 2006). Unfortunately, meaningful differences in vocabulary development emerge within the first years of life (Hart & Risley, 1995, 2003) and often persist despite formal schooling (Biemiller & Slonim, 2001; Christian, Morrison, Frazier, & Massetti, 2000), with severe academic disadvantages for those children with a restricted vocabulary (Walker, Greenwood, Hart, & Carta, 1994). Consequently, effective and timely interventions are needed to target early vocabulary gaps (Biemiller, 2003, 2006).

Over the last twenty years, shared-reading intervention programs have become popular to foster young children's vocabulary development (e.g., Dickinson & Smith, 1994; Hargrave & Sénéchal, 2000; Whitehurst, Arnold, Epstein, Angell, Smith, & Fischel, 1994). According to recent meta-analyses, shared-reading interventions seem to be a fairly effective means of increasing children's vocabulary (d = 0.54 to 0.88; see Marulis & Neuman, 2010, 2013; Mol, Bus, & de Jong, 2009; Mol, Bus, de Jong, & Smeets, 2008; Stahl & Fairbanks, 1986). Although the use of questions during shared-reading is regarded as a key factor (Marulis & Neuman, 2010, 2013; Mol et al., 2008), findings are unclear regarding the role of question demand-level (i.e., their level of abstraction) and placement, and also their interaction with children's language skills. How and when should questions be asked to maximize learning gains? Research in this field typically draws on conflicting hypotheses from broader frameworks such as "aptitude-treatment-interaction" or "scaffolding" (i.e., gradually reducing assistance, while the learning becomes more and more competent and autonomous). Consequently, depending on the theoretical background, different predictions regarding the effects of question placement and demand-level are made (e.g., Blewitt, Rump, Shealy, & Cook, 2009; Brabham & Lynch-Brown, 2002; Reese & Cox, 1999; Walsh, Sánchez, & Burnham, 2016). The current study aims at comparing and testing predictions regarding the role of question placement and demand-level, as well as to determine which parameters are best suited to foster children’s novel word learning from listening to stories.

Demand-Level of Questions

Regarding the demand-level of questions, a distinction is typically made between low- and high-demand questions (e.g., Blewitt et al., 2009; Reese & Cox, 1999; Walsh et al., 2016), which are sometimes also termed perceptual and conceptual questions (e.g., Justice, 2002). Whereas low-demand/perceptual questions are located on a low level of abstraction and focus on descriptions and repetitions of given information (e.g., “Whom is the dog barking at?” or “What is the dog doing?”), high-demand/conceptual questions target more abstract knowledge and typically require the child to make inferences, predictions, or evaluations (e.g., “Why is the dog barking so loudly?” or “Do you think that the dog will bite the postman?”) (Walsh et al., 2016; for a more fine-grained differentiation of demand-level, see van Kleeck, Gillam, Hamilton, & McGrath, 1997).

Three different hypotheses regarding the effect of the demand-level of questions on novel word acquisition during shared-reading have been proposed so far. The first hypothesis, proposed by Reese and Cox (1999), draws on the aptitude-treatment-interaction-framework (e.g., Connor, Morrison, & Katch, 2004; Connor et al., 2009) and predicts that children of higher performance levels would profit more from high-demand questions, and less-skilled children would advance more with low-demand questions. Regarding the aptitude-treatment-interaction hypothesis, studies yield mixed results. On the one hand, in Reese and Cox's (1999) study, in which four-year-old children heard approximately 30 stories in a one-on-one setting across the span of six weeks, demand-level interacted with children's vocabulary knowledge and question placement. Children with low vocabulary profited most from low-demand questions that were interspersed within the story, whereas children with better vocabulary gained most through high-demand questions placed after the story. On the other hand, no other studies tested for (Justice, 2002; Walsh et al., 2016) or found corresponding interactions between demand-level of questions and language skills, such as vocabulary knowledge (Blewitt et al., 2009).

The other hypotheses incorporate the distinction between fast and slow mapping processes in word learning (Carey, 1978). Fast mapping refers to a first step in word acquisition, by which some of its phonological, syntactic, or semantic features enter the mental lexicon. One or a few encounters with a word can suffice to acquire initial mental representations. Slow mapping, in contrast, takes more time and encounters, leading to deeper, more elaborated, and better structured word knowledge (Carey, 2010).

Hypotheses that are based on the distinction between fast and slow mapping processes argue that questions provide sufficient information for fast mapping irrespective of demand-level, but that the demand-level of questions has differential effects on the extension of partial knowledge and deeper encoding processes (Blewitt et al., 2009; Walsh et al., 2016). In line with the distinction between fast and slow mapping processes, the second hypothesis, proposed by Walsh et al. (2016), claims that after an initial encounter with novel words, high-demand questions, namely questions that are located on a higher level of abstraction and focus on conceptual aspects of a word, are especially beneficial as they allow children to create richer word meanings.

Regarding the high-demand hypothesis, findings are equally mixed. Although results of observational studies suggest that questions should be cognitively challenging (e.g., Dickinson & Smith, 1994; Haden, Reese, & Fivush, 1996), experiments conducted by Blewitt et al. (2009) and Justice (2002) reported no differences in receptive or expressive measures of target vocabulary between low- and high-demand question conditions. Additionally, Reese and Cox's (1999) data (with receptive vocabulary as dependent variable) generally favored low-demand questions, aside from the aforementioned interaction with children's language skills. Conversely, lending some support for the high-demand hypothesis, Walsh et al. (2016) found higher gains for the high-demand conditions regarding expressive but not receptive target vocabulary. However, these conditions resulted in no greater gains in vocabulary than a just-reading control group.

The third hypothesis, proposed by Blewitt et al. (2009), draws on a scaffolding framework (Wood, Bruner, & Ross, 1976) and argues that after an initial encounter with novel words a scaffolding-like approach, consisting of a gradual transition from low- to high-demand questions, might be best suited to bolster deeper word learning processes. Consistent with the scaffolding hypothesis, Blewitt et al. (2009; Study 2) reported an advantage of the scaffolding-like condition for deeper word processing (assessed via expressive word definitions), but not for fast mapping processes which they claimed to assess via a receptive measure of word recognition, that required choosing the right picture corresponding to a given target-word. However, more work is required to evaluate the scaffolding hypothesis, with Blewitt et al.'s (2009) study being the only one to have employed a scaffolding-like condition.

In sum, the picture portrayed by studies targeting the effect of demand-level of questions is inconclusive, favoring neither of the proposed hypotheses. Moreover, integrating the findings is difficult as the studies differed regarding important design features, such as the question conditions examined in the study, the number of books used, the number of reading sessions per book, or the mean age and age span of the participants. More work is needed from comprehensive studies testing the various positions.

Placement of Questions

In terms of question placement, two suggestions dominate the literature, namely recommendations to embed questions within stories, or place these prior to or following the stories, with placement after, not before, being the more common approach (Brabham & Lynch-Brown, 2002; Blewitt et al., 2009; Reese & Cox, 1999; Walsh et al., 2016).

Regarding a potential impact of question placement on novel word learning while listening to stories, at least three different outcomes are conceivable. Firstly, asking questions after the end of the story (i.e., the no-interruption hypothesis) might be better because the story flow is not disrupted and is consequently perceived as a whole (Dickinson & Smith, 1994; Brabham & Lynch-Brown, 2002). Secondly, interspersing questions within the stories (i.e., the interruption hypothesis) may instead offer the possibility to reflect on and clarify difficult aspects of the story as they arise, and thus enhance comprehension and learning of novel words (Brabham & Lynch-Brown, 2002).

A third hypothesis draws again on the aptitude-treatment-interaction framework, and argues that interspersing questions within the stories might be particularly helpful for lower performing children (Reese & Cox, 1999).

Unfortunately, research has also been inconclusive regarding the effects of question placement. Although an observational study in kindergarten classrooms conducted by Dickinson and Smith (1994) provided evidence that a non-disruptive reading style with questions woven around the stories seemed to best foster general vocabulary development, an experimental study with first and third graders came to an opposite conclusion (Brabham & Lynch-Brown, 2002). Additionally, two studies conducted with preschool children in one-on-one (Blewitt et al., 2009) or small group reading sessions (Walsh et al., 2016) found no significant differences between the reading styles, concluding that both types of placement were equally effective.

In a long-term study by Reese and Cox (1999), question placement interacted with children's vocabulary knowledge. Reese and Cox found that easy questions interspersed within the story produced the highest overall vocabulary gains, but children with better vocabulary profited most from difficult questions after the story. Both, however, were not significantly different from a condition in which difficult questions were interspersed within the story. As in Reese and Cox's (1999) study, the design was not fully factorial, namely because the condition with easy questions after the story was omitted, effects of placement and demand-level are difficult to separate. Yet, closely examining the descriptive results of all three conditions, demand-level and not question placement seemed to result in greater vocabulary gains. Again, due to the heterogeneity of design features, it is difficult to integrate findings across studies into a consistent picture, favoring any of the proposed hypotheses. Consequently, there is clear need for additional research.

Current Study

Despite shared-reading being used extensively to foster language development in young children and research showing the general effectiveness of book reading interventions, there is a lack of research concerning the effectiveness of specific features (Wasik, Hindman, & Snell, 2016). Particularly, more work is needed to gain insight into differential effects of demand-level and placement of questions on novel word acquisition during shared-reading.

Based on the proposed hypotheses and mixed results of previous studies on effects of question demand-level and placement, we designed an experiment to integrate the different approaches into one coherent experimental framework. We tried to extend existing studies by (a) rigorously controlling for interaction/feedback provided by the experimenters, (b) including a full-factorial design, (c) larger sample sizes per condition, (d) receptive and expressive measures for different levels of word processing, and (e) a follow-up test for retention of learning gains. The three research questions we tried to target with the current study were:

1) Do questions increase word learning from listening to stories?
2) Do question demand-level or its interaction with children's initial language skills influence word learning from listening to stories?
3) Do question placement or its interaction with children's initial language skills influence word learning from listening to stories?

First, in line with relevant meta-analyses (Marulis & Neuman, 2010, 2013; Mol et al., 2008) we assume that although some word learning will occur from book reading irrespective of the use of questions or not (Hypothesis 1a), gains are higher for words accompanied by questions (Hypothesis 1b). This should be evidenced via a between-subjects comparison between the question conditions, in which questions are asked, and the just-reading control group, in which no questions are included. Here, we expect to find a difference in favor of the question conditions.

Second, regarding demand-level of questions, we tested predictions made by the presented hypotheses. The aptitude-treatment-interaction hypothesis (Reese & Cox, 1999) predicts an interaction between children's initial language skills (i.e., general vocabulary) and demand-level. Higher performing children should profit from high-demand questions, whereas lower-performing children should advance more from low-demand questions (Hypothesis 2a). In contrast, the high-demand hypothesis (Walsh et al., 2016) predicts a general superiority of high-demand questions (Hypothesis 2b), and the scaffolding hypothesis (Blewitt et al., 2009) a superiority of a scaffolding-like procedure with question demand-level increasing from low- to high-demand over the course of the reading sessions (Hypothesis 2c). However, as both the high-demand and scaffolding hypotheses assume that children need to have acquired some knowledge of novel words before differential effects of question demand-level emerge, differences are only expected on measures that reflect deeper levels of word processing, such as on word definition tasks (Blewitt et al., 2009; Walsh et al., 2016).

Third, regarding question placement there seem to be at least three possibilities. The no-interruption hypothesis (Dickinson & Smith, 1994) predicts an overall benefit of non-interruptive question placement conditions (Hypothesis 3a). In contrast, the interruption hypothesis (Brabham & Lynch-Brown, 2002) assumes that interspersing stories within the stories helps to improve understanding and acquisition of novel words

(Hypothesis 3b). Finally, the aptitude-treatment-interaction hypothesis predicts an interaction between children's skill level (i.e., general vocabulary) and placement of questions. Higher performing children should thus profit most from questions after the stories and less-skilled children should advance most through questions within the stories (Hypothesis 3c).

Methods

Participants

Five kindergartens in a middle-sized city (approximately 130,000 inhabitants) in Germany were invited to distribute letters of participation to the parents of four- to six-year-old children. Participating parents were offered feedback regarding their children's language performance. In total, the parents of 94 children provided written consent for their child's participation. From this sample, six children were not included in the intervention (three had insufficient language abilities and did not understand the instruction of the tests, two switched kindergarten before the intervention phase began, one letter of consent arrived only after the start of the study) and two children had to be excluded from analyses (one was younger than four years, and one refused to continue after the language pre-tests). Of the remaining 86 children 52% were female and had a mean age of 61.60 months (SD = 8.06; Min = 48; Max = 77).

All children except one were born in Germany, although 14% had one parent and 6% both parents born in a country other than Germany, and 9% of the children spoke a second language at home. All children except one had at least one parent who left school with a formal educational qualification. Seventy-two % of the mothers and 78% of the fathers had received a high-school diploma or a university degree. Thus, the educational level of our sample was noticeably higher than that of the German population (Statistisches Bundesamt, 2017).

Design

The design of the study was 3 x 2, with demand-level (low, high, scaffolding-like) as a between-subjects factor, and question placement (within the stories, after the stories) as a within-subjects factor. In addition, two kinds of control, a between-subjects control as well as a within-subjects control, were included in the experiment. The within-subjects control was that each question condition comprised selected words that were accompanied by questions (target-words), but also a selection of words that were not accompanied (control-words). The between-subjects control was a just-reading control group, in which none of the target-words was accompanied by questions. A depiction of the study design can be found in Figure 1.

We included three dependent variables to assess different levels of target-word acquisition from listening to stories. Receptive target-word acquisition was measured with word recognition, expressive target-word acquisition with word definitions as well as with picture naming. Target- and control-word recognition and definitions were assessed at pre-test, immediate post-test, and delayed post-test. Picture naming of target- and control-words was measured only twice, namely at immediate and delayed post-test. In addition, to test for aptitude-treatment interactions, continuous measures of receptive and expressive vocabulary knowledge were included.

An effort was made to make the study single-blind, thus the preschool teachers, parents and children received no information as to the goals of the study. Due to the design, the experimenters, however, read the stories, conducted the pre- and post-tests, and were, consequently, not blind to the children's experimental condition assignment.

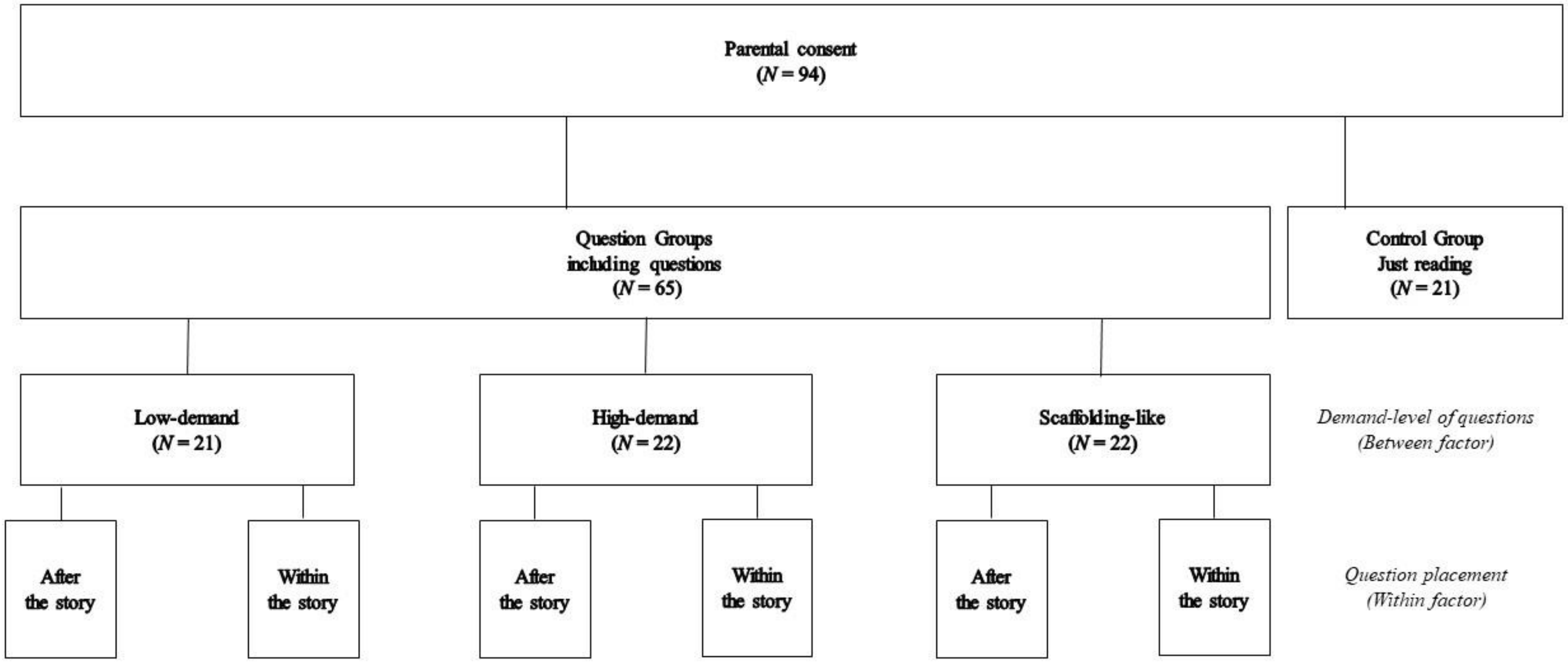

Figure 1. Overview of the study design.

Stimuli

Stories. Four fictional stories were chosen from different children's books that were written for the age span between four and six years. To reduce memory load and the length of the experimental sessions, we shortened the stories to around 850-1000 words per story. The readability index LIX (Björnsson, 1968) of the stories ranged between 23 and 39, denoting "very easy" to "easy" texts (using the calculation program of W. Lenhard & Lenhard, 2014). Each of the four stories contained six different target-words (four nouns and two verbs) and three different control-words (two nouns and one verb), resulting in 24 different target-words (16 nouns and eight verbs) and 12 different control-words (eight nouns and four verbs). In the question conditions, the target-words were accompanied by questions, whereas the control-words were presented without questions. In the just-reading control condition, none of the words was highlighted by questions. Target- and control-words appeared only once per story. Consistent with research examining word learning from stories and to avoid ceiling effects, target- and control-words were selected to have a low frequency of occurrence in everyday language as well as in children's books, as determined by the occurrence in lexical databases DWDS (Geyken, 2007; Heister et al., 2011) and ChildLex (Schroeder, Würzner, Heister, Geyken, & Kliegl, 2015). A list of target- and control-words with their respective lemma frequencies in both corpora is included in the appendix (see Appendix A1). Target-words and control-words did not differ significantly regarding lemma frequency (per million tokens/words), ChildLex: $M_{\text{control-words}} = 19.33$, $SD_{\text{control-words}} = 23.24$, $M_{\text{target-words}} = 9.79$, $SD_{\text{target-words}} = 12.38$, $t(34) = 1.62$, $p = .115$, DWDS: $M_{\text{control-words}} = 225.83$, $SD_{\text{control-words}} = 300.23$, $M_{\text{target-words}} = 123.42$, $SD_{\text{target-words}} = 121.21$, $t(12.825) = 1.14$, $p = .277$. Accordingly, both target- and control-words were expected to be largely, and to a similar extent across stories, unfamiliar to the participating children. Examples are *Buddel* (a rarely used word for bottle), *Eisbrecher* (icebreaker), or *sinnieren* (a rarely used word which means “to muse”). To increase ecological validity, five pictures were included per story, with three target-words and one control-word being depicted once.

Questions. In the question conditions, in each of the three reading sessions each of the six target-words, that were included per story, was accompanied by one question, either directly after the sentence that contained the target-word or at the end of the story. Questions differed according to demand-level conditions (low-demand, high-demand, and scaffolding-like). In line with previous studies (e.g., Blewitt et al., 2009; Justice, 2002; Walsh et al., 2016), low-demand questions focused on perceptual descriptions and repetitions of given information (e.g., “Where does the lion lurk?” or “What does the lion’s mane look like?”), whereas high-demand questions required the child to make inferences (e.g., “Why is the warthog afraid of the lion?” or “Why does the warthog prefer to bathe alone?”). The scaffolding-like condition was operationalized through gradually increasing demand-level of questions. To do so, in the first reading session all target-words were accompanied by low-demand questions, and in the second reading session half of the target-words were highlighted by low-demand, and the other half by high-demand questions. Finally, in the third reading session only high-demand questions were employed. The experimenters were instructed to encourage the children to answer the questions, but they did not provide any factual feedback or elaborate the children’s answers.

Measures

All materials, including the parental questionnaire and the different language tests were in German.

Demographics. Parents completed a questionnaire regarding their own and their children's country of birth, languages spoken at home and the parents' highest educational qualification. The response rate was 100%.

Vocabulary. To assess children's receptive vocabulary, a German adaption of the Peabody Picture Vocabulary Test 4 (PPVT-4; A. Lenhard, Lenhard, Segerer, & Suggate, 2015) was used. Children had to point the correct picture matching the spoken word. Split-half reliability (odd-even split of administered items) was .88. To assess children's expressive vocabulary, the subtest Expressive Vocabulary of the German version of the Wechsler Preschool and Primary Scale of Intelligence-III (WPPSI-III; Petermann, 2011) was employed. Cronbach's α was .71. As the tests were highly correlated (r = .67), to obtain a single score for general vocabulary the results of both tests were z-standardized and averaged.

Phonological working memory. Phonological working memory was included as a potential control variable, as it is an important prerequisite for the processing of novel sound patterns and language learning (Baddeley, 2003). Phonological working memory was measured by two different tests, namely the subtest Phonological Working Memory for Non-Words of the Speech Development Test for Three- to Five-Year-Old Children (SETK 3-5; Grimm, 2010), in which children are required to repeat non-words, as well as the subtest Digit Span of the German version of the Kaufman Assessment Battery for Children (K-ABC; Melchers & Preuß, 2009), in which digits have to be recounted in the right order. Cronbach's α was .69 for Phonological Working

Memory for Non-Words and .74 for Digit Span. Both tests were moderately correlated (r = .42).

Tests of target- and control-word acquisition. To assess different levels of target-word acquisition, tests of word recognition (receptive), picture naming (expressive), and word definition (expressive) were constructed. To check for implementation fidelity of the conditions, namely to rule out that inadvertent differences in quality of story reading or a better relationship between experimenter and child through higher degrees of interaction may affect the results, control-words were included alongside the target-words in each task. In each task a maximum score of 36 was possible, whereby 24 was the maximum for target-words and 12 for control-words.

Word recognition task. The word recognition task was used to assess shallower receptive target- and control-word acquisition and followed a format similar to the PPVT-4. Children had to indicate, from one of four pictures, the picture that best matched the spoken word. Someone looking thoughtfully into the air, for example, illustrated the verb *sinnieren* ("to muse"). Distractors were pictures in which a person was either singing, sleeping, or eating. Cronbach's α was .60 at pre-test, .79 at immediate post-test, and .81 at delayed post-test.

Picture naming task. The picture naming task assessed shallower expressive target- and control-word knowledge and required the children to label illustrations of the target- and control-words presented one at a time. As synonyms exist for each of our target- and control-words, the labelling task was only administered at immediate and delayed post-test. If a synonym of a word was given, the experimenter was instructed to ask again (e.g., "That is correct. But one can also use another word for that object. Do you still remember the word that I used in the story?"). Cronbach's α was .76 at immediate post-test, .78 at delayed post-test.

Word definition task. The word definition task was used to assess deeper expressive target- and control-word knowledge and required the children to define the target- and control-words. They were asked: "Can you explain, what XY means?" followed by another prompt: "Do you know anything else about a XY?" One point was awarded if a correct synonym or a complete explanation was provided. Partial explanations (e.g., naming superordinates, properties, or associated concepts) received 0.5 points. Incorrect answers or answers that relied solely on phonological similarity (e.g., *Pfuhl/puddle* was explained by the word *Pool/pool*) were awarded zero points. All answers were coded independently by two raters. Inter-rater reliability was assessed using a two-way mixed, average-measures intraclass correlation (ICC) (McGraw & Wong, 1996). ICC for consistency was .98 at the pre-test, .99 at the immediate post-test, and .99 at the delayed post-test, indicating excellent agreement (Cicchetti, 1994). Finally, the codings of the two experimenters were averaged, resulting in scores of .0, .25, .5, .75 or 1 per target-word definition, with a minimum of zero und a maximum of 36 for the total scale. Cronbach's α was .74 at the pre-test, .88 at immediate post-test, and .90 at delayed post-test.

Procedure

The study took place in separate rooms in the kindergartens and was conducted by trained student research assistants and the first author. All test and reading sessions were conducted individually and had a duration of approximately 30 minutes.

Reading sessions. In total, there were six reading sessions per child, each on different days across the course of four weeks (days between reading sessions: *m* =

4.67, *SD* = 3.35). Each story was read three times. In the first three reading sessions, an experimenter read two of the four stories, one after another, either with the questions interspersed within the stories or after the stories. In reading sessions 4 to 6, the other two stories were read in the respective other within-condition of question placement. To avoid order and story effects both the order of the stories as well as their assignment to the story conditions were randomized. Experimenters were instructed to encourage the children to answer the questions, but they did not provide any factual feedback or elaborate the children's answers. Figure 2 provides an overview of the procedure.

Pre- and post-tests of target- and control-words. The pre-test of the first two stories was conducted before the stories were read for the first time. The immediate post-test was administered immediately after the third reading of the stories. The same procedure was applied to the second set of stories. The delayed post-test was conducted approximately four weeks after the last reading session (*M* = 28.81 days, *SD* = 3.04). In order to minimize effects of earlier target-word tests on later ones, target-word tests were always administered in the following order: picture naming, word definitions, and word recognition.

Language tests. The majority of the covariates (PPVT-4, Expressive Vocabulary, Phonological Working Memory for Non-Words) was gathered in a separate session approximately four and a half weeks before the reading sessions started (*M* = 32.08 days, *SD* = 3.63). Digit Span was assessed during the second reading sessions of the second set of stories.

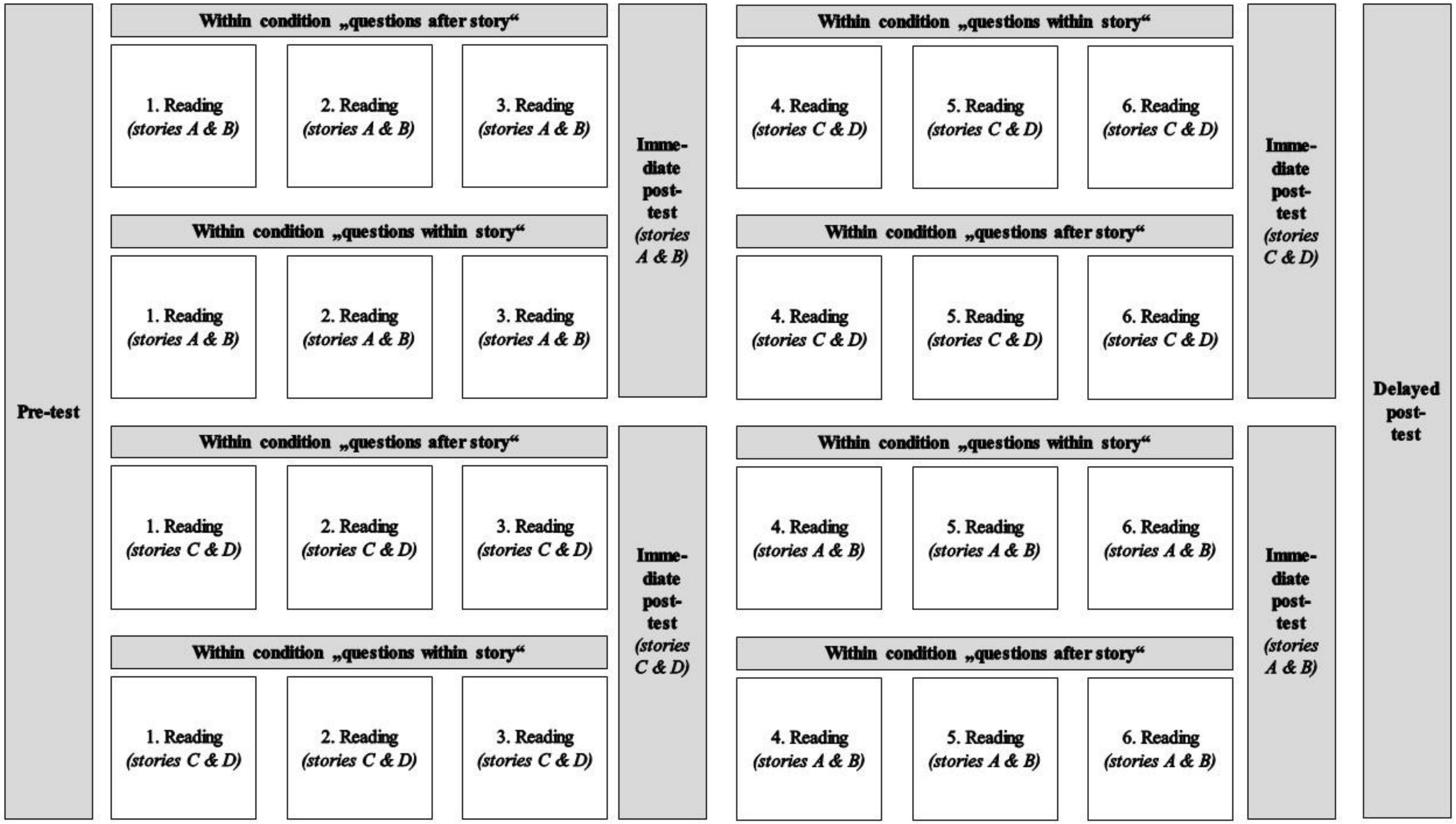

Figure 2. Overview of the experimental procedure.

Data Analysis

Analyses were conducted using mixed-effects models for experimental research designs (Hoffman & Rovine, 2007). To maintain better comparability with prior research in shared-reading mechanisms (e.g., Blewitt et al., 2009; Reese & Cox, 1999; Walsh et al., 2016), we modeled learning gains on the scale level.

At first, analyses were conducted with the package *lme4* (version 1.1-14; Bates, Mächler, Bolker, & Walker, 2015) of *R* (version 3.3.3; R Core Team, 2017). However, as the inspection of the residuals indicated violations of homoscedasticity and normality assumption, we used the package *robustlmm* (version 2.1-3; Koller, 2016), which provides robust estimates of linear mixed models. This is achieved by using lower weights for outlying values. As the estimating equations of the *robustlmm* package do not correspond to any likelihood or pseudo-likelihood, information criteria and tests based on the log-likelihood statistic are not available. In consequence, model comparisons, which are additionally recommended to assess the importance of specific predictors (Hoffman & Rovine, 2007), could not be employed. Additionally, as *robustlmm* does not provide *p*-values for predictors, $t > |2.0|$ is considered indicating a statistically significant effect (Gelman & Hill, 2007).

Our goal was confirmatory hypothesis testing. Consequently, we included in a first step all random slopes for fixed effects that varied within subjects, thereby specifying a maximal random structure (Barr, 2013; Barr, Levy, Scheepers, & Tily, 2013). However, due to the relative low number of data points on the scale level, we had to simplify the random structures of our models (Bates, Kliegl, Vasishth, & Baayen, 2015). For the comparisons between control and question groups as well as between question groups (i.e., demand-level), only random intercept for subjects were included.

For those within the question groups (i.e., question placement), we first modeled random intercepts as well as time and question placement as random slopes in our *lme4* models. When switching to robust estimation, perfect correlations between random intercept and slopes indicated overparametrization. We therefore only kept random intercepts in the random structure. As we were interested in learning differences between specific groups, we dummy-coded all variables (Cohen, Cohen, West, & Aiken, 2003). The estimates (unstandardized betas) therefore represent simple effects or simple interaction effects, which allow a direct comparison between groups and their learning gains. The *R* codes are included in the appendix (see Appendix B1, B2, B3).

Missing data. A small proportion of children were missing for the language tests (between 1-3%) or the respective tests of target- and control-word acquisition (between 2-5% of the children). An advantage of mixed models, however, is that they are very robust against small amounts of missing data (Quené & van den Bergh, 2008; Twisk, de Boer, de Vente, & Heymans, 2013). Consequently, no imputation method was used in the present study.

Results

To enhance readability, test-statistics are preferably displayed in tables. Table 1 provides an overview of the descriptive statistics of age, language covariates, and measures of target- and control-word acquisition in each between-subjects condition. The groups did not differ regarding age or any of the language covariates ($p > .05$) – so age and measures of phonological working memory were not included as control variables in the following analyses.

Table 1

Descriptive Statistics for the Just-Reading Control and the Question Groups

	Just-reading (*n* = 21)	**Low-demand (*n* = 21)**	**High-demand (*n* = 22)**	**Scaffolding-like (*n* = 22)**
	M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Age (months)	59.76 (8.53)	62.00 (7.11)	61.77 (8.01)	62.82 (8.73)
Language covariates				
Peabody Picture Vocabulary Test-4	104.71 (21.37)	109.95 (27.01)	111.00 (18.29)	114.27 (23.72)
Expressive Vocabulary (WPPSI-III)	18.95 (2.52)	20.05 (3.09)	19.86 (2.73)	19.77 (2.88)
Working Memory for Non-Words (SETK 3-5)	11.29 (3.23)	12.15 (3.08)	11.86 (2.83)	11.05 (3.37)
Digit Span (K-ABC)	7.90 (2.34)	8.25 (1.94)	8.50 (2.15)	7.95 (2.33)
Target-word tests				
Word recognition (pre-test)	10.16 (3.29)	11.35 (3.90)	10.01 (2.43)	10.24 (2.95)
Word recognition (immediate post-test)	13.20 (3.24)	15.00 (4.71)	15.09 (4.00)	14.29 (4.34)
Word recognition (delayed post-test)	12.95 (3.75)	14.80 (4.87)	14.62 (4.19)	13.95 (4.10)
Picture naming (immediate post-test)	3.10 (1.64)	5.75 (3.39)	5.59 (2.94)	5.48 (3.17)
Picture naming (delayed post-test)	2.81 (1.54)	5.50 (3.89)	4.48 (3.11)	4.71 (2.39)
Word definition (pre-test)	0.96 (1.10)	2.00 (1.96)	2.20 (1.63)	1.77 (1.60)
Word definition (immediate post-test)	3.46 (1.79)	6.59 (4.54)	7.32 (3.95)	5.50 (3.87)
Word definition (delayed post-test)	3.07 (2.04)	6.83 (5.01)	6.00 (3.76)	4.80 (3.78)
Control-word tests				
Word recognition (pre-test)	4.32 (1.63)	4.25 (1.52)	4.55 (1.60)	4.48 (1.72)

Word recognition (immediate post-test)	4.80 (1.64)	5.50 (2.67)	5.55 (1.82)	5.71 (1.76)
Word recognition (delayed post-test)	4.70 (1.59)	5.35 (2.25)	5.38 (2.20)	4.81 (2.09)
Picture naming (immediate post-test)	0.76 (0.83)	0.85 (0.75)	0.77 (0.61)	1.19 (1.40)
Picture naming (delayed post-test)	0.71 (0.72)	1.00 (1.17)	0.76 (0.62)	0.81 (1.12)
Word definition (pre-test)	0.19 (0.32)	0.89 (1.06)	0.59 (0.66)	0.99 (1.27)
Word definition (immediate post-test)	0.74 (0.71)	1.98 (1.75)	1.94 (1.47)	1.75 (1.94)
Word definition (delayed post-test)	0.82 (0.73)	2.13 (2.23)	1.61 (l.39)	1.64 (1.61)

Note. For language covariates, raw scores are reported. Target-word tests: maximum score = 24. Control-word tests: maximum score = 12. WPPSI-III = Wechsler Preschool and Primary Scale of Intelligence-III; SETK 3-5 = Speech Development Test for Three- to Five-Year-Old Children; K-ABC = Kaufman Assessment Battery for Children.

Effects of Questions: Comparing the Just-Reading Control Group and the Question Groups

To examine the general effect of story readings irrespective of the use of questions (Hypothesis 1a) and the beneficial effect of adding questions (Hypothesis 1b), we compared the just-reading control group and the question groups regarding target-word learning. To control for integrity of our conditions, we also compared them regarding control-word acquisition. We included time and between-group as fixed effects and a random intercept for subject. The final models for the target-words are displayed in Table 2, those for the control-words in Table 3.

Regarding the target-word tests (see Table 2), the final models showed that in the control group significant learning of target-vocabulary did occur from pre-test to immediate post-test and to delayed post-test. This is in accord with Hypothesis 1a, which assumed that word learning form listening to stories can occur irrespective of their presentation.

In line with our assumption that questions enhance learning outcomes (Hypothesis 1b), the question groups yielded higher gains at the immediate and delayed post-test regarding picture naming and word definitions of target-words (picture naming at the delayed post-test: *Estimate* = 1.85, *SE* = 0.67, t = 2.76). The target-word recognition measure displayed only a descriptive, marginally significant advantage in favor of the question groups at both measurement points.

Table 2

The Effect of Questions on Acquisition of Target-Words

	Receptive vocabulary (word recognition)[a]			Expressive vocabulary (word definitions)[a]			Expressive vocabulary (picture naming)[b]		
	Estimate (95%-CI)	*SE*	*t*	*Estimate (95%-CI)*	*SE*	*t*	*Estimate (95%-CI)*	*SE*	*t*
Fixed Parts									
(Intercept)	9.94 (8.13–11.74)	0.92	10.79*	0.96 (-0.45–2.37)	0.72	1.34	3.00 (1.83–4.11)	0.58	5.10*
Time (immediate post-test)[c]	2.79 (1.67–3.91)	0.57	4.88*	2.34 (1.09–3.59)	0.64	3.67*			
Time (delayed post-test)[c]	2.62 (1.48–3.76)	0.58	4.49*	1.89 (0.64–3.14)	0.64	2.96*	-0.25 (-1.05–0.56)	0.41	-0.60
Group (question conditions)[d]	0.96 (-1.10–3.03)	1.05	0.91	0.82 (-0.81–2.44)	0.83	0.99	2.37 (1.06–3.66)	0.67	3.54*
Time (immediate post-test) X Group	1.18 (-0.10–2.46)	0.65	1.80	2.13 (0.70–3.57)	0.73	2.91*			
Time (delayed post-test) X Group	1.11 (-0.19–2.41)	0.66	1.68	1.94 (0.50–3.37)	0.73	2.64*	-0.52 (-1.45–0.40)	0.47	-1.11

Random Parts			
Intercept variance	12.868	6.010	4.839
Residual variance	2.962	3.861	1.610
Observations	245	247	165
$N_{subjects}$	83	83	83

Note. The final models were fitted with the *rlmer*-function of the *robustlmm* package. Each model comprised a random intercept for subject as well as time and group as fixed effects. The factor Time was dummy-coded (word recognition and word definition: pre-test = 0; picture naming: immediate post-test = 0), as was the factor Group (just reading control group = 0).

[a] The intercept represents the score of the control group at the pre-test. The effect of time represents the change in the control group from pre-test to immediate or delayed post-test. [b] As no pre-test was conducted for the picture naming task, the intercept represents the score of the control group at the immediate post-test. The effect of time represents the change in the control group from immediate to delayed post-test. [c] reference group = pre-test (word recognition and word definitions) or immediate post-test (picture naming). [d] reference group = just reading control group.

* $|t| > 2.0$ is considered indicating a statistically significant effect (Gelman & Hill, 2007).

Table 3

The Effect of Questions on Acquisition of Control-words

	Receptive vocabulary (word recognition)[a]			Expressive vocabulary (word definitions)[a]			Expressive vocabulary (picture naming)[b]		
	Estimate (95%-CI)	*SE*	*t*	*Estimate (95%-CI)*	*SE*	*t*	*Estimate (95%-CI)*	*SE*	*t*
Fixed Parts									
(Intercept)	4.24 (3.35–5.13)	0.46	9.31*	0.19 (-0.33–0.71)	0.27	0.70	0.68 (0.31–1.04)	0.19	3.63*
Time (immediate post-test)	0.50 (-0.39–1.40)	0.46	1.10	0.53 (0.05–1.00)	0.24	2.17*			
Time (delayed post-test)	0.17 (-0.73–1.07)	0.46	0.37	0.59 (0.11–1.06)	0.24	2.43*	-0.03 (-0.34–0.28)	0.16	-0.20
Group (question conditions)	0.12 (-0.90–1.14)	0.52	0.24	0.50 (-0.10–1.10)	0.30	1.64	0.14 (-0.28–0.56)	0.21	0.66
Time (immediate post-test) X Group	0.66 (-0.36–1.68)	0.52	1.27	0.47 (-0.08–1.02)	0.28	1.69			
Time (delayed post-test) X Group	0.62 (-0.41–1.65)	0.53	1.17	0.26 (-0.29–0.80)	0.28	0.92	-0.04 (-0.40–0.31)	0.18	-0.24

Random Parts			
Intercept variance	1.911	0.778	0.423
Residual variance	1.881	0.558	0.238
Observations	245	247	165
$N_{subjects}$	83	83	83

Note. The final models were fitted with the *rlmer*-function of the *robustlmm* package. Each model comprised a random intercept for subject as well as time and group as fixed effects. The factor Time was dummy-coded (word recognition and word definition: pre-test = 0; picture naming: immediate post-test = 0), as was the factor Group (just reading control group = 0).

[a] The intercept represents the score of the control group at the pre-test. The effect of time represents the change in the control group from pre-test to immediate or delayed post-test. [b] As no pre-test was conducted for the picture naming task, the intercept represents the score of the control group at the immediate post-test. The effect of time represents the change in the control group from immediate to delayed post-test. [c] reference group = pre-test (word recognition and word definitions) or immediate post-test (picture naming). [d] reference group = just reading control group.

* $|t| > 2.0$ is considered indicating a statistically significant effect (Gelman & Hill, 2007).

To check integrity of our conditions, we compared the just-reading control and the question conditions regarding control-word learning. For control-words we assumed that there should not be any differences between the just-reading control and the question groups. In line with this, gains were comparable in size for the three measures of control-word acquisition at both measurement points (see Table 3; picture naming at the delayed post-test: *Estimate* = 0.10, *SE* = 0.21, t = 0.45).

Effects of Question Demand-Level

To examine predictions made by the different hypotheses regarding demand-level of questions, we compared the effects of low-demand, high-demand, and scaffolding-like conditions on target-word acquisition, and tested for an interaction with children's language skills. The aptitude-treatment-interaction hypothesis predicted that children with better general vocabulary profited most from high-demand questions and lower skilled children from low-demand questions (Hypothesis 2a). The high-demand hypothesis predicted high-demand questions (Hypothesis 2b) and the scaffolding hypothesis a scaffolding-like increase of question demand-level (Hypothesis 2c) to result in most learning gains, but only for slow mapping processes (word definitions) and not fast mapping processes (word recognition and word naming).

At first, we constructed basic models by including time and demand-level as fixed effects and a random intercept for subject (Table 4). In an additional model for each measure of target-word acquisition, we included general vocabulary knowledge to examine potential interaction effects with children's language skills.

Table 4

The Effect of Question Demand-Level on Acquisition of Target-Words

	Receptive vocabulary (word recognition)[a]			Expressive vocabulary (word definitions)[a]			Expressive vocabulary (picture naming)[b]		
	Estimate (95%-CI)	*SE*	*t*	*Estimate (95%-CI)*	*SE*	*t*	*Estimate (95%-CI)*	*SE*	*t*
Fixed Parts									
(Intercept)	11.39 (9.48–13.30)	0.98	11.67*	1.58 (-0.10–3.26)	0.86	1.85	5.52 (4.16–6.88)	0.69	7.96*
Time (immediate post-test)[c]	3.73 (2.63–4.84)	0.56	6.63*	4.74 (3.33–6.15)	0.72	6.61*			
Time (delayed post-test)[c]	3.46 (2.36–4.56)	0.56	6.15*	4.99 (3.58–6.39)	0.72	6.95*	-0.38 (-1.19–0.46)	0.41	-0.93
Demand-level (scaffolding-like)[d]	-1.09 (-3.76–1.59)	1.36	-0.80	0.08 (-2.25–2.41)	1.19	0.07	-0.26 (-2.16–1.63)	0.97	-0.27
Demand-level (high-demand)[d]	-0.39 (-3.03–2.26)	1.35	-0.29	0.59 (-1.71–2.90)	1.18	0.51	-0.04 (-1.92–1.84)	0.96	-0.04
Time (immediate post-test) X Demand-level (scaffolding-like)	0.22 (-1.32–1.76)	0.79	0.28	-1.09 (-3.03–0.86)	0.99	-1.09			

Time (delayed post-test) X Demand-level (scaffolding-like)	0.26 (-1.29–1.80)	0.79	0.33	-2.06 (-4.01–(-0.11))	0.99	-2.08*	-0.25 (-1.39–0.88)	0.58	-0.44
Time (immediate post-test) X Demand-level (high-demand)	0.46 (-1.06–1.98)	0.78	0.59	0.43 (-1.49–2.36)	0.98	0.44			
Time (delayed post-test) X Demand-level (high-demand)	0.54 (-0.99–2.08)	0.78	0.69	-1.16 (-3.10–0.78)	0.99	-1.17	-0.80 (-1.93–0.33)	0.58	-1.38
Random Parts									
Intercept variance	15.117			8.857			7.523		
Residual variance	3.011			4.720			1.624		
Observations	188			187			125		
$N_{subjects}$	63			63			63		

Note. The final models were fitted with the *rlmer*-function of the *robustlmm* package. Each model comprised a random intercept for subject as well as time and demand-level as fixed effects. The factor Time was dummy-coded (word recognition and word definition: pre-test = 0; picture naming: immediate post-test = 0), as was the factor Demand-level (low-demand = 0).

[a] The intercept represents the score of the low-demand condition at pre-test. The effect of time represents the change in the low demand condition from pre-test to immediate post-test or delayed post-test. [b] As no pre-test was conducted for the picture naming task, the intercept represents the score of the low-demand condition at immediate post-test. The effect of time represents the change in the low-demand condition from immediate to delayed post-test. [c] dummy-coded: reference group = pre-test (word recognition and word definitions) or immediate post-test (picture naming). [d] dummy-coded: reference group = low-demand.

* $|t| > 2.0$ is considered indicating a statistically significant effect (Gelman & Hill, 2007).

Results indicated that demand-level conditions were comparable regarding word recognition from pre-test to immediate post-test (high-demand vs. scaffolding-like: *Estimate* = 0.24, *SE* = 0.77, *t* = 0.31) as well as from pre-test to delayed post-test (high-demand vs. scaffolding-like: *Estimate* = 0.29, *SE* = 0.77, *t* = 0.37).

The same was also true for the picture naming task. Demand-level conditions did not differ at the immediate post-test (high-demand vs. scaffolding-like: *Estimate* = 0.22, *SE* = 0.95, *t* = 0.24) nor at the delayed post-test (scaffolding-like vs. low-demand: *Estimate* = -0.52, *SE* = 0.97, *t* = -0.54; high-demand vs. low-demand: *Estimate* = -0.84, *SE* = 0.96, *t* = -0.87; high-demand vs. scaffolding-like: *Estimate* = -0.32, *SE* = 0.95, *t* = -0.34).

With regard to word definitions, the groups did not differ regarding learning gains from pre-test to immediate post-test (high-demand vs. scaffolding-like: *Estimate* = 1.52, *SE* = 0.96, *t* = 1.58). Although there were no differences regarding learning gains at the delayed post-test between high-demand and low-demand conditions or high-demand and scaffolding-like conditions (*Estimate* = 0.90, *SE* = 0.97, *t* = 0.93), the scaffolding-like group showed significantly smaller target-word retention than the low-demand group.

Adding general vocabulary knowledge to the aforementioned basic models, results indicated, that there were no interactions between demand-level conditions and general vocabulary (all comparisons: $|t| < 2.0$).

To examine the general relationship between target-word knowledge, target-word learning and general vocabulary, while including demand-level, we reran the aforementioned models using effect coding for demand-level. Vocabulary knowledge was associated with higher target-word knowledge at the pre-test (word recognition:

Estimate = 2.45, *SE* = 0.41, *t* = 6.03; word definitions: *Estimate* = 0.95, *SE* = 0.33, *t* = 2.85) as well as with higher gains at the immediate post-test (word recognition: *Estimate* = 0.96, *SE* = 0.33, *t* = 2.88; word definitions: *Estimate* = 2.00, *SE* = 0.32, *t* = 6.18; picture naming: *Estimate* = 1.97, *SE* = 0.32, *t* = 6.11) and at the delayed post-test (word recognition: *Estimate* = 0.83, *SE* = 0.34, *t* = 2.49; word definitions: *Estimate* = 2.12, *SE* = 0.32, *t* = 6.54; picture naming: *Estimate* = 2.00, *SE* = 0.32, *t* = 6.18).

Effects of Question Placement

To examine the predictions made by the different hypotheses regarding the effects of question placement on target-word acquisition, we compared question placement, namely after the stories (Hypothesis 3a) versus within the stories (Hypothesis 3b), and its potential interaction with language skills, namely general vocabulary (Hypothesis 3c), regarding target-word learning.

For each test of target-word acquisition, in a first basic model we included time and question placement as fixed effects and a random intercept for subject. As question placement was operationalized as a within-factor and assignment of story combination could not be balanced perfectly, we controlled for story effects by including story combination and its interaction with time as covariates. For better interpretability, story combination was effect coded. In a second model, we added general vocabulary as fixed effect. The basic models of each measure of target-word acquisition are displayed in Table 5.

Table 5

The Effect of Question Placement on Acquisition of Target-Words

	Receptive vocabulary (word recognition)[a]			Expressive vocabulary (word definitions)[a]			Expressive vocabulary (picture naming)[b]		
	Estimate (95%-CI)	*SE*	*t*	*Estimate (95%-CI)*	*SE*	*t*	*Estimate (95%-CI)*	*SE*	*t*
Fixed Parts									
(Intercept)	5.33 (4.75–5.90)	0.29	18.24*	0.95 (0.48–1.42)	0.24	3.97*	2.55 (2.13–2.97)	0.22	11.81*
Time (immediate post-test)[c]	2.31 (1.83–2.79)	0.24	9.46*	2.19 (1.76–2.61)	0.22	10.09*			
Time (delayed post-test)[c]	2.12 (1.64–2.61)	0.25	8.59*	2.09 (1.66–2.52)	0.22	9.56*	-0.26 (-0.65–0.13)	0.20	-1.30
Placement (after)[d]	0.17 (-0.31–0.65)	0.25	0.70	-0.13 (-0.55–0.30)	0.22	-0.57	0.28 (-0.11–0.67)	0.20	1.41
Story Set (second combination)[e]	0.54 (0.30–0.78)	0.12	4.39*	-0.20 (-0.42–0.01)	0.11	-1.87	-0.28 (-0.48–0.09)	0.10	-2.86*
Time (immediate post-test) X Placement	-0.55 (-1.23–0.14)	0.35	-1.57	-0.00 (-0.61–0.60)	0.31	-0.01			

Time (delayed post-test) X Placement	-0.49 (-1.17–0.20)	0.35	-1.40	-0.38 (-0.99–0.23)	0.31	-1.23	-0.18 (-0.73–0.38)	0.28	-0.63
Time (immediate post-test) X Story Set	0.27 (-0.07–0.62)	0.17	1.58	0.19 (-0.11–0.49)	0.15	1.24			
Time (delayed post-test) X Story Set	0.17 (-0.18–0.51)	0.17	0.96	-0.03 (-0.33–0.28)	0.16	-0.18	-0.11 (-0.39–0.17)	0.14	-0.79
Random Parts									
Intercept variance	3.399			2.097			1.650		
Residual variance	1.815			1.427			1.186		
Observations	379			378			251		
$N_{subjects}$	65			65			64		

Note. The final models were fitted with the *rlmer*-function of the *robustlmm* package. Each model comprised a random intercept for subject as well as time, placement, story set, and their interactions as fixed effects. The factor Time was dummy-coded (word recognition and word definition: pre-test = 0; picture naming: immediate post-test = 0), as was the factor placement (within = 0). The factor Story Set was effect-coded (first story set = -1; second story set = 1).

[a] The intercept represents the score of the within-placement condition at pre-test. The effect of time represents the change in the within-placement condition from pre-test to immediate post-test or delayed post-test. [b] As no pre-test was conducted for the picture naming task, the intercept represents the score of the within-placement condition at immediate post-test. The effect of time represents the change in the within-placement condition from immediate to delayed post-test. [c] dummy-coded: reference group = pre-test (word recognition and word definitions) or immediate post-test (picture naming). [d] dummy-coded: reference group = within-placement. [e] effect-coded: first set = -1; second story set = 1.

* $|t| > 2.0$ is considered indicating a statistically significant effect (Gelman & Hill, 2007).

Results of the basic models indicated that question placement did not have any influence on target-word learning at the immediate post-test nor at the delayed post-test (picture naming: *Estimate* = 0.10, *SE* = 0.20, t = 0.51).

Adding general vocabulary to the aforementioned basic models, results indicated that there were no interactions between placement conditions and general vocabulary (all comparisons: $|t| < 2.0$).

To examine the general relationship between target-word knowledge, target-word learning and general vocabulary, while including question placement, we reran the aforementioned models using effect coding for question placement. Vocabulary knowledge was associated with higher target-word knowledge at the pre-test (word recognition: *Estimate* = 1.22, *SE* = 0.20, t = 6.13; word definitions: *Estimate* = 0.50, *SE* = 0.17, t = 3.00) as well as with higher gains at the immediate post-test (word recognition: *Estimate* = 0.44, *SE* = 0.18, t = 2.40; word definitions: *Estimate* = 0.99, *SE* = 0.15, t = 6.73; picture naming: *Estimate* = 0.99, *SE* = 0.17, t = 5.88) and at the delayed post-test (word recognition: *Estimate* = 0.45, *SE* = 0.18, t = 2.45; word definitions: *Estimate* = 1.01, *SE* = 0.15, t = 6.86; picture naming: *Estimate* = 0.99, *SE* = 0.17, t = 5.87).

Discussion

The current study aimed to extend previous studies on the effects of demand-level and placement of questions on learning words from listening to stories. In terms of demand-level, we systematically investigated different hypotheses proposed by the literature on the effects of questions on language learning (e.g., Blewitt et al., 2009; Reese & Cox, 1999; Walsh et al., 2016). Our focus was on several approaches, revolving around broader frameworks such as aptitude-treatment interactions and scaffolding. For question placement, we contrasted questions interspersed within the stories (i.e., the interruption hypothesis) as well as placement after the stories, and examined a potential interaction with general vocabulary knowledge (i.e., an aptitude-treatment-interaction hypothesis). To do so, we read four stories three times to young children in individual reading sessions and tested for acquisition of target-words, which we inserted into the plot of the stories. In addition, to control for the integrity of our conditions we included control-words of which the presentation (i.e., without questions) was identical across conditions.

Book Reading as a Mean to Foster Vocabulary Development

In line with our first hypothesis (Hypothesis 1a), which stated that repeated readings of stories should foster word learning, we found that target-word learning through listening to stories occurred irrespective of their presentation (i.e., with or without questions). This is in accord with findings from other studies (e.g., Blewitt et al., 2009; Biemiller & Boote, 2006; Elley, 1989; Vaahtoranta, Suggate, Jachman, Lenhart, & Lenhard, 2018). More importantly, in our study we were also able to show that learning gains were maintained, although somewhat reduced, until delayed post-test four weeks later.

Also in accordance with the bulk of relevant literature (e.g., Blewitt et al., 2009; Brabham & Lnych-Brown, 2002) and recent meta-analyses (Marulis & Neuman, 2010, 2013; Mol et al., 2008), it was possible to enhance learning gains by including accompanying questions, as proposed by Hypothesis 1b. In the just-reading control group, gains amounted at the immediate post-test to 10-20% of possible gain, whereas asking questions added another 10% to the performance of the experimental conditions (calculated from Table 2 by dividing the actual gain through the maximum possible gain). These numbers strongly resemble those obtained by Biemiller and Boote (2006).

As also indicated by our results, the effect of questions seemed to be more pronounced for expressive (picture naming and word definitions) than for receptive word learning (word recognition). A possible explanation might be that the production of newly learned words is more demanding than their mere recognition. Consequently, the question conditions that provided one additional encounter and focused on these target-words might have been more helpful for active retrieval in contrast to more passive word recognition.

As the just-reading control condition and the question conditions did not differ regarding control-word learning, other potential explanations such as inadvertent differences in quality of story reading or a better relationship between experimenter and child through higher degrees of interaction are unlikely. However, given that target- and control words did not differ regarding lemma frequency in both lexical databases, fewer control-words were acquired than target-words. In the just-reading control group, the respective values were approximately twice as high for target-words as for control-words. This, of course, could contribute to the lack of differences observed for control-words. As the focus in the present study was on the comparison between the experimental groups, we chose not to include any counter-balancing of target- and control words.

In addition, our results indicated that irrespective of story presentation higher general vocabulary knowledge led to higher gains in target-word learning. Therefore, our study provides further support for the so-called “Matthew-effect” in the domain of book reading, meaning that the larger a child’s language skill, the more readily the child learns new words from listening to stories (e.g., Blewitt et al, 2009; Penno, Wilkinson, & Moore, 2002).

Effects of Demand-Level and Placement of Questions

By including a low-demand, high-demand, scaffolding-like demand condition as well as a continuous measure of general vocabulary knowledge, we were able to systematically contrast the high-demand (Walsh et al., 2016), the scaffolding hypothesis (Blewitt et al., 2009), as well as the aptitude-treatment-interaction hypothesis (Reese & Cox, 1999), and test their predictions made for effects of demand-level. As we could not detect any differences between the conditions nor an interaction with children's language skills, the results are in contrast with predictions made by these hypotheses.

The same is true for the effects of question placement, where we tested predictions made by the interruption hypothesis (Brabham & Lynch-Brown, 2002), the no-interruption hypothesis (Dickinson & Smith, 1994), and the aptitude-treatment-interaction hypothesis (Reese & Cox, 1999). Again, contrary to proposed hypotheses, we could detect neither a general effect of question placement nor an interaction with children's language skills.

Theoretical and Practical Implications

At a first glance, our study seems to add to a confusing research picture, where empirical findings cannot be explained by one of the proposed hypotheses alone (e.g., Blewitt et al., 2009; Reese & Cox, 1999; Walsh et al., 2016), nor by a reasonable combination of two or more of them. However, one possibility to reconcile findings regarding placement and demand-level of questions may be by reflecting on the role of the experimenter in the literature. Most of the relevant studies (e.g., Blewitt et al., 2009; Justice, 2002; Reese & Cox, 1999; our current study) were conducted in one-to-one sessions, whereby the experimenters guide attention to specific aspects of a story by asking a set of predefined questions, namely those defining the specific question conditions. As the goal was to causally link these conditions to learning gains, the experimenters are typically not allowed to answer children's questions or provide any other information or feedback regarding the correctness of children's answers. Thus, the child-experimenter interaction represents only a small part of the dyadic cycle observed in naturally occurring shared-reading situations, in which asking questions is followed by dialogic interaction, such as parental feedback or explanations (Ninio & Bruner, 1978). Accordingly, the experimental situation may lack the meaningful interactions required to really invoke effects of different question placements and demand-levels.

Therefore, on a conceptual level, pure effects of placement or demand-level of questions have to be separated from effects of dyadic or whole-group discussion that may be stimulated by different types of questions. We propose that in the aforementioned cases, namely without the experimenter participating in the discussion and providing factual feedback, the effect of asking questions may – for children as young as in our sample – mainly consist of an attention focusing function on specific aspects of a story. Consequently, neither the placement nor the demand-level of questions per se might produce differential effects.

Moreover, differences might emerge through differential interaction and accompanying cognitive processing following different kinds of questions. Here the Brabham and Lynch-Brown (2002) study seems pertinent, which represents the only instance in which differences between placement conditions were found. There the adult narrator, who was the children's teacher, was allowed to stimulate interaction and to actively participate in the discussion.

Although Reese and Cox (1999), who had adults read books to individual children while adhering to a strict reading protocol, made sure that no extra-textual discussion could take place, this study differed from the other studies regarding intensity. With a duration of approximately six weeks, with two to three reading sessions per week, and reading 32 books in total, it might be possible that children habituated to specific types of questions. The long duration and the expectation of being asked similar questions each time might have caused the children to reflect on the questions and by doing so to prepare for or preempt these in the next session. Alternatively, the high intensity might have stimulated discussion between parents and children or encouraged children to require shared-reading by their parents in the way learnt from the numerous intervention sessions. These alternate explanations need to be addressed with further research.

To date, it seems that the only results which could not be explained by a differentiation between a mainly attention guiding effect of questions and effects of differential discussion following different questions is Blewitt et al.'s (2009) finding that a scaffolding-like approach was superior to either low- or high-demand questions regarding deeper word processing. Offering an alternative interpretation of Blewitt et al.'s finding, it is also conceivable that their scoring procedure might have had a large share in their results. According to their scoring procedure, separate points were awarded for every information unit (i.e., superordinate category, synonyms, perceptual properties, functional properties, or parts). By doing so, it could be possible that scoring favored the scaffolding-like condition, where questions differed in terms of whether conceptual or perceptual information was requested. Asking only perceptual questions (i.e., low-demand) or only conceptual questions (i.e., high-demand) might therefore have provided the children with similar pieces of information that could be repeated by the children, when they were asked to explain the target-words. Apart from that, there is concern that the fine-grained scoring of each information unit might pose problems regarding validity. In a recent review on assessing vocabulary learning in early childhood, Hoffman, Teale, and Paciga (2014) demonstrated that naming a higher number of information units does not necessarily represent deeper word knowledge. In addition, the only approach with which Hoffman et al. were able to achieve a reasonable degree of inter-rater reliability was on a three-point scale like that used by Biemiller and Boote (2006), Biemiller and Slonim (2001).

In summary, we propose that without real interaction, the effect of questions may mainly consist in their attention guiding function to specific aspects within a story. This attention guiding property does not depend on specifics like demand-level or placement of questions. However, if experimenters are allowed to mimic in full the dialogic circle observed during naturally book reading (Ninio & Bruner, 1978), namely by giving feedback on children's responses, answering children's questions, and extending children's utterances, different placement as well as demand-level conditions might stimulate differential dyadic interaction. Differential interaction in turn might lead to differential effects on language learning from these situations (e.g., Blewitt & Langan, 2016; Brabham & Lnych-Brown, 2002).

Limitations and Future Research

Several limitations have to be kept in mind when interpreting the results of the current study. The first limitation is the use of the *robustlmm* package (Koller, 2016), which, due to computational requirements, does not offer the possibility to evaluate the importance of specific predictors by means of model comparisons. However, as our sample was only medium-sized and inspection of the residuals or of the random effects indicated a violation of assumptions, we had to select robust model estimation, as provided by *robustlmm* package, which is achieved by using lower weights for outlying values.

A second limitation is the fact that by using a strict reading protocol, our study design was, albeit deliberately, artificial. In real-world contexts, parents or teachers naturally answer children's questions and give feedback on their utterances (e.g., Dickinson & Smith, 1994; Ninio, 1980; Ninio & Bruner, 1978). Yet, to maintain comparability with existing research (e.g., Blewitt et al., 2009; Justice, 2002; Reese & Cox, 1999) we deemed it necessary to separate the pure effect of question features from effects of discussion following these questions. Yet, this separation is also a strength of the current study, as systematic studies comparing effects of discussion stimulated by different kinds of questions are clearly needed.

A third limitation is that aptitude was not experimentally manipulated, for example by systematically providing different amounts of knowledge to different children, rendering tests of aptitude-treatment interaction only quasi-experimental. Instead, we used measures of children's vocabulary as indicators of aptitude.

Finally, a fourth limitation is that our study was restricted to a limited set of (shorter) children's stories. The effect of questions, and of different types of questions, could vary depending on the words being learned, the language complexity of the books, and the supports provided by the stories. Consequently, the results observed may be specific to easy stories and the type of target-words employed here, perhaps not being generalizable to all words that children learn from books.

Future research could focus on a number of areas to better understand the optimal conditions for shared-reading leading to language development. One clear direction is to extend the current study design by including more real-world interaction and discussion. By gradually decreasing the amount of experimental control and simultaneously increasing the amount and type of such more socially-valid interaction, it may be possible to disentangle effects of individual cognitive processing, attention guiding cues, and differential discussion stimulated by different question features. Further work could also seek to move beyond word learning to examine other semantic (e.g., narrative) and syntactic aspects of language and perhaps also consider knowledge acquisition. Finally, with regards to the huge impact of initial language skills on learning gains, just asking preschool children questions without providing sufficient feedback and assistance, seems to only widen the word gap. Future work needs to find better ways to foster at-risk children's vocabulary knowledge and to help those that are most disadvantaged.

References

Baddeley, A. (2003). Working memory and language: An overview. Journal of *Communication Disorders*, *36*(3), 189–208. https://doi.org/10.1016/S0021-9924(03)00019-4

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*, 328. https://doi.org/10.3389/fpsyg.2013.00328

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. Retrieved from https://arxiv.org/abs/1506.04967v1

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Biemiller, A. (2003). Vocabulary: Needed if more children are to read well. *Reading Psychology*, *24*(3-4), 323–335. https://doi.org/10.1080/02702710390227297

Biemiller, A. (2006). Vocabulary development and instruction: A prerequisite for school learning. In D. K. Dickinson & S. B. Neuman (Eds.), *Handbook of early literacy research* (Vol. 2, pp. 41–51). New York, NY: The Guilford Press.

Biemiller, A., & Boote, C. (2006). An effective method for building meaning vocabulary in primary grades. *Journal of Educational Psychology*, *98*(1), 44–62. https://doi.org/10.1037/0022-0663.98.1.44

Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of Educational Psychology*, *93*(3), 498–520. https://doi.org/10.1037/0022-0663.93.3.498

Björnsson, C.-H. (1968). *Lesbarkeit durch LIX*. (Tech. Rep. No. 6) [Readability through LIX]. Stockholm, Sweden: Pedagogiskt Centrum.

Blewitt, P., & Langan, R. (2016). Learning words during shared book reading: The role of extratextual talk designed to increase child engagement. *Journal of Experimental Child Psychology, 150*, 404–410. https://doi.org/10.1016/j.jecp.2016.06.009

Blewitt, P., Rump, K. M., Shealy, S. E., & Cook, S. A. (2009). Shared book reading: When and how questions affect young children's word learning. *Journal of Educational Psychology, 101*(2), 294–304. https://doi.org/10.1037/a0013844

Brabham, E. G., & Lynch-Brown, C. (2002). Effects of teachers' reading-aloud styles on vocabulary acquisition and comprehension of students in the early elementary grades. *Journal of Educational Psychology, 94*(3), 465–473. https://doi.org/10.1037/0022-0663.94.3.465

Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan, & G. A. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 265–293). New York, NY: Cambridge University Press.

Carey, S. (2010). Beyond fast mapping. *Language Learning and Development*, *6*(3), 184–205. https://doi.org/10.1080/15475441.2010.484379

Christian, K., Morrison, F. J., Frazier, J. A., & Massetti, G. (2000). Specificity in the nature and timing of cognitive growth in kindergarten and first grade. *Journal of*

Cognition and Development, *1*(4), 429–448. https://doi.org/10.1207/S15327647JCD0104_04

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Connor, C. M., Morrison, F. J., & Katch, L. E. (2004). Beyond the reading wars: Exploring the effect of child-instruction interactions on growth in early reading. *Scientific Studies of Reading*, *8*(4), 305–336. https://doi.org/10.1207/s1532799xssr0804_1

Connor, C. M., Piasta, S. B., Fishman, B., Glasney, S., Schatschneider, C., Crowe, E.,. . . Morrison, F. J. (2009). Individualizing student instruction precisely: Effects of child x instruction interactions on first graders' literacy development. *Child Development*, *80*(1), 77–100. https://doi.org/10.1111/j.1467-8624.2008.01247.x

Cromley, J. G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology*, *99*(2), 311–325. https://doi.org/10.1037/0022-0663.99.2.311

Dickinson, D. K., & Smith, M. W. (1994). Long-term effects of preschool teachers' book readings on low-income children's vocabulary and story comprehension. *Reading Research Quarterly*, *29*(2), 104–122. https://doi.org/10.2307/747807

Elley, W. B. (1989). Vocabulary acquisition from listening to stories. *Reading Research Quarterly, 24*(2), 174–187. https://doi.org/10.2307/747863

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Geyken, A. (2007). The DWDS corpus: A reference corpus for the German language of the 20th century. In C. Fellbaum (Ed.), *Idioms and collocations. Corpus-based linguistic and lexicographic studies* (pp. 23–41). London, UK: Continuum.

Grimm, H. (2010). *Sprachentwicklungstest für drei- bis fünfjährige Kinder (SETK 3–5)* [Speech Development Test For Three- to Five-Year-Old Children] (2nd ed.). Göttingen, Germany: Hogrefe.

Haden, C. A., Reese, E., & Fivush, R. (1996). Mothers' extratextual comments during storybook reading: Stylistic differences over time and across texts. *Discourse Processes*, *21*(2), 135–169. https://doi.org/10.1080/01638539609544953

Hargrave, A. C., & Sénéchal, M. (2000). A book reading intervention with preschool children who have limited vocabularies: The benefits of regular reading and dialogic reading. *Early Childhood Research Quarterly*, *15*(1), 75–90. https://doi.org/10.1016/S0885-2006(99)00038-1

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Paul H. Brookes.

Hart, B., & Risley, T. R. (2003). The early catastrophe: The 30 million word gap. *American Educator, 27*, 4–9.

Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., & Kliegl, R. (2011). dlexDB – eine lexikalische Datenbank für die psychologische und linguistische Forschung [dlexDB – A lexical database for the psychological and linguistic research]. *Psychologische Rundschau*, *62*(1), 10–20. https://doi.org/10.1026/0033-3042/a000029

Hoffman, J. L., Teale, W. H., & Paciga, K. A. (2014). Assessing vocabulary learning in early childhood. *Journal of Early Childhood Literacy*, *14*(4), 459–481. https://doi.org/10.1177/1468798413501184

Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods*, *39*(1), 101–117. https://doi.org/10.3758/BF03192848

Justice, L. M. (2002). Word exposure conditions and preschoolers' novel word learning during shared storybook reading. *Reading Psychology*, *23*(2), 87–106. https://doi.org/10.1080/027027102760351016

Kail, R., & Pellegrino, J. W. (1985). *Human intelligence: Perspectives and prospects*. New York, NY: Freeman.

Koller, M. (2016). robustlmm: An R package for robust estimation of linear mixed-effects models. *Journal of Statistical Software, 75*(6), 1–24. https://doi.org/10.18637/jss.v075.i06

Lenhard, A., Lenhard, W., Segerer, R., & Suggate, S. (2015). *PPVT 4: Peabody Picture Vocabulary Test - 4 (German Adaption)*. Frankfurt am Main, Germany: Pearson Assessment.

Lenhard, W., & Lenhard, A. (2014). Berechnung des Lesbarkeitsindex LIX nach Björnsson [Calculation of Björnsson's readability index LIX]. Bibergau, Germany: Psychometrica. Retrieved from http://www.psychometrica.de/lix.html

Marulis, L. M., & Neuman, S. B. (2010). The effects of vocabulary intervention on young children's word learning: A meta-analysis. *Review of Educational Research*, *80*(3), 300–335. https://doi.org/10.3102/0034654310377087

Marulis, L. M., & Neuman, S. B. (2013). How vocabulary interventions affect young children at risk: A meta-analytic review. *Journal of Research on Educational Effectiveness*, *6*(3), 223–262. https://doi.org/10.1080/19345747.2012.755591

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Melchers, P., & Preuß, U. (2009). *K-ABC: Kaufman Assessment Battery for Children - German version* (8th ed.). Frankfurt am Main, Germany: Pearson Assessment.

Mol, S. E., Bus, A. G., & de Jong, M. T. (2009). Interactive book reading in early education: A tool to stimulate print knowledge as well as oral language. *Review of Educational Research*, *79*(2), 979–1007. https://doi.org/10.3102/0034654309332561

Mol, S. E., Bus, A. G., de Jong, M. T., & Smeets, D. J. H. (2008). Added value of dialogic parent–child book readings: A meta-analysis. *Early Education and Development*, *19*(1), 7–26. https://doi.org/10.1080/10409280701838603

Ninio, A. (1980). Picture-book reading in mother-infant dyads belonging to two subgroups in Israel. *Child Development*, *51*(2), 587–590. https://doi.org/10.2307/1129299

Ninio, A., & Bruner, J. (1978). The achievement and antecedents of labelling. *Journal of Child Language*, *5*(1), 1–15. https://doi.org/10.1017/S0305000900001896

Penno, J. F., Wilkinson, I. A. G., & Moore, D. W. (2002). Vocabulary acquisition from teacher explanation and repeated listening to stories: Do they overcome the Matthew effect? *Journal of Educational Psychology*, *94*(1), 23–33. https://doi.org/10.1037/0022-0663.94.1.23

Petermann, F. (2011). *Wechsler Preschool and Primary Scale of Intelligence-III: WPPSI-III - German version* (2nd ed.). Frankfurt am Main, Germany: Pearson Assessment.

Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*(4), 413–425. https://doi.org/10.1016/j.jml.2008.02.002

R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Reese, E., & Cox, A. (1999). Quality of adult book reading affects children's emergent literacy. *Developmental Psychology*, *35*(1), 20–28. https://doi.org/10.1037/0012-1649.35.1.20

Schroeder, S., Würzner, K.-M., Heister, J., Geyken, A., & Kliegl, R. (2015). childLex: A lexical database of German read by children. *Behavior Research Methods, 47*(4), 1085–1094. https://doi.org/10.3758/s13428-014-0528-1

Sénéchal, M., Ouellette, G., & Rodney, D. (2006). The misunderstood giant: On the predictive role of early vocabulary to future reading. In D. K. Dickinson & S. B. Neuman (Eds.), *Handbook of early literacy research* (Vol. 2, pp. 173–182). New York, NY: The Guilford Press.

Snow, C. E., & Goldfield, B. A. (1983). Turn the page please: Situation-specific language acquisition. *Journal of Child Language, 10*(3), 551–569. https://doi.org/10.1017/S0305000900005365

Stahl, S. A., & Fairbanks, M. M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research, 56*(1), 72–110. https://doi.org/10.3102/00346543056001072

Statistisches Bundesamt. (2017). *Bildungsstand der Bevölkerung 2016* [Educational level of the population 2016]. Retrieved from https://www.destatis.de/DE/Publikationen/Thematisch/BildungForschungKultur/Bildungsstand/BildungsstandBevoelkerung5210002167004.pdf?__blob=publicationFile

Twisk, J., de Boer, M., de Vente, W., & Heymans, M. (2013). Multiple imputation of missing values was not necessary before performing a longitudinal mixed-model analysis. *Journal of clinical epidemiology, 66*(9), 1022–1028. https://doi.org/10.1016/j.jclinepi.2013.03.017

Vaahtoranta, E., Suggate, S., Jachmann, C., Lenhart, J., & Lenhard, W. (2018). Can explaining less be more? Enhancing vocabulary through explicit versus elaborative storytelling. *First Language, 38*(2), 198–217. https://doi.org/10.1177/0142723717737452

van Kleeck, A., Gillam, R. B., Hamilton, L., & McGrath, C. (1997). The relationship between middle-class parents' book-sharing discussion and their preschoolers' abstract language development. *Journal of Speech Language and Hearing Research, 40*(6), 1261–1271. https://doi.org/10.1044/jslhr.4006.1261

Walker, D., Greenwood, C., Hart, B., & Carta, J. (1994). Prediction of school outcomes based on early language production and socioeconomic factors. *Child Development, 65*(2), 606–621. https://doi.org/10.1111/j.1467-8624.1994.tb00771.x

Walsh, B. A., Sánchez, C., & Burnham, M. M. (2016). Shared storybook reading in Head Start: Impact of questioning styles on the vocabulary of Hispanic dual language learners. *Early Childhood Education Journal, 44*(3), 263–273. https://doi.org/10.1007/s10643-015-0708-3

Wasik, B. A., Hindman, A. H., & Snell, E. K. (2016). Book reading and vocabulary development: A systematic review. *Early Childhood Research Quarterly, 37*, 39–57. https://doi.org/10.1016/j.ecresq.2016.04.003

Whitehurst, G. J., Arnold, D. S., Epstein, J. N., Angell, A. L., Smith, M., & Fischel, J. M. (1994). A picture book reading intervention in day care and home for children from low-income families. *Developmental Psychology, 30*(5), 679–689. https://doi.org/10.1037/0012-1649.30.5.679

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, *17*(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Appendix 1

Target- and Control-Words Used in the Current Study

Target- and control-words (German)	Approximate English translation	Lemma frequency ChildLex	Lemma frequency DWDS
Target-words			
anketten	to chain	0	38
Atoll	atoll	4	43
aushändigen	to hand (something)	11	481
Barke	skiff	0	104
Besserwisser	know-all	7	50
Buddel	bottle	6	24
Eisbrecher	icebreaker	5	94
Geäst	branches	23	166
Hauer	fang	7	133
Heimstätte	homestead	1	188
Kanapee	settee	2	176
Katapult	catapult	2	44
Klaue	claw	19	255
Korkenzieher	corkscrew	6	45
Leu	lion	3	31
Lurch	amphibian	35	104
Ödland	wasteland	2	60
Pfuhl	puddle	0	46
Reisig	brushwood	20	148
Schemen	spectre	0	0
schlummern	to doze	42	356
skandieren	to chant	0	36
(sich) suhlen	to wallow	6	28
zanken	to squabble	34	312

Control-words			
(sich) amüsieren	to enjoy oneself	26	491
Beistand	support	7	816
hechten	to dive headlong	79	14
Robe	gown	6	153
säubern	to cleanse	44	796
Schwätzchen	chin wag	0	0
sinnieren	to muse	0	69
Sud	brew	7	81
trällern	to warble	29	96
Tümmler	porpoise	0	22
Warze	wart	20	128
zusammenbrauen	to concoct	14	44

Note. Lemma frequencies represent absolute frequencies per million tokens/words in the respective corpora.

Highlights

- Effects of question placement and demand-level during storytelling were examined.
- Asking questions during storytelling enhanced vocabulary learning.
- Placement and demand-level did not exert differential effects on word learning.
- Placement or demand-level did not interact with language skills.
- Better vocabulary was associated with greater gains.

7 Study 4: Shared-Reading in Small Groups: Examining the Effects of Question Demand-Level and Placement

A version of this chapter was submitted as:

Lenhart, J., Suggate, S., Lenhard, W., & Vaahtoranta, E. (submitted). Shared-reading in small groups: Examining the effects of question demand-level and placement. *Manuscript submitted for publication.*

Draft version 15.02.2019. This paper has not (yet) been published and is not therefore the authoritative document of record. Please do not copy or cite without author's permission.

Abstract

Shared-reading is an effective means of fostering preschool children's vocabulary development, with questions being widely used to increase learning outcomes. Theoretical explanations on effects of specific question features are, however, often contradictory. To examine different predictions relating to question demand-level and placement, two stories were read in a pre-post-design to a sample of 91 four- to six-year-old preschool children in small groups (five to seven children), whereby target-word learning was measured via word recognition and definition tasks. Demand-level (low-demand vs. high-demand vs. scaffolding-like) was operationalized as a between-subjects factor and placement (during vs. after the story) as a within-subjects factor. In addition, as controls, one group received a just-reading condition and control words (not accompanied by questions) were interspersed in the stories. Results indicate that children's target-word learning profited from asking questions. However, contrary to predictions made by different theories, question placement, demand-level or its interaction with children's vocabulary size or phonological working memory had no influence on learning gains.

Keywords: questioning style; shared-reading; word learning; read-aloud; vocabulary acquisition

Educational Impact and Implications Statement

The present study shows that reading stories to kindergarten children enhances vocabulary development. We found that asking children questions during the stories bolstered the effectiveness of shared reading, regardless of whether questions are placed during or after the story, or whether these are easy or difficult. The results underline the central role that adults play not only for engaging children through questions, but also for providing corrective feedback and for elaborating children's utterances.

Introduction

Children's early language skills are a key determinant of their academic success (Scarborough, 1998), with vocabulary size at school entry being an important predictor of later achievement (Cromley & Azevedo, 2007; Cunningham & Stanovich, 1997; Sénéchal, Ouellette, & Rodney, 2006). However, children from socioeconomically disadvantaged backgrounds display considerably lower vocabulary knowledge than their peers (Biemiller & Slonim, 2001; Locke, Ginsborg, & Peers, 2002). These differences emerge during early childhood (Fernald, Marchman, & Weisleder, 2013; Hart & Risley, 1995, 2003), and tend to persist if not targeted by intervention (Biemiller & Slonim, 2001; Christian, Morrison, Frazier, & Massetti, 2000).

For young children, interventions embedded within shared-reading situations are frequently used to address vocabulary and other language deficits. These either aim to improve parents' natural shared-reading behavior (e.g., Aram, Fine, & Ziv, 2013; Niklas & Schneider, 2015) or to implement shared-reading sessions in kindergarten, preschool, and school environments (e.g., Beck & McKeown, 2007; Biemiller & Boote, 2006). Both types of shared-reading interventions effectively help to foster vocabulary knowledge (meta-analyses report overall effect sizes between d = 0.42 to 0.88; see Marulis & Neuman, 2010, 2013; Mol, Bus, & de Jong, 2009; Mol, Bus, de Jong, & Smeets, 2008; Stahl & Fairbanks, 1986).

Regarding specific intervention types, numerous studies have demonstrated positive effects of providing explicit definitions, teaching, or elaborations (e.g., Ard & Beverly, 2004; Biemiller & Boote, 2006; Blewitt & Langan, 2016; Silverman, 2007). The influence of asking children questions, in contrast, is less well researched. Although questions seem to have a positive impact on vocabulary learning, generally their effect is enhanced substantially through additional explanations (Ard & Beverly, 2004) or subsequent extended discussion with the narrator (Blewitt & Langan, 2016). However, as highlighted by recent reviews, the contribution and importance of specific factors and moderators need further clarification (Flack, Field, & Horst, 2018; R. L. Walsh & Hodge, 2018; Wasik, Hindman, & Snell, 2016).

Such clarification is particularly needed with respect to the effects of questioning during shared-reading, which is regarded as a key factor for successful interactive book reading. Although the use of questions generally enhances learning gains compared to just-reading conditions (Marulis & Neuman, 2010, 2013; Mol et al., 2008; Wasik et al., 2016), the importance of question characteristics has still to be resolved. Particularly, research needs to determine: (a) how challenging or demanding such questions should be (i.e., “demand-level”; e.g., Blewitt, Rump, Shealy, & Cook, 2009; B. A. Walsh, Sánchez, & Burnham, 2016), (b) when questions should be asked (i.e., “placement”; e.g., Brabham & Lynch-Brown, 2002; Blewitt et al., 2009), and (c) whether question demand-level or placement interact with children’s language skills (e.g., Blewitt et al., 2009; Lenhart, Lenhard, Vaahtoranta, & Suggate, 2019; Reese & Cox, 1999).

Several theoretical and instructional frameworks exist that are commonly used to explain and predict effects of question demand-level and question placement on vocabulary acquisition. As outlined next, these different approaches lead to conflicting assumptions on how to best instruct children.

Question Demand-Level during Shared-Reading

In general, the demand-level of questions during shared-reading is defined by the complexity of cognitive processing required to provide an appropriate answer

(Haden, Reese, & Fivush, 1996). With regard to methodology, studies typically use a dichotomized categorization of demand-level (e.g., Blewitt et al., 2009; Hassinger-Das et al., 2016; Reese & Cox, 1999). Low-demand questions require only shallow cognitive processing such as recognizing, reproducing, or repeating information, explicitly stated in the story. High-demand questions, in contrast, might require children to link story information to prior knowledge, combine distinct ideas in the story, or draw inferences (Blewitt et al., 2009; Justice, 2002; B. A. Walsh et al., 2016). Thus, low-demand questions often involve simple perceptual descriptions (e.g., "What is she doing?", "Where is he doing that?", "How is he doing it?", "Can you point to xyz on the picture?") whereas high-demand questions involve conceptual inferences, predictions, and evaluations (e.g., "Why is she doing that?"). In terms of the effect of question demand-level on the acquisition of novel words during shared-reading, three theoretical explanations resulting in differing predictions have been proposed so far.

Aptitude-treatment-interaction approach. One theory with regard to the role of demand level (Reese & Cox, 1999) constitutes a variant of the well-known aptitude-treatment-interaction (e.g., Connor, Morrison, & Katch, 2004; Connor et al., 2009). This aptitude-treatment approach assumes that demand-level of questions and children's language skills (such as general vocabulary) interact, whereby low-performing children profit most from low-demand questions, whereas high-performing children learn most from high-demand questions.

The empirical evidence in shared-reading research for the treatment-aptitude approach is, at best, mixed. Reese and Cox's (1999) study, in which 48 (n_{group} = 16) four-year-old children were read approximately 30 stories in a one-on-one setting across the span of six weeks, found an interaction between demand-level and children's receptive vocabulary. Children with good vocabulary gained most through high-demand questions placed after the story, whereas children with low vocabulary profited most from low-demand questions that were interspersed within the story. Other studies, however, did not replicate this finding (Blewitt et al., 2009; Lenhart et al., 2019). Lenhart et al. (2019), for instance, read four stories, three times in succession, in one-on-one settings to children aged between four and six years (N = 86, n_{group} = 21 to 22). Although they found a significant effect of general vocabulary knowledge on the acquisition of novel words from listening to stories, there was no evidence for an interaction between vocabulary and question demand-level. Similarly, Blewitt et al. (2009) who read books individually to a group of three-year-olds (Study 1: N = 60, n_{groups} = 11 to 12; Study 2: N = 50, n_{groups} = 16 to 17) did not find any evidence for an interaction between children's language performance and question demand-level.

High-demand approach. A second theory, the high-demand theory proposed by B. A. Walsh et al. (2016), incorporates the distinction between surface and deeper word learning, guided by Carey's (1978) fast and slow mapping concepts in novel word acquisition. The term fast mapping tries to capture the extremely fast acquisition of words and word features in young children, such as the phonological sound pattern and word meaning, even after minimal exposure (Carey & Bartlett, 1978). In contrast, during slow mapping, children build systematically on the initial learning process to derive richer lexical understanding (Carey, 2010; Swingley, 2010).

Based on the distinction between surface and deeper word knowledge, B. A. Walsh et al. (2016) argue that questions provide sufficient information for initial word learning (i.e., fast mapping) irrespective of demand-level, but that demand-level of questions may have differential effects on deeper encoding processes and on the extension of partial word knowledge. They assume that high-demand questions should be particularly beneficial for deeper word knowledge (i.e., ability to explain words), as they require deeper cognitive processing and therefore should help children to create richer concepts of newly learned words. Word surface knowledge (i.e., simple word recognition), in contrast, should not be influenced by question demand-level and could be built by a sufficient number of encounters with novel words.

Results of observational studies in schools (e.g., Dickinson & Smith, 1994; Gonzalez et al., 2014) and at home (e.g., Haden et al., 1996) indeed support the notion that cognitively challenging questions may foster language development. Experiments that focus on effects of question demand-level on young children's word learning from shared-reading have, however, not provided support for the high-demand theory (Blewitt et al., 2009; Justice, 2002, N = 23, n_{groups} = 11 to 12; Lenhart et al., 2019). In these studies, low- and high-demand questions differed neither on measures of surface word learning, such as word recognition, nor on measures of deeper word learning, such as word definitions.

The only shared-reading study, lending some support for the high-demand theory, is B. A. Walsh et al.'s (2016) in which a sample of three- to five-year-old Hispanic dual language learners enrolled in Head Start (N = 57, n_{group} = 9 to 16) showed greater deeper, but not surface, word learning in high-demand conditions over low-demand question conditions (mean d = 1.51). However, none of the question conditions resulted in statistically significant greater vocabulary gains than a just-reading control group, casting doubt on the extent to which the findings actually support the high-demand hypothesis.

Scaffolding-like approach. A third approach (Blewitt et al., 2009) centering on the concept of scaffolding (see Wood, Bruner, & Ross, 1976) also distinguishes between surface and deeper word knowledge. Similar to the high-demand approach, the theory focuses on deeper word learning. Blewitt et al. argue that a scaffolding-like approach, in the sense of a transition from low- to high-demand questions, might best facilitate the acquisition of deeper word knowledge. In contrast, and also in line with the assumptions of the high-demand theory, word surface knowledge (i.e., simple word recognition) should not be influenced by question demand-level and should be acquired by mere encounters with novel words.

Consistent with the scaffolding hypothesis, Blewitt et al. (2009; Study 2: $N = 50$, $n_{\text{groups}} = 16$ to 17) found that the scaffolding condition did not differ from either low- or high-demand conditions regarding performance on a measure of surface learning (word recognition), but that it was beneficial for building deeper word knowledge (word definitions: mean $d = 0.89$). In contrast, Lenhart et al. (2019) tried to replicate Blewitt et al.'s study with slightly older children ($N = 86$, $n_{\text{group}} = 21$ to 22) and reported no significant differences between the conditions. However, as these two are the only studies to have examined a scaffolding-like presentation of question demand-level in shared book reading, coupled with concerns as to the validity of Blewitt et al.'s (2009) word definition scoring method (see Hoffman, Teale, & Paciga, 2014), more work is needed to evaluate Blewitt's scaffolding hypothesis.

Question Placement during Shared-Reading

Regarding question placement, literature does not distinguish between surface and deeper word learning and typically reports two conditions. Questions are either interspersed in the flow of the story or placed around the story, namely before and/or after the story (Brabham & Lynch-Brown, 2002; Blewitt et al., 2009; Reese & Cox, 1999; B. A. Walsh et al., 2016). Accordingly, two contrasting assumptions regarding the influence of question placement on word learning from listening to stories can be found in the shared-reading literature (see Brabham & Lynch-Brown, 2002). The first assumes that a non-interrupted story presentation facilitates story flow and focusing on the story gist, producing subsequent benefits for word learning. A second school of thought favors interspersing questions within the story, to provide better opportunities to reflect on and clarify difficult aspects, thereby enhancing comprehension and learning of novel words.

In addition to these two approaches, a third proposition might draw on the aptitude-treatment-interaction framework (see Wood et al., 1976). Specifically, question placement may not have a general effect on word learning, but asking questions or providing information during the stories versus after the stories might interact with children's cognitive abilities, such as their vocabulary knowledge (Reese & Cox, 1999) or phonological working memory (Jimenez & Saylor, 2017).

As with studies on question demand-level, empirical research on the effects of question placement has also been somewhat inconclusive. An observational study in 25 kindergarten classrooms of four-year-old low-income children (Dickinson & Smith 1994), for example, reported that a non-disruptive reading style with high-demand questions before and after story presentation seemed to best foster general vocabulary development. Conversely, the results of an experimental study conducted in whole classroom settings with first and third graders (N = 246, n_{group} = 39 to 48) favored the condition with questions interspersed throughout the stories (mean effect size in first grade for vocabulary d = 1.05; Brabham & Lynch-Brown, 2002). Furthermore, experimental studies, not allowing any experimenter feedback or elaboration conducted with kindergarten children in one-on-one (Blewitt et al., 2009; Lenhart et al., 2019: within condition n = 65) or small group reading settings (B. A. Walsh et al., 2016), found no differences between both types of question placement.

Turning to a potential aptitude-treatment interaction, Reese and Cox's study (1999) seems to suggest that question placement may interact with children's vocabulary knowledge. Although they reported that a "describer style", namely easy questions asked within the stories, produced the highest overall vocabulary gains, children with better vocabulary profited most from a "performance-oriented style", marked by difficult questions after the story. However, both styles were not significantly different from a condition in which difficult questions were interspersed within the story (i.e., "comprehender style"). Unfortunately, as the study design did not comprise a condition with easy questions after each story, effects of question placement and demand-level are difficult to separate.

To separate placement from demand level, Jimenez and Saylor (2017) examined only effects of question placement on novel word learning and story comprehension. Although in their sample of three- to six-year-old children (N = 83; n_{group} = 41 to 42) they found no general differences as to whether the questions were placed inside or outside the stories, they reported a significant interaction between verbal memory and placement on word learning (β = .29). There was a stronger relationship between memory and vocabulary learning in the "inside" condition than in the "outside" condition (β_{inside} = .43, β_{outside} = .08, Cohen's q = .38). Consequently, Jimenez and Saylor argue that interactive instruction, such as questions and/or elaborations, during the stories may put a larger demand on children's cognitive load than vocabulary instruction after the stories, and may therefore only be suitable for children with better phonological working memory.

However, comparable to research on question demand-level, the heterogeneity of design features (sample size, individual vs. group settings, sample characteristics, pre-post vs. posttest only) impedes integration of findings into a coherent picture favoring any of the proposed theories. Consequently, there is a clear need for additional research.

The Current Study

Summing up the studies on question demand-level and placement, there seems to be little empirical evidence strongly favoring any of the theories or approaches, in terms of both demand-level and placement. Methodologically, one reason may be that most studies contained small sample sizes, for instance Blewitt et al. (2009), Justice (2002), Reese and Cox (1999), or B. A. Walsh et al. (2016) report sample sizes of less than 18 children per group, sometimes even with groups as small as only nine to eleven children (e.g., Blewitt et al., 2009, Study 1; Justice, 2002, B. A. Walsh et al., 2016). Consequently, in terms of evaluating the competing theoretical frameworks, studies may have lacked sufficient statistical power. Furthermore, although constituting important work, some study designs did not allow for teasing out placement from demand-level effects (e.g., Dickinson & Smith, 1994; Reese & Cox, 1999). In short, our understanding of how to best enhance the important medium of shared story reading via question demand-level and placement is unsatisfactory.

Accordingly, in the current study, we aim to test different frameworks while addressing methodological limitations. Further, we seek to bridge the gap between highly-controlled experimental studies (e.g., Blewitt et al., 2009; Lenhart et al., 2019) and less-controlled field studies (e.g., Dickinson & Smith, 1994; Gonzalez et al., 2014). Therefore, we used a small group design, in which children were encouraged to discuss the questions, thereby stimulating natural group interaction. Here, the adult narrators' job was to read the stories, present a predefined set of questions, and to initiate and sustain children's discussion of their answers to the questions.

In this context, we tested the aptitude-treatment, the high-demand, and the scaffolding theory for question demand-level as well as a non-interruptive, an interruptive, and an aptitude-treatment question placement approach. To do so, we used a comprehensive experimental design, that included question demand-level as a between subjects condition, with selected words presented in low-demand, high-demand, or scaffolding-like (i.e., increasing from low- to high-demand) question conditions, all of which also varied by question placement (within vs. after the story), thus constituting a within-subjects factor. Additionally, we included two types of control. As a between-subjects control we added a just-reading group, in which no questions were asked, and as a within-subjects control each text also comprised control words, which were not accompanied by questions in any of the conditions.

In line with previous research, we examined the following research questions and hypotheses. First, we assume that questions generally facilitate children's word learning in comparison to a just-reading control group (Hypothesis 1a). Reflecting this, there should be no differences between question and just-reading conditions regarding words that were not targeted by questions (i.e., control words) (Hypothesis 1b). Additionally, to exclude the possibility that enhanced target-word learning in the question groups might only be an effect of target-word repetition within the question (for an elaboration of this point, see for example Blewitt et al., 2009), we included target-word repetition (one occurrence vs. repetition) as a within-subjects factor in the just-reading control group:

- H1a: Children in the question groups show better target-word learning than their peers in the just-reading control group.
- H1b: Children in the question groups and in the just-reading control group show no differences regarding control-word learning.

Second, regarding demand-level of questions, we tested the aptitude-treatment theory (Reese & Cox, 1999), the high-demand theory (B. A. Walsh et al., 2016), and the scaffolding theory (Blewitt et al., 2009). The high-demand theory assumes that children's deeper word learning (i.e., measured via word definitions) is best supported through high-demand questions (Hypothesis 2a), whereas the scaffolding theory posits deeper word learning for questions that gradually increase from low- to high-demand (Hypothesis 2b). Both theories have in common that they assume no differences between conditions regarding surface learning of novel words (i.e., word recognition). The aptitude-treatment theory, in contrast, makes no such differentiation between surface and deeper word learning and predicts that children with good language skills generally (i.e., regarding word recognition as well as word definitions) learn best via high-demand questions and those with poor language skills through low-demand questions (Hypothesis 2c):

- H2a (high-demand theory): Although there may be no differences in surface target-word learning between demand-level groups, the high-demand condition best fosters deep target-word learning.

- H2b (scaffolding-like theory): Although there are no differences regarding surface target-word learning between demand-level groups, the scaffolding-like condition fosters deep target-word learning best.
- H2c (aptitude-treatment theory): Children with good language skills acquire most surface and deeper target-word knowledge from high-demand questions and those with poor language skills from low-demand questions.

Third, regarding question placement, we examined the influence of a non-interruptive (e.g., Dickinson & Smith, 1994) versus an interruptive (e.g., Brabham & Lynch-Brown, 2002) question placement condition on novel word learning. Although the non-interruptive approach favors questions either before or after the story reading (Hypothesis 3a), the interruptive approach predicts better novel word acquisition through questions interspersed within the stories (Hypothesis 3b). In addition, we also examined a potential interaction of children's phonological working memory with question placement, namely that question within the stories increase cognitive load and should therefore better suitable for children exhibiting good memory (e.g., Jimenez & Saylor, 2017):

- H3a (non-interruptive approach): Questions after or before the story best foster target-word learning.
- H3b (interruptive approach): Questions embedded within the story best foster target-word learning.
- H3c (aptitude-treatment theory): Children with good phonological working memory learn most words from questions interspersed within the story and those with poor working memory from questions after the story.

Fourth, we tested whether there was any interaction between question demand-level and placement. Reese and Cox's (1999) study as well as an observational study by Dickinson and Smith (1994) point to the possibility that there may indeed by an interaction between both factors, or even a higher order interaction that includes children's language skills. In particular, high-demand questions may be especially beneficial if they are placed after the story (Dickinson & Smith, 1994; Hypothesis 4a), which in turn may be moderated by children's language abilities, with children displaying good vocabulary knowledge profiting most from high-demand questions after each story (Reese & Cox, 1999; Hypothesis 4b):

- H4a: There is an interaction between question placement and demand-level, with high-demand questions asked after the story resulting in highest target-word learning.
- H4b: There is an interaction between question placement, demand-level, and children's language skills, with children displaying a good general vocabulary profiting most from questions asked after the story.

Methods

Design

The study comprised a pre-post-design (see Figure 1) with four between groups, namely a just-reading control group and three question groups, in which selected words (i.e., target words) were accompanied by either low-demand, high-demand, or scaffolding-like (i.e., increasing from low- to high-demand) questions. In the question conditions, we included question placement (within the stories vs. after the stories) as a within-subjects factor.

Figure 1

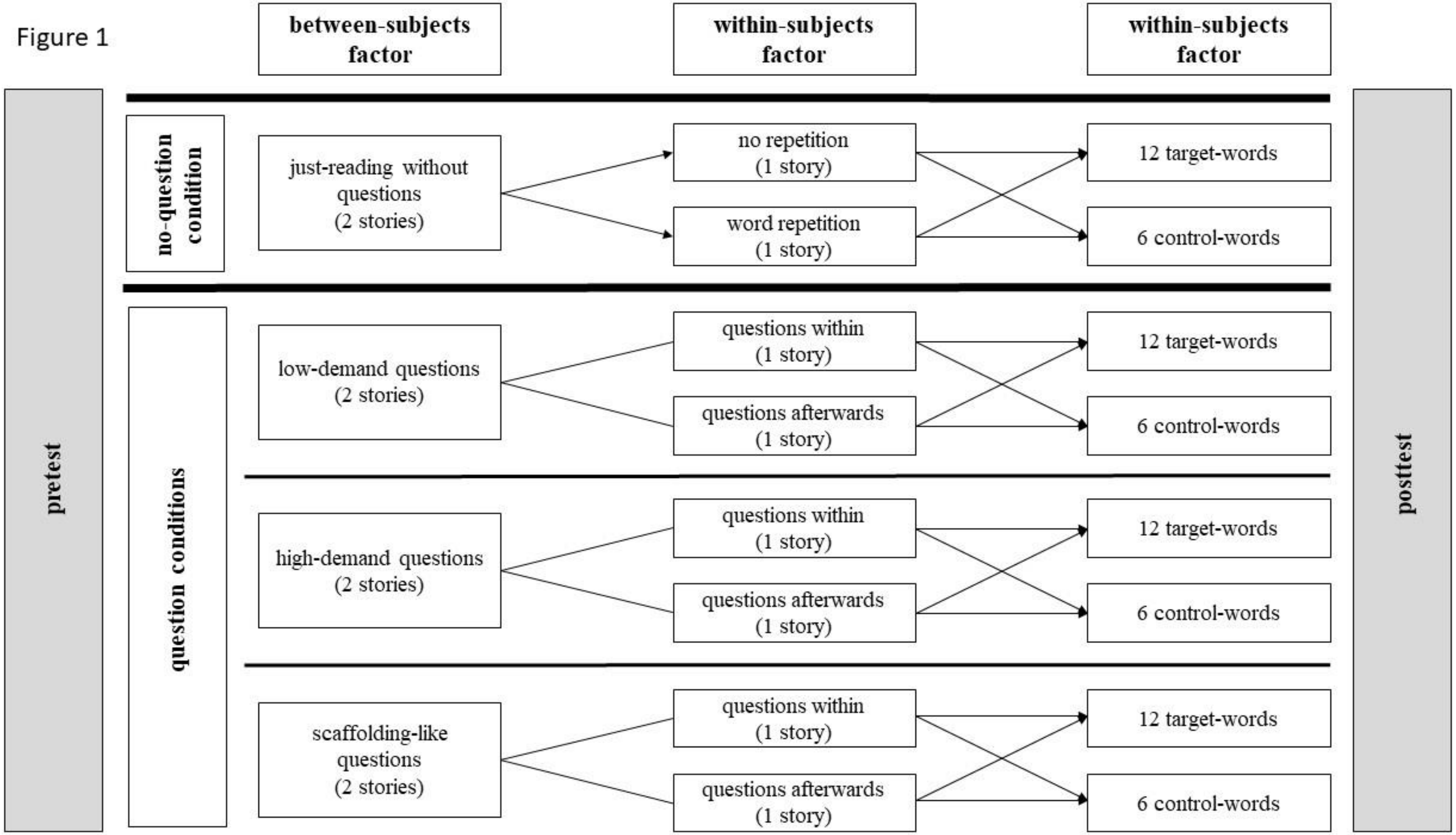

Figure 1. Overview of the study design.

To examine the general effects of questions more closely, we also included two types of additional controls in our study design. First, to investigate whether questions affect only the story aspects to which they directly refer (i.e., the target words), we added control words in each story that were not accompanied by questions. Second, to ensure that the questions and not simply the pure number of target-word repetitions resulted in better target-word learning, half the target words in the just-reading control group were repeated in a sentence following their first mention (within-subjects factor: one word occurrence per story = no repetition vs. word repetition).

Following the proposed distinction between surface and deeper word learning, we adhered to testing procedures in Blewitt et al.'s (2009) and B. A. Walsh et al.'s (2016) studies and included a word recognition test at the pre- and the posttest to measure surface learning and a word definition test to assess deeper word knowledge. In addition, to examine a potential aptitude-treatment interaction, we added continuous measures of general receptive and expressive vocabulary knowledge and of phonological working memory.

Power Analysis

We conducted a power analysis with *power analysis for general ANOVA designs* (PANGEA, v.02; Westfall, 2016). To date, there is no generally accepted analytical way to calculate power analyses for complex generalized linear mixed-effects models and procedures that rely on simulation require presuppositions on various components of the models. We therefore chose to make power analyses for the corresponding ANOVA designs, which can be expected to constitute a conservative estimate of power for our study design.

Based on previous studies that used a just-reading control group (Blewitt et al., 2009, Study 1: immediate posttest mean $d = 1.12$; Brabham & Lynch-Brown, 2002: mean $d = 1.01$; Lenhart et al., 2019: immediate posttest mean $d = 0.91$) we expected an effect size of at least $d = 0.80$ for target-word learning differences between just-reading control and question groups. With $d = 0.80$, alpha at .05, and power = .80, 17 participants per group (i.e., 68 in total) would be needed for the analysis of word-learning differences between just-reading control group and questions groups, which represents a 2 (pre- vs. posttest) x 4 (just-reading vs. high-demand vs. low-demand vs. scaffolding-like group) mixed-design.

Studies that lend support to the respective theories on effects of question placement and demand-level report even larger effect sizes with $d > 0.80$ (Blewitt et al., 2009, Study 2; Brabham & Lynch-Brown, 2002; B. A. Walsh et al., 2016). For a conservative estimate, we included studies that did not report significant effects for some or all of the comparisons (e.g., Blewitt et al., 2009, Study 1; Justice, 2002; Lenhart et al., 2019), resulting in a mean effect size $d = 0.57$ across relevant studies for target-word learning differences between question conditions. With $d = 0.57$, alpha at .05, and power = .80, 21 participants per group were needed for the analysis of target-word learning differences between question demand-level and placement conditions and their interaction, which represents a 2 (pre- vs. posttest) x 2 (question during vs. question after the story) x 3 (high-demand vs. low-demand vs. scaffolding-like group) mixed-design.

Participants

Parents of 97 children across six kindergartens in a German city (approximately 150 000 inhabitants) provided written consent for their child's participation. As six of those children missed two or all of the three book reading sessions, they were not included in the analyses. The remaining 91 children (53% female) had a mean age of 64.14 months (*SD* = 7.60; *Min* = 49; *Max* = 79). The sample consisted primarily of native German speakers, with only 13% of the children having one parent and 12% both parents born in a country other than Germany. The primary language was German, with 76 % of the children speaking only German at home. The mean educational level of our sample was with 74% of the mothers and 80% of the fathers having received a high-school diploma or a university degree, which is noticeably higher than that of the German population (Federal Bureau of Statistics, 2017).

Procedure

All sessions of the study took place in a separate room in the kindergartens and were conducted by trained student research assistants. The pre- and posttest sessions were conducted in one-on-one settings, whereas the shared-reading session took place in 16 small groups of five to seven children. Children were randomly assigned to these shared-reading groups in their kindergartens. The groups in turn were randomly assigned to one of the four book-reading conditions. Although we endeavored to assign the narrators (*N* = 9) randomly to the individual group sessions, scheduling constraints in the kindergartens made it necessary to deviate sometimes from this procedure. However, there was no significant relationship between specific narrators and condition assignment (*p* = .698, two-tailed Fisher's exact test).

The sequence of the study was as follows: First, the pretest was conducted individually in a separate session approximately two weeks before the reading intervention started (*M* = 15.90 days, *SD* = 5.88). Then three reading sessions were held in the course of two weeks in small groups. In the reading sessions, the same two stories were read each time, so that both stories were read three times in total. Finally, the posttest was administered in a separate session approximately one week after completion of the reading intervention (*M* = 7.93 days, *SD* = 2.46).

To avoid order and story effects, both the order of the two stories as well as story assignment to the within-subjects condition in the question conditions (i.e., question placement) and to the within-subjects condition in the just-reading control group (i.e., word repetition) were randomized and counterbalanced. Thus, in the question conditions, in each session one of the two stories was read with questions following the story, the other story with questions interspersed within the story. In the just-reading control group, in each session one of the two stories was read with repetition of target-words, the other story without repetitions.

Finally, in the question conditions, narrators were instructed to encourage the children to answer and discuss the questions, but they did not provide any factual feedback or elaborate the children's answers. For instance, if a child asked for the meaning of a word, the experimenters were instructed to redirect the question into the group ("That is an interesting question. Has anyone got an idea?").

The project was approved by the ethical review committee of our Department and all participants were treated in accordance with the ethical principles of the American Psychological Association.

Stimuli and Experimental Conditions

Stories. Two fiction stories were selected from children's books suited for preschool children. We aligned the stories to a roughly comparable length (837 and 999 words). Each story contained six different target words (four nouns and two verbs) and three different control words (two nouns and one verb), resulting in 12 different target words (eight nouns and four verbs) and six different control words (four nouns and two verbs).

In the question conditions, the target words were accompanied by questions, whereas the control words were presented without questions. In the just-reading condition, none of the words was highlighted by a question. Target and control words appeared only once per story and following procedures in relevant literature (e.g., Lenhart et al., 2019)—were selected to have a low frequency of occurrence both in everyday language as well as in children's books, as determined by the occurrence in lexical databases DWDS (Geyken, 2007; Heister et al., 2011) and ChildLex (Schroeder, Würzner, Heister, Geyken, & Kliegl, 2015). Accordingly, target and control words were expected to be largely unfamiliar to the participating children. Examples are *Sud* ("broth"), *Hauer* ("fang"), or *sinnieren* (a rarely used word which means "to muse").

Target and control words did not differ regarding their frequency of occurrence in both lexical databases (lemma frequency per million tokens), ChildLex: $M_{\text{target words}} = 11.08$, $SD_{\text{target words}} = 14.09$, $M_{\text{control words}} = 9.33$, $SD_{\text{control words}} = 10.95$, $t(16) = 0.27$, $p = .794$; DWDS: $M_{\text{target words}} = 119.92$, $SD_{\text{target words}} = 149.52$, $M_{\text{control words}} = 95.17$, $SD_{\text{control words}} = 39.82$, $t(16) = 0.39$, $p = .700$. Finally, to increase ecological validity, five pictures were added for each story, with three pictures depicting target words, one picture comprising a control word, and another picture depicting the main protagonist of the stories.

Questions. In the question conditions, each target word was accompanied by a question, resulting in six questions per story. Depending on the story's question placement condition, the questions were placed either directly after the sentence that contained the target word or after the story was read. Questions differed according to demand-level conditions (low-demand, high-demand, and scaffolding-like). In line with previous studies (e.g., Blewitt et al., 2009; Justice, 2002; Lenhart et al., 2019; B. A. Walsh et al., 2016), low-demand questions focused on perceptual descriptions and repetitions of given information (e.g., "Where does the lion lurk?" or "What does the lion's mane look like?"), whereas high-demand questions required the child to make inferences (e.g., "Why is the warthog afraid of the lion?" or "Why does the warthog prefer to bathe alone?"). The scaffolding-like condition was operationalized through gradually increasing demand-level of questions: In the first reading session, only low-demand questions were asked, in the second reading session, half of the target words were highlighted by low-demand, the other half by high-demand questions, and finally, in the third reading session, only high-demand questions were asked.

The experimenters were instructed to initiate and to sustain children's discussion following the questions, but they were not allowed to provide any content-related input or elaborate on the children's answers. An additional observer recorded the time of the reading sessions and of the group discussions, and made sure that the narrators were adhering to the protocol correctly. No violations of treatment fidelity were reported.

Target-word repetition. To check if enhanced target-word learning in the question groups might only be an effect of target-word repetition within the question, we included target-word repetition (one occurrence vs. repetition) as a within-subjects factor in the just-reading control group. To do so, a short sentence that included a repetition of the target word was added directly after each sentence that contained a target word in the story. Importantly, however, the additional sentence provided no additional information on the target word. Consequently, each child in the just-reading control group heard one “normal” story and one story that included target-word repetitions. For each group, the stories were randomly assigned to the conditions.

Measures

All materials, including the parental questionnaire and the different language tests were in German.

Demographics. All parents completed a questionnaire regarding their own and their children’s country of birth, languages spoken at home and the parents’ highest educational qualification.

Tests of target- and control-word acquisition. To assess different facets of novel word acquisition, a test of word recognition (receptive word knowledge) and a word definition task (expressive word knowledge) were constructed. In each test, a maximum score of 18 was possible, whereby 12 was the maximum for target words and six for control words.

Word recognition task. The word recognition task was used to assess shallower, receptive novel word acquisition and followed a format similar to the *Peabody Picture Vocabulary Test – Fourth Edition (PPVT-4*; Dunn & Dunn, 2007*)*. Here, children had to indicate, from one of four pictures, the picture that best matched a spoken word. Similar tasks have been widely used in shared-reading research (e.g., Blewitt et al., 2009; Lenhart et al., 2019; Vaahtoranta, Suggate, Jachmann, Lenhart, & Lenhard, 2018; B. A. Walsh et al., 2016). Cronbach’s α was .70 at posttest.

Word definition task. In accord with shared-reading research (e.g., Blewitt et al., 2009; Lenhart et al., 2019; B. A. Walsh et al., 2016), a word definition task was used to assess deeper, expressive word knowledge. The task required the children to define the target and control words. They were asked: “Can you explain what XY means?” followed by another prompt: “Do you know anything else about a XY?” Following recent scoring suggestions (Hoffman et al., 2014), we used a three-point scale. One point was awarded if a correct synonym or a complete explanation was provided. Partial explanations (e.g., naming superordinates, properties, or associated concepts) received 0.5 points. Incorrect answers were awarded zero points. All answers were coded independently by two raters. Interrater reliability was assessed using a two-way mixed, average-measures intraclass correlation (ICC) (McGraw & Wong, 1996). For target words ICC (consistency) was .95 at the pretest and .98 at the posttest, for control words ICC (consistency) was .75 at the pretest and .97 at the posttest, indicating acceptable agreement (Cicchetti, 1994). In the case of disagreement between the raters, a third rater rated the respective items independently and the results were discussed until agreement was reached. The final ratings resulted in scores of .00, .50, or 1.00 per word definition, with a minimum of zero und a maximum of 18 for the total scale. Cronbach’s α was .80 at posttest.

Language covariates. We included several language measures, namely expressive and receptive vocabulary, phonological working memory, speech comprehension, and grammatical knowledge to check for equivalence of experimental groups, and to examine potential aptitude-treatment interaction effects.

Receptive and expressive vocabulary. To test for a potential aptitude-treatment interaction in novel word learning from shared-reading, we included measures of receptive and expressive vocabulary knowledge. For the assessment of children’s receptive vocabulary, a German adaption of the *Peabody Picture Vocabulary Test – Fourth Edition* (*PPVT-4*; Lenhard, Lenhard, Segerer, & Suggate, 2015) was used. Split-half reliability (odd-even split of administered items) was .90. To measure children’s expressive vocabulary, the subtest Expressive Vocabulary of the German version of the *Wechsler Preschool and Primary Scale of Intelligence – Third Edition* (*WPPSI-III*; Petermann, 2011) was employed. Cronbach's α was .79. As both tests were highly correlated (r = .69), the results of both tests were z-standardized and averaged, resulting in a single vocabulary score.

Phonological working memory. The subtests *Phonological Working Memory for Non-Words* and *Memory for Sentences* of the *Speech Development Test for Three- to Five-Year-Old Children* (*SETK 3-5*; Grimm, 2010), in which children are required to repeat non-words or sentences, were administered to measure phonological working memory. Cronbach's α was .74 for *Phonological Working Memory for Non-Words* and .87 for *Memory for Sentences*. Both tests were moderately correlated (r = .48).

Speech comprehension. The subtest *Sentence Comprehension* of the *SETK 3-5* (Grimm, 2010) was used to assess speech comprehension. Here, the children had to manipulate a given set of objects according to spoken instructions. Cronbach's α was .62.

Grammatical knowledge. The subtest *Morphological Rule-Making* of the *SETK 3-5* (Grimm, 2010), in which the children had to deduce correct pluralization for a set of words, was employed to assess grammatical knowledge. Cronbach's α was .71.

Data Analysis

Data preparation and analyses were conducted using IBM SPSS 23 and R (version 3.5.2; R Core Team, 2018). To check for equivalence of experimental groups at the pretest we conducted one-way ANOVAs.

Word learning was modelled at the item level, with participants and target/ control words being included as subject X item crossed random effects (Baayen, Davidson, & Bates, 2008; Quené & van den Bergh, 2008). In addition, if indicated by the data structure, we nested the subjects within reading groups, as children were assigned to small groups during the reading sessions, resulting in the following random structure: (variables | reading group) + (variables | reading group: subject) + (variables | item).

Prior to analyses, continuous variables were grand mean centered and categorical independent variables were effect-coded, with single variable estimates representing main effects and interaction estimates denoting interaction effects (Cohen, Cohen, West, & Aiken, 2003). As the word recognition task resulted in binary outcomes, namely 0 or 1 point per word, we used the *glmer*-function of the R package *lme4* (version 1.1-19; Bates, Maechler, Bolker, & Walker, 2015). For the analysis of the word definition task that was scored on an ordinal scale, we used the R package *ordinal* (version 2018.8-25; Christensen, 2018). Contrasts comparing individual groups were computed using the R package *emmeans* (version 1.3.2; Lenth, 2018). Tables for mixed models were created using the R package *sjPlot* (version 2.6.2; Lüdecke, 2018). Effect sizes for both types of mixed-model analyses are reported as *Odds Ratio* (*OR*).

For random structure model simplification, we followed guidelines provided by Bates, Kliegl, Vasishth, and Baayen (2015). In a first step, we used principal component analysis of the random effect structure (provided in the *RePsychLing* package, version 0.0.4; Baayen, Bates, Kliegl, & Vasishth, 2015) to identify the number of variance components supported by the data (the procedure is not available for the *ordinal* package). In a second step, we compared the goodness of fit with likelihood ratio tests and AIC-values, starting with dropping the highest order interaction term. According to

Bates, Kliegl, et al. (2015), the final models represent the optimal linear mixed models for the data.

Missing data. A small proportion of the target- and control-word items as well as of the phonological working memory test was missing (2% of the phonological working memory scores, 2% of the word recognition items, and 1% of the word definition items). Imputation was not used because mixed models have the advantage that they are very robust against small numbers of missing data (Quené & van den Bergh, 2008; Twisk, de Boer, de Vente, & Heymans, 2013).

Results

Table 1 provides an overview of the descriptive statistics of sex composition, age, language tests, and measures of target- and control-word acquisition in each between-subjects condition. The groups did not differ regarding sex composition, age or any of the language test variables ($p > .05$) – so none of the variables was included as a co-variate in the following analyses.

Table 1

Descriptive Statistics for the Just-Reading Control and the Question groups

	Just-reading (*n* = 24) M (SD)	**Low-demand (*n* = 23)** M (SD)	**High-demand (*n* = 21)** M (SD)	**Scaffolding-like (*n* = 23)** M (SD)
Sex (female)	58%	48%	57%	48%
Age (months)	62.63 (6.09)	62.35 (8.30)	65.00 (6.74)	66.74 (8.58)
Language tests				
Receptive vocabulary (PPVT-4)	114.67 (21.11)	113.39 (28.48)	117.86 (18.23)	121.43 (26.99)
Expressive Vocabulary (WPPSI-III)	21.08 (2.93)	20.61 (3.81)	19.86 (3.48)	20.61 (3.10)
Working Memory for Non-Words (SETK 3-5)	11.25 (3.25)	12.55 (2.46)[a]	12.67 (4.02)	12.68 (3.55)[a]
Memory for Sentences (SETK 3-5)	100.88 (15.14)	100.82 (15.39)[a]	94.35 (20.24)[b]	94.82 (18.54)[a]
Speech Comprehension (SETK 3-5)	12.21 (1.64)	11.82 (1.92)[a]	11.81 (2.29)	11.30 (2.48)
Morphological Rule-Making (SETK 3-5)	25.29 (6.07)	23.27 (4.64)[a]	24.05 (5.34)	24.36 (4.74)[a]
Target-word tests				
Word recognition (pretest)	6.25 (2.27)	5.48 (2.00)	5.24 (2.07)	5.32 (1.78)[a]
Word recognition (posttest)	7.38 (2.46)	8.32 (2.25)[a]	7.76 (2.43)	7.83 (2.79)
Word definition (pretest)	0.83 (0.95)	0.52 (0.61)	0.79 (0.87)	0.57 (0.93)
Word definition (posttest)	2.46 (1.84)	3.23 (2.79)[a]	2.74 (2.52)	2.35 (1.95)
Control-word tests				
Word recognition (pretest)	1.46 (1.14)	1.61 (1.37)	1.57 (1.12)	1.73 (1.08)[a]
Word recognition (posttest)	2.08 (1.47)	2.32 (1.36)[a]	2.05 (1.20)	2.13 (1.52)

Word definition (pretest)	0.08 (0.24)	0.30 (0.49)	0.17 (0.37)	0.04 (0.21)
Word definition (posttest)	0.35 (0.56)	0.48 (0.57)[a]	0.31 (0.51)	0.33 (0.61)
Story session features				
Story duration (seconds)	458.90 (30.63)	565.23 (42.86)	566.51 (76.53)	623.39 (72.60)
Discussion duration per question (seconds)	-	17.89 (0.68)	24.09 (3.74)	29.14 (11.27)

Note. For language covariates, raw scores are reported. Target-word tests: maximum score = 12. Control-word tests: maximum score = 6. PPVT-4 = Peabody Picture Vocabulary Test 4; WPPSI-III = Wechsler Preschool and Primary Scale of Intelligence-III; SETK 3-5 = Speech Development Test for Three- to Five-Year-Old Children. [a] n = 22; [b] n = 20.

Regarding story duration and question discussion time (see Table 1), the between-subjects conditions differed on shared-reading duration per story, $F(3, 87) = 32.79$, $p < .001$, $\eta^2_p = .531$. As to be expected, story sessions were shortest in the just-reading control condition. Sessions were longest in the scaffolding-like condition, whereas low- and high-demand conditions did not differ. These differences seem to reflect mainly the time children used to answer and discuss the questions that were asked in the question conditions. Here, we also found differences between conditions, $F(2, 64) = 15.17$, $p < .001$, $\eta^2_p = .322$. Discussion following the low-demand questions was significantly shorter than following high-demand questions. Both in turn were shorter than the discussion in the scaffolding-like condition. However, as discussion time was not significantly related to target-word learning across question conditions, word recognition task: $r(62) = -.04$, $p = .735$, word definition task: $r(63) = .06$, $p = .630$, we did not include it as a covariate in subsequent analyses.

Examining the Effect of Questions

Comparing the just-reading control group and the question groups regarding target-word and control-word learning. Our first hypothesis states that children's learning for target words should be greater in the question conditions compared to the just-reading control group (Hypothesis 1a), but that there should be no differences regarding control words (Hypothesis 1b). To examine this hypothesis, we calculated mixed models with time (pre- vs. posttest) and group (just-reading vs. low-demand vs. high-demand vs. scaffolding-like group) as fixed effects. Analyses of the random structures indicated that for control words (Model 2 and 4), models including only random intercepts for subject (without nesting within reading groups) and item were appropriate. For target words in both tasks (Model 1 and 3), the nesting of subjects within reading groups had to be included, and a random item slope for time had to be added for the definition task.

Tables 2 and 3 display results for target-word learning in the word recognition task (Model 1) and the word definition task (Model 3), respectively. As expected, in both tasks there was a significant main effect of time (word recognition: *OR* = 1.66; word definition: *OR* = 3.86), that was modified by an interaction between time and group, with the just-reading control group (word recognition: *OR* = 0.78; word definition: *OR* = 0.72) and the low-demand group (word definition: *OR* = 1.42) deviating significantly from mean vocabulary acquisition (see Figure 2a and 2b).

Contrasts indicate that significant target-word learning occurred from pre- to posttest in each of the groups on word recognition (control group: log-odds = 0.51, *SE* = 0.20, z = 2.63, p = .008; low-demand: log-odds = 1.23, *SE* = 0.21, z = 5.91, p < .001; high-demand: log-odds = 1.14, *SE* = 0.21, z = 5.35, p < .001; scaffolding-like: log-odds = 1.18, *SE* = 0.21, z = 5.70, p < .001) and on word definition (control group: log-odds = -2.04, *SE* = 0.38, z = 5.40, p < .001; low-demand: log-odds = 3.40, *SE* = 0.43, z = 7.84, p < .001; high-demand: log-odds = 2.54, *SE* = 0.41, z = 6.21, p < .001; scaffolding-like: log-odds = 2.82, *SE* = 0.44, z = 6.46, p < .001).

In line with Hypothesis 1a, children in the question groups displayed greater target-word gains than their peers in the just-reading control group regarding word recognition (low-demand: log-odds = 0.71, *SE* = 0.28, z = 2.52, p = .012; high-demand: log-odds = 0.62, *SE* = 0.29, z = 2.17, p = .030; scaffolding-like: log-odds = 0.67, *SE* = 0.28, z = 2.37, p = .018). For word definitions, a similar pattern emerged. However, only for the low-demand and the scaffolding-like group the difference was significant or respectively marginally significant (log-odds = 1.37, *SE* = 0.46, z = 2.98, p = .003; log-odds = 0.79, *SE* = 0.46, *z* = 1.71, *p* = .087), whereas for high-demand there was only a superiority on the descriptive level (log-odds = 0.50, *SE* = 0.44, *z* = 1.14, *p* = .256).

Regarding control-word learning, we also found a significant main effect of time for both the word recognition (Model 2; see Table 2) and the word definition task (Model 4; see Table 3) (word recognition: *OR* = 1.30; word definition: *OR* = 2.14). In accordance with Hypothesis 1b, children in the question groups did not differ from their peers in the just-reading control group on control-word recognition (low-demand: log-odds = 0.07, *SE* = 0.41, *z* = 0.17, *p* = .868; high-demand: log-odds = -0.19, *SE* = 0.41, *z* = -0.45, *p* = .655; scaffolding-like: log-odds = -0.17, *SE* = 0.41, *z* = -0.42, *p* = .677; see Figure 2c) or on control-word definitions (low-demand: log-odds = -1.17, *SE* = 0.86, *z* = -1.35, *p* = .177; high-demand: log-odds = -0.73, *SE* = 0.98, *z* = -0.74, *p* = .458; scaffolding-like: log-odds = 0.91, *SE* = 1.31, *z* = 0.69, *p* = .488; see Figure 2d).

Table 2

Effects of Questions on Target- and Control-Word Acquisition in the Word Recognition Task

	Model 1 (Target Words)				**Model 2 (Control Words)**			
Fixed Effects	*Odds Ratios*	*SE*	*95%-CI*	*p*	*Odds Ratios*	*SE*	*95%-CI*	*p*
(Intercept)	1.39	0.33	0.72 – 2.68	0.324	0.36	0.42	0.16 – 0.82	0.015
Time (posttest)	1.66	0.05	1.50 – 1.84	<0.001	1.30	0.07	1.12 – 1.50	<0.001
Group (just-reading)	1.04	0.24	0.65 – 1.66	0.867	0.89	0.17	0.64 – 1.24	0.494
Group (low-demand)	1.03	0.24	0.64 – 1.65	0.911	1.05	0.17	0.75 – 1.47	0.781
Group (high-demand)	0.95	0.24	0.59 – 1.53	0.822	0.99	0.18	0.70 – 1.40	0.966
Time X Group (just-reading)	0.78	0.09	0.66 – 0.92	0.003	1.04	0.12	0.81 – 1.32	0.772
Time X Group (low-demand)	1.11	0.09	0.93 – 1.32	0.236	1.07	0.13	0.84 – 1.37	0.580
Time X Group (high-demand)	1.06	0.09	0.89 – 1.27	0.500	0.94	0.13	0.73 – 1.22	0.660
Random Effects								
Residual variance	3.29				3.29			
Intercept variance$_{\text{Reading Group}}$	0.18							
Intercept variance$_{\text{Reading Group:Subject}}$	0.50				0.40			
Intercept variance$_{\text{Item}}$	1.10				0.98			

$ICC_{Reading\ Group}$	.04	
$ICC_{Reading\ Group:Subject}$	.10	.09
ICC_{Item}	.22	.21
Observations	2149	1067
Marginal R^2 / Conditional R^2	.052 / .385	.016 / .307

Notes: Models were calculated using the *glmer* function of the *lme4* package (link function = logit). All factors were effect-coded (Time: pretest = -1; Group: scaffolding-like group = -1).

Table 3

Effects of Questions on Target- and Control-Word Acquisition in the Word Definition Task

	Model 3 (Target-Words)				**Model 4 (Control-Words)**			
Fixed Effects	*Odds Ratios*	*SE*	*95%-CI*	*p*	*Odds Ratios*	*SE*	*95%-CI*	*p*
(Intercept 0|0.5)	18.21	0.47	7.28 – 45.52	<0.001	103.10	0.83	20.42 – 520.56	<0.001
(Intercept 0.5|1)	45.54	0.47	17.97 – 115.36	<0.001	258.66	0.85	48.55 – 1378.03	<0.001
Time(posttest)	3.86	0.15	2.88 – 5.16	<0.001	2.14	0.20	1.45 – 3.15	<0.001
Group (just-reading)	1.18	0.34	0.61 – 2.29	0.619	0.87	0.40	0.40 – 1.90	0.722
Group (low-demand)	0.89	0.35	0.45 – 1.78	0.750	2.34	0.36	1.16 – 4.71	0.018
Group (high-demand)	1.14	0.35	0.57 – 2.27	0.715	1.11	0.40	0.51 – 2.41	0.796
Time X Group (just-reading)	0.72	0.13	0.55 – 0.93	0.013	1.13	0.32	0.61 – 2.10	0.698
Time X Group (low-demand)	1.42	0.15	1.06 – 1.90	0.019	0.63	0.26	0.38 – 1.05	0.078
Time X Group (high-demand)	0.92	0.14	0.70 – 1.22	0.574	0.79	0.31	0.43 – 1.44	0.435
Random Effects								
Residual variance			3.29				3.29	
Intercept variance$_{\text{Reading Group}}$			0.14					

Intercept variance$_{Reading\ Group:Subject}$	2.05	1.41
Intercept variance$_{Item}$	1.50	2.83
Time slope variance$_{Item}$	0.31	
ICC$_{Reading\ Group}$		
ICC$_{Reading\ Group:Subject}$		0.19
ICC$_{Item}$		0.38
Observations	2167	1081

Notes: Models were calculated using the *clmm* function of the *ordinal* package (link function = logit; threshold = flexible). All factors were effect-coded (Time: pretest = -1; Group: scaffolding-like group = -1). ICCs are not provided for models that comprise random slopes.

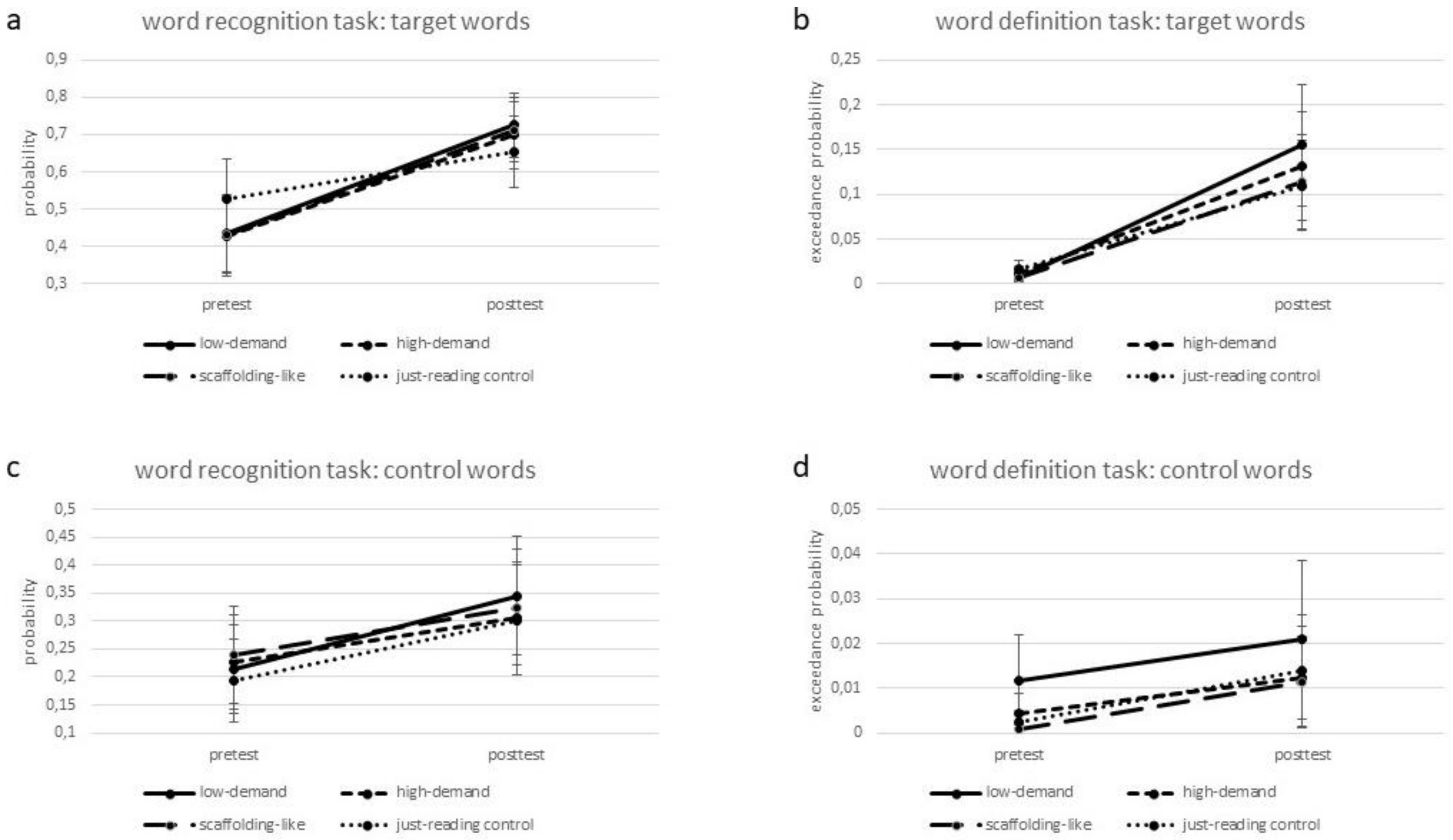

Figure 2. Target- and control-word learning in the question and the just-reading control groups.

Note: Error Bars = SE. For word recognition, the dependent variable is the probability to recognize a specific word. For word definition, the dependent variable is the exceedance probability at each threshold.

Examining effects of mere target-word repetition in the just-reading control group. To rule out the methodological concern that the beneficial effect of questions on target-word learning could simply be a side effect of target-word repetition within the question (see for example Blewitt et al., 2009), we analyzed the effect of target-word repetition in the just-reading control group. We calculated mixed models with time (pre- vs. posttest) and target-word repetition (without vs. with repetition) as fixed effects for target words in the word recognition and the word definition task. Analyses of the random structures indicated that for word definition (Model 6) a model including only random intercepts for subject (without nesting into reading groups) and item was sufficient. For word recognition (Model 5), random intercepts for reading groups, subjects (nested within reading groups), and item, as well as a random subject slope for word repetition were appropriate.

Although both target-word tests (Model 5 and 6; see Table 4) indicated significant target-word learning in the just-reading control group (word recognition: *OR* = 1.32; word definition: *OR* = 2.33), a single repetition of the target words per story did not significantly enhance learning (word recognition: *OR* = 1.09; word definition: *OR* = 1.28). This lends further support to Hypothesis 1a, because it indicates that the effect observed for the use of questions was not just an effect of mere repetition of target words.

Table 4

Effects of Target-Word Repetition on Word Recognition and Definition Learning in the Just-Reading Control Group

	Model 5 (Word Recognition - Target-Words)				**Model 6 (Word Definition - Target-Words)**			
Fixed Effects	*Odds Ratios*	*SE*	*95%-CI*	*p*	*Odds Ratios*	*SE*	*95%-CI*	*p*
(Intercept)	1.47	0.48	0.57 – 3.79	0.421				
(Intercept 0|0.5)					14.62	0.55	5.02 – 42.57	<0.001
(Intercept 0.5|1)					31.87	0.56	10.56 – 96.19	<0.001
Time(posttest)	1.32	0.10	1.09 – 1.61	0.006	2.33	0.15	1.72 – 3.14	<0.001
Target-Word Repetition (repetition)	1.09	0.14	0.83 – 1.43	0.536	1.28	0.15	0.96 – 1.71	0.096
Time X Target-Word Repetition	1.13	0.10	0.93 – 1.38	0.219	1.27	0.15	0.95 – 1.69	0.105
Random Effects								
Residual variance	3.29				3.29			
Intercept variance$_{\text{Reading Group}}$	0.31							
Intercept variance$_{\text{Reading Group:Subject}}$	0.56				1.86			
Intercept variance$_{\text{Item}}$	1.43				1.94			
Repetition slope variance$_{\text{Reading Group:Subject}}$	0.21							

$ICC_{Reading\ Group}$		
$ICC_{Reading\ Group:Subject}$		.26
ICC_{Item}		.27
Observations	576	576
Marginal R^2 / Conditional R^2	.017 / .442	

Notes: The model for word recognition was calculated using the *glmer* function of the *lme4* package (link function = logit). The model for word definition was calculated using the *clmm* function of the *ordinal* package (link function = logit; threshold = flexible). All factors were effect-coded (Time: pretest = -1; Word Repetition: no repetition = -1). ICCs are not provided for models that comprise random slopes.

Effects of Question Demand-Level and Placement

To test predictions made by different theories regarding question demand-level (H2a, H2b) and placement (H3a, H3b) as well as their interaction (H4a), we calculated mixed models for target-word learning with time (pre- vs. posttest), question placement (within vs. after the story), question demand-level (low vs. high vs. scaffolding-like demand) as fixed effects for word recognition and word definition. Analysis of the random structure indicated that for word recognition (Model 7) including random intercepts for subject (nested within reading groups) and item were appropriate. For the word definition task (Model 8), random intercepts for item and subject (without the nesting within reading groups) were sufficient.

Table 5 displays the results for word recognition that was used as an indicator for surface word learning and the results for word definition that was used as an indicator for deeper word learning. Although there were significant learning effects across conditions (word recognition: $OR = 1.80$; word definition: $OR = 3.77$), neither question demand-level nor placement nor their interaction had any significant effect on target-word acquisition in the word recognition or the word definition task. Consequently, concerning question demand-level, these results do not provide evidence for the high-demand (Hypothesis 2a) or the scaffolding theory (Hypothesis 2b) and regarding question placement, our data supported neither the non-interruptive (Hypothesis 3a) nor the interruptive approach (Hypothesis 3b). Moreover, our data do also not support Hypothesis 4a that assumes best effects for high-demand questions asked after the story.

Table 5

Effects of Question Demand-Level and Placement on Target-Word Learning in the Word Recognition and the Word Definition Task

	Model 7 (Word Recognition - Target-Words)				**Model 8 (Word Definition - Target-Words)**			
Fixed Effects	*Odds Ratios*	*SE*	*95%-CI*	*p*	*Odds Ratios*	*SE*	*95%-CI*	*p*
(Intercept)	1.37	0.33	0.71 – 2.63	0.345				
(Intercept: 0|0.5)					17.14	0.45	7.09 – 41.44	<0.001
(Intercept: 0.5|1)					45.39	0.46	18.40 – 112.01	<0.001
Time (posttest)	1.80	0.06	1.59 – 2.03	<0.001	3.77	0.11	3.05 – 4.67	<0.001
Demand-Level (low-demand)	1.05	0.22	0.68 – 1.60	0.837	0.93	0.32	0.50 – 1.74	0.825
Demand-Level (high-demand)	0.96	0.22	0.62 – 1.47	0.846	1.24	0.32	0.66 – 2.30	0.504
Placement (after the story)	1.07	0.06	0.95 – 1.20	0.285	0.93	0.10	0.76 – 1.13	0.478
Time X Demand-Level (low-demand)	1.02	0.08	0.87 – 1.21	0.801	1.28	0.14	0.97 – 1.70	0.080
Time X Demand-Level (high-demand)	0.98	0.09	0.83 – 1.15	0.784	0.82	0.14	0.63 – 1.08	0.161
Time X Placement	1.04	0.06	0.92 – 1.17	0.527	1.18	0.10	0.97 – 1.44	0.093
Demand-Level (low-demand) X Placement	1.05	0.08	0.89 – 1.23	0.596	0.98	0.15	0.74 – 1.31	0.914

Demand-Level (high-demand) X Placement	1.01	0.09	0.85 – 1.19	0.904	1.12	0.14	0.85 – 1.47	0.435
Time X Demand-Level (low-demand) X Placement	0.92	0.08	0.78 – 1.08	0.294	1.04	0.14	0.78 – 1.37	0.803
Time X Demand-Level (high-demand) X Placement	1.10	0.09	0.93 – 1.30	0.279	0.95	0.14	0.72 – 1.24	0.695
Random Effects								
Residual variance		3.29				3.29		
Intercept variance$_{\text{Reading Group}}$		0.15						
Intercept variance$_{\text{Reading Group:Subject}}$		0.50				2.46		
Intercept variance$_{\text{Item}}$		1.04				1.70		
ICC$_{\text{Reading Group}}$		.03						
ICC$_{\text{Reading Group:Subject}}$		.10				.33		
ICC$_{\text{Item}}$		.21				.23		
Observations		1573				1591		
Marginal R^2 / Conditional R^2		.067 / .383						

Notes: The model for word recognition was calculated using the *glmer* function of the *lme4* package (link function = logit). The model for word definition was calculated using the *clmm* function of the *ordinal* package (link function = logit; threshold = flexible). All factors were effect-coded (Time: pretest = -1; Demand-Level: scaffolding-like = -1; Placement: within the story = -1).

Aptitude-Treatment Interaction with Question Demand-Level and Placement

We also examined the prediction of the aptitude-treatment theory regarding question demand-level (Hypothesis 2c), question placement (Hypothesis 3c), and their combination (Hypothesis 4b). To test respective predictions, we calculated mixed models for target-word learning in the low- and high-demand groups, dropping the scaffolding-like demand condition. Time (pre- vs. posttest), demand-level (low- vs. high-demand), question placement (within vs. after the story) and general vocabulary or phonological working memory were included as fixed effects. Analyses of the random structures (Model 9, 10, 11 and 12) indicated that including random intercepts for subject (without nesting within reading groups) and item was appropriate. In addition, for word recognition (Model 9), a random item slope for general vocabulary knowledge had to be included.

Aptitude-treatment interaction with general vocabulary knowledge. Although general vocabulary knowledge was significantly positively related to target-word knowledge (word recognition: *OR* = 2.43; word definition: *OR* = 5.49) and learning (word recognition: *OR* = 1.20; word definition: *OR* = 1.48), it did not significantly interact with question demand-level (low- vs. high-demand), placement, or their combination (see Table 6). Consequently, our results do not support either Hypothesis 2c or Hypothesis 4b.

Table 6

Effects of Question Demand-Level, Placement and General Vocabulary Knowledge on Target-Word Learning in the Word Recognition and Definition Task

	Model 9 (Word Recognition - Target-Words)				**Model 10 (Word Definition - Target-Words)**			
Fixed Effects	*Odds Ratios*	*SE*	*95%-CI*	*p*	*Odds Ratios*	*SE*	*95%-CI*	*p*
(Intercept)	1.43	0.33	0.75 – 2.74	0.282				
(Intercept 0|0.5)					16.84	0.46	6.87 – 41.26	<0.001
(Intercept 0.5|1)					48.61	0.47	19.20 – 123.06	<0.001
Time (posttest)	1.84	0.08	1.58 – 2.14	<0.001	3.23	0.16	2.37 – 4.40	<0.001
Demand-Level (high-demand)	0.97	0.10	0.79 – 1.18	0.736	1.23	0.20	0.84 – 1.81	0.286
Placement (after the story)	1.09	0.08	0.94 – 1.27	0.236	1.08	0.15	0.80 – 1.46	0.604
General Vocabulary Knowledge	2.43	0.18	1.72 – 3.45	<0.001	5.49	0.26	3.33 – 9.06	<0.001
Time X Demand-Level	0.99	0.08	0.85 – 1.15	0.873	0.88	0.15	0.65 – 1.18	0.390
Time X Placement	1.04	0.08	0.89 – 1.20	0.646	1.00	0.15	0.74 – 1.35	0.993
Demand-Level X Placement	0.97	0.08	0.83 – 1.14	0.703	1.17	0.15	0.86 – 1.58	0.321

Time X General Vocabulary Knowledge	1.20	0.09	1.00 – 1.44	0.048	1.48	0.18	1.05 – 2.10	0.027
Demand-Level X General Vocabulary Knowledge	1.05	0.12	0.83 – 1.32	0.684	1.03	0.24	0.64 – 1.67	0.888
Placement X General Vocabulary Knowledge	1.01	0.09	0.84 – 1.21	0.930	0.86	0.18	0.60 – 1.22	0.394
Time X Demand-Level X Placement	1.09	0.08	0.94 – 1.27	0.235	0.95	0.15	0.71 – 1.28	0.742
Time X Demand-Level X General Vocabulary Knowledge	1.01	0.09	0.84 – 1.20	0.956	0.93	0.18	0.66 – 1.31	0.679
Time X Placement X General Vocabulary Knowledge	0.95	0.09	0.79 – 1.13	0.560	1.35	0.18	0.96 – 1.92	0.088
Demand-Level X Placement X General Vocabulary Knowledge	0.96	0.10	0.79 – 1.18	0.728	0.83	0.18	0.58 – 1.18	0.291
Time X Demand-Level X Placement X General Vocabulary Knowledge	0.90	0.09	0.75 – 1.07	0.238	1.06	0.18	0.75 – 1.49	0.757
Random Effects								
Residual variance			3.29				3.29	
Intercept variance$_{\text{Reading Group}}$								
Intercept variance$_{\text{Reading Group:Subject}}$			0.19				0.53	
Intercept variance$_{\text{Item}}$			1.22				1.86	

General Vocabulary Knowledge slope $\text{variance}_{\text{Item}}$	0.20	
$\text{ICC}_{\text{Reading Group}}$		
$\text{ICC}_{\text{Reading Group:Subject}}$		.09
ICC_{Item}		.33
Observations	1040	1044
Marginal R^2 / Conditional R^2	.192 / .434	

Notes: The model for word recognition was calculated using the *glmer* function of the *lme4* package (link function = logit). The model for word definition was calculated using the *clmm* function of the *ordinal* package (link function = logit; threshold = flexible). All factors were effect-coded (Time: pretest = -1; Demand-Level: low-demand = -1; Placement: within the story = -1). ICCs are not reported for models that comprise random slopes.

Table 7

Effects of Question Demand-Level, Placement, and Phonological Working Memory on Target-Word Learning in the Word Recognition and Definition Task

	Model 11 (Word Recognition - Target-Words)				**Model 12 (Word Definition - Target-Words)**			
Fixed Effects	*Odds Ratios*	*SE*	*95%-CI*	*p*	*Odds Ratios*	*SE*	*95%-CI*	*p*
(Intercept)	1.45	0.33	0.76 – 2.75	0.258				
(Intercept 0|0.5)					17.21	0.52	6.26 – 47.33	<0.001
(Intercept 0.5|1)					48.96	0.53	17.28 – 138.73	<0.001
Time (posttest)	1.79	0.08	1.54 – 2.08	<0.001	4.03	0.14	3.08 – 5.28	<0.001
Demand-Level (high-demand)	0.89	0.13	0.69 – 1.16	0.396	1.12	0.31	0.62 – 2.04	0.712
Placement (after the story)	1.10	0.07	0.95 – 1.27	0.216	0.96	0.13	0.75 – 1.23	0.768
Phonological Working Memory	1.08	0.04	0.99 – 1.18	0.078	1.18	0.11	0.95 – 1.47	0.125
Time X Demand-Level	0.99	0.07	0.85 – 1.14	0.872	0.82	0.13	0.64 – 1.05	0.110
Time X Placement	1.04	0.07	0.90 – 1.20	0.581	1.20	0.12	0.94 – 1.54	0.140
Demand-Level X Placement	0.99	0.07	0.85 – 1.14	0.878	1.06	0.13	0.83 – 1.37	0.637

Time X Phonological Working Memory	1.04	0.03	0.99 – 1.09	0.127	1.01	0.04	0.93 – 1.10	0.848
Demand-Level X Phonological Working Memory	1.01	0.04	0.93 – 1.11	0.749	0.95	0.11	0.77 – 1.17	0.624
Placement X Phonological Working Memory	1.00	0.03	0.95 – 1.05	0.915	1.07	0.04	0.98 – 1.17	0.120
Time X Demand-Level X Placement	1.09	0.07	0.95 – 1.26	0.231	0.99	0.12	0.77 – 1.26	0.918
Time X Demand-Level X Phonological Working Memory	0.98	0.03	0.94 – 1.03	0.491	1.00	0.04	0.92 – 1.08	0.940
Time X Placement X Phonological Working Memory	0.99	0.03	0.94 – 1.04	0.714	0.95	0.04	0.87 – 1.04	0.254
Demand-Level X Placement X Phonological Working Memory	0.98	0.03	0.93 – 1.03	0.447	0.97	0.04	0.89 – 1.05	0.429
Time X Demand-Level X Placement X Phonological Working Memory	1.01	0.03	0.96 – 1.06	0.647	1.00	0.04	0.92 – 1.09	0.984
Random Effects								
Residual variance			3.29				3.29	
Intercept variance$_{\text{Reading Group}}$								
Intercept variance$_{\text{Reading Group:Subject}}$			0.50				3.04	
Intercept variance$_{\text{Item}}$			1.08				1.84	
ICC$_{\text{Reading Group}}$								

$ICC_{Reading\ Group:Subject}$	.10	.37
ICC_{Item}	.22	.22
Observations	1028	1032
Marginal R^2 / Conditional R^2	.087 / .383	

Notes: The model for word recognition was calculated using the *glmer* function of the *lme4* package (link function = logit). The model for word definition was calculated using the *clmm* function of the *ordinal* package (link function = logit; threshold = flexible). All factors were effect-coded (Time: pretest = -1; Demand-Level: low-demand = -1; Placement: within the story = -1).

Aptitude-treatment interaction with phonological working memory. Table 7 displays the results for target-word acquisition in the word recognition and definition tasks. Both analyses show that phonological working memory was not significantly related to target-word knowledge or learning and that there was no significant interaction between phonological working memory and question placement, demand-level, or their combination. Thus, the results do not provide evidence for an aptitude-treatment interaction as proposed by Hypothesis 3c.

Discussion

In the current study, we examined multiple research questions. First, we wanted to replicate the finding that asking questions generally has a positive effect on novel word learning from shared-reading. To this end, we contrasted the question groups and the just-reading control group regarding target- and control-word learning. In addition, we included target-word repetition as a within-subjects factor in the just-reading control group, to control for the possibility that effects of asking questions could be due to the mere repetition of target words within the questions. Second, we examined effects of question demand-level and tested predictions made by three different theoretical accounts: the high-demand theory (B. A. Walsh et al., 2016), the scaffolding theory (Blewitt et al., 2009), and the aptitude-treatment interaction theory (Reese & Cox, 1999). Third, we examined the effects of question-placement and tested two widely used approaches, an interruptive and a non-interruptive story delivery (e.g., Brabham & Lynch-Brown, 2002), as well as an aptitude-treatment interaction proposed by Jimenez and Saylor (2017). Finally, we examined a potential interaction between question demand-level and placement as indicated by Dickinson and Smith's study (1994) as well as a potential three-way-interaction with children's language skills as indicated by Reese and Cox's study (1999).

Regarding the first research question, our results indicate that although word learning occurred across conditions, target-word gains were greater in the question conditions compared to the just-reading control group. In contrast, control and question conditions displayed no differences regarding learning of control words that were not accompanied by questions. Thus, these findings support the notion that asking questions is indeed an effective means to increase children's word learning from shared-reading and that our experimental paradigm was effective.

Moreover, we also addressed a critical point discussed by Blewitt et al. (2009), namely that the effect of questions on word learning may be due to the mere repetition of these words within the questions. However, the analysis of the just-reading control group, in which we contrasted target words that were repeated and those that were presented in a normal way (i.e., only once per story), showed that a simple repetition of target words in the stories – without providing additional information about those words – had no significant effect on learning gains. Consequently, the beneficial effect of asking questions on word learning does not seem to be a simple effect of mere word repetition within the questions.

Regarding the second research question, namely effects of question demand-level, we could not find any differences between the low-demand, high-demand, or the scaffolding-like (gradually increasing from low- to high-demand questions) conditions. As this was true for both the so-called surface word learning and the deeper word learning measures, our results support neither the high-demand theory (B. A. Walsh et al., 2016) nor the scaffolding theory (Blewitt et al., 2009). Moreover, in the current study, we could not find any evidence for an interaction between question demand-level (low- vs. high-demand) and children's general vocabulary. Consequently, our results do not support the aptitude-treatment interaction theory (Reese & Cox, 1999). This pattern of results is in accordance with Justice (2002) and Lenhart et al. (2019) who did not find any differences between question conditions.

Concerning the third research question, namely the effects of question placement, we found no difference between questions asked after the story reading and questions interspersed within the story. Thus, in line with most other experimental studies (e.g., Blewitt et al., 2009; Jimenez & Saylor, 2017; Lenhart et al., 2019; B. A. Walsh et al., 2016), our findings support neither the non-interruptive nor the interruptive question placement approach. However, we could also not replicate Jimenez and Saylor's (2017) results that reported an interaction between instruction placement and children's verbal memory. Finally, regarding target-word acquisition, our study did not provide evidence for any meaningful interaction between question placement, demand-level, and children's language skills or memory capacity. As the present study's sample size was larger (e.g., Blewitt et al., 2009; Justice, 2002; Reese & Cox, 1999; B. A. Walsh et al., 2016) or very similar to that in most studies (e.g., Jimenez & Saylor, 2017; Lenhart et al. 2019) and power was sufficient to detect reported effect sizes, our results indicate that there were no meaningful differential effects of question demand-level and placement on vocabulary acquisition during book reading.

Reflections on the Potential Role of Elaborative Adult Feedback

Taken together, the results of our study do not favor any of the proposed theories on question demand-level or on question placement, but at a first glance seem to add to an already confusing research picture. However, as recently proposed by Lenhart et al.

(2019), it may help to reflect on the nature of the shared-reading process that takes place in experimental studies and on the role the narrator performed in those studies.

Typically, in well-controlled experimental designs (e.g., Blewitt et al., 2009; Justice, 2002; Lenhart et al., 2019), narrators are not allowed to participate in answering the questions or to provide any actual feedback. Their only role is to ask a set of predefined questions and to motivate the children to reflect on and to answer the questions. Consequently, the narrator draws attention to the target words by asking questions, but all other potentially beneficial effects of questions have to originate from the child. Although studies using this type of well-controlled experimental design typically find a superiority of question conditions compared to just-reading control groups, most of them do not find differences between question demand-level or placement conditions (e.g., Justice, 2002; Lenhart et al., 2019).

As discussed by Lenhart et al. (2019), the Reese and Cox (1999) study may represent a special case. Although the readers were not allowed to give corrective feedback, the reading of approximately 30 books with three reading sessions per week across six weeks might have given rise to the expectation of being asked similar questions each time. This in turn might have habituated the children to the specific type of questions and encouraged them to demand shared-reading by their parents in the way learned from the numerous reading sessions.

In contrast to those very strict experimental designs, studies permitting the narrator to provide corrective feedback and to actively steer and participate in the discussion following the questions typically report differences between question placement (e.g., Brabham & Lynch-Brown, 2002) or demand-level conditions (e.g., Dickinson & Smith, 1994). As well as drawing children’s attention to specific story aspects, in these studies different types of questions may lead to differential reader-listener interaction. Consequently, potentially beneficial effects of different types of questions do not necessarily have to originate from the child, but they can emerge because of differential interaction between reader and listener.

Narrator-child interaction during shared-reading may have a positive effect on children's language development because children and adults participate in a "ritualized dialogue", in which the adult can elicit attention, ask questions, give feedback, and explain and elaborate the child's utterances (Ninio & Bruner, 1978, p. 1). Although studies report beneficial effects of explicit explanations and elaborations provided by the adult narrator (e.g., Ard & Beverly, 2004; Biemiller & Boote, 2006; Blewitt & Langan, 2016; Silverman, 2007), the specific type of interaction pattern seems to matter, with different interaction patterns producing differential learning gains (e.g., Gonzalez et al., 2014; Haden et al., 1994).

Consequently, asking different types of questions, such as low- vs. high-demand questions, may lead to differential adult input. Dyadic interactions following high-demand questions may cause the narrator to explain or elaborate on a more abstract level compared to low-demand questions where discussion may only target less abstract descriptions and repetitions of displayed story content. To examine this possible pattern, it would be necessary to allow the adult to participate in the interaction following the questions (see for example Blewitt & Langan, 2016). Moreover, besides these differences in interaction quality, it may also be possible that different types of question placement or demand-level lead to differences in interaction quantity, such as time spent on the task or the amount of adult input.

However, the problem with studies in which the reader is free to participate in the discussion, is that they contain a larger degree of uncontrolled adult input, complicating the search for causal links between question types and children's word learning. Therefore, in the current study, we used a controlled small group design to try to bridge the gap between highly controlled experiments that employ one-on-one reading sessions (e.g., Blewitt et al., 2009; Lenhart et al., 2019) and more naturalistic whole-classroom interaction studies (e.g., Brabham & Lynch-Brown, 2002; Dickinson & Smith, 1994). Although similar to highly controlled one-on-one designs, our narrators were not allowed to answer the questions, they were permitted to stimulate and help sustain group discussion in the question groups.

Taken together, findings from well-controlled experiments (e.g., Justice, 2002; Lenhart et al., 2019), less well-controlled studies (e.g., Brabham & Lynch-Brown, 2002; Dickinson & Smith, 1994), and our current study, lead to the tentative conclusion that—at least for preschool children—differences regarding question placement or demand-level might not be sufficient to produce differential learning effects, when questions are not followed by extended meaningful discussion, feedback, or elaboration by the adult narrator. Consequently, the interaction between narrators and children may be the truly relevant aspect for vocabulary acquisition, with question demand-level and placement proving supplemental to this interaction.

Limitations and Directions for Future Research

Although we combined advantages of well-controlled one-on-one designs and more naturalistic studies by using well-controlled small-group reading sessions, our study suffers from several limitations that are in part connected to the study design. First, due to the experimental restrictions on narrator input, we cannot draw inferences potential differential effects of question demand-level and placement in real-world settings. In these settings, adult narrators typically provide feedback and elaborate on children's answers (e.g., Ninio & Bruner, 1978). However, this constraint was necessary to allow comparability with most of the other studies examining these question characteristics (e.g., Blewitt et al., 2009; Justice, 2002; Lenhart et al., 2019; Reese & Cox, 1999; B. A. Walsh et al., 2016) and to allow the examination of causal links with novel word acquisition.

Second, we did not use audio or video recordings of the reading sessions. Although we measured the time of the story readings and the duration of children answering and discussing the questions, we do not have information about the content or dynamics of these small-group interactions. For example, we do not know whether the 1.5 to approximately 3 minutes that children spent answering questions (i.e., approximately 19-28% of the story reading duration) was filled with helpful, irrelevant or even unhelpful information or whether its duration resulted mainly from the adult narrators trying to engage the children in answering the questions. Finally, we only used two different children's stories. The effect of questions and of different types of questions could vary depending on the words to be acquired as well as on the language complexity and the supports (e.g., number of pictures) provided by the stories.

Future studies need to tackle these issues. First and foremost, quantifying and classifying the interaction between children as well as between narrator and children, possibly via audio or video recordings, would extend the findings of the current study. Another possible extension of the current design is to allow the adult narrators to participate in a meaningful manner in the interaction following the questions (see for example Blewitt & Langan, 2016). For instance, this could be done by instructing the narrators to focus their discussion input on descriptive aspects of the story in the low-demand condition and on more abstract, inferential aspects of the story in the high-demand condition. Finally, future work could also try to move beyond word learning and examine effects of different questions and associated interactions on story comprehension, reasoning, and the development of broader socio-cognitive competencies.

Summary and Implications for Practice

To summarize, our study extends previous research and sheds further light on the effects of question placement and demand-level on children's word learning from listening to stories. The current experiment demonstrated not only that story-reading is an essential component of fostering vocabulary development, but that questions—regardless of difficulty and placement—further bolster the educational contribution. Consequently, the presumption that some types of questions might inherently result in better outcomes for preschool children seems to be unjustified, underlining the central role that adults play not only for engaging children through skillful and appropriate questions, but also for providing corrective feedback and for elaborating children's utterances.

Finally, our study underlines the need for critically examining and testing predictions on these question features before they are implemented in any large-scale shared-reading intervention. Although we agree with R. L. Walsh and Hodge's (2018, p. 289–290) conclusion that at present "there appear to be too few experimental studies to allow any generalizable comments about the optimal level of demand, placement of questions, frequency of questioning or group size", we are optimistic that future research can build on this fertile groundwork.

References

Aram, D., Fine, Y., & Ziv, M. (2013). Enhancing parent–child shared book reading interactions: Promoting references to the book's plot and socio-cognitive themes. *Early Childhood Research Quarterly, 28*, 111–122. https://doi.org/10.1016/j.ecresq.2012.03.005

Ard, L. M., & Beverly, B. L. (2004). Preschool word learning during joint book reading: Effect of adult questions and comments. *Communication Disorders Quarterly, 26*, 17–28. https://doi.org/10.1177/15257401040260010101

Baayen, H., Bates, D., Kliegl, R., & Vasishth, S. (2015). RePsychLing: Data sets from psychology and linguistics experiments. R package version 0.0.4. Retrieved from https://github.com/dmbates/RePsychLing

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. Retrieved from https://arxiv.org/abs/1506.04967v1

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48. https://doi/10.18637/jss.v067.i01.

Beck, I. L. & McKeown, M. G. (2007). Increasing young low-income children's oral vocabulary repertoires through rich and focused instruction. *The Elementary School Journal, 107*, 251–271. https://doi/10.1086/511706

Biemiller, A., & Boote, C. (2006). An effective method for building meaning vocabulary in primary grades. *Journal of Educational Psychology*, *98*, 44–62. https://doi.org/10.1037/0022-0663.98.1.44

Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of Educational Psychology*, *93*, 498–520. https://doi.org/10.1037/0022-0663.93.3.498

Blewitt, P., & Langan, R. (2016). Learning words during shared book reading: The role of extratextual talk designed to increase child engagement. *Journal of Experimental Child Psychology*, *150*, 404–410. https://doi.org/10.1016/j.jecp.2016.06.009

Blewitt, P., Rump, K. M., Shealy, S. E., & Cook, S. A. (2009). Shared book reading: When and how questions affect young children's word learning. *Journal of Educational Psychology*, *101*, 294–304. https://doi.org/10.1037/a0013844

Brabham, E. G., & Lynch-Brown, C. (2002). Effects of teachers' reading-aloud styles on vocabulary acquisition and comprehension of students in the early elementary grades. *Journal of Educational Psychology*, *94*, 465–473. https://doi.org/10.1037/0022-0663.94.3.465

Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan, & G. A. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 265–293). New York, NY: Cambridge University Press.

Carey, S. (2010). Beyond fast mapping. *Language Learning and Development*, *6*(3), 184–205. https://doi.org/10.1080/15475441.2010.484379

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development, 15*, 17–29.

Christensen, R. H. B. (2018). ordinal – Regression Models for Ordinal Data. R package version 2018.8-25. Retrieved from http://www.cran.r-project.org/package=ordinal/.

Christian, K., Morrison, F. J., Frazier, J. A., & Massetti, G. (2000). Specificity in the nature and timing of cognitive growth in kindergarten and first grade. *Journal of Cognition and Development*, *1*, 429–448. https://doi.org/10.1207/S15327647JCD0104_04

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Connor, C. M., Morrison, F. J., & Katch, L. E. (2004). Beyond the reading wars: Exploring the effect of child-instruction interactions on growth in early reading. *Scientific Studies of Reading*, *8*, 305–336. https://doi.org/10.1207/s1532799xssr0804_1

Connor, C. M., Piasta, S. B., Fishman, B., Glasney, S., Schatschneider, C., Crowe, E.,. . . Morrison, F. J. (2009). Individualizing student instruction precisely: Effects of child x instruction interactions on first graders' literacy development. *Child Development*, *80,* 77–100. https://doi.org/10.1111/j.1467-8624.2008.01247.x

Cromley, J. G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology*, *99*, 311–325. https://doi.org/10.1037/0022-0663.99.2.311

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology*, *33*, 934–945. https://doi.org/10.1037/0012-1649.33.6.934

Dickinson, D. K., & Smith, M. W. (1994). Long-term effects of preschool teachers' book readings on low-income children's vocabulary and story comprehension. *Reading Research Quarterly*, *29*, 104–122. https://doi.org/10.2307/747807

Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test – Fourth Edition*. Bloomington, MN: Pearson.

Federal Bureau of Statistics (2017). *Bildungsstand der Bevölkerung 2016* [Educational level of the population 2016]. Retrieved from https://www.destatis.de/DE/Publikationen/Thematisch/BildungForschungKultur/Bildungsstand/BildungsstandBevoelkerung5210002167004.pdf?__blob=publicationFile

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science*, *16*, 234–248. https://doi.org/10.1111/desc.12019

Flack, Z. M., Field, A. P., & Horst, J. S. (2018). The effects of shared storybook reading on word learning: A meta-analysis. *Developmental Psychology*, *54*, 1334–1346. https://doi.org/10.1037/dev0000512

Geyken, A. (2007). The DWDS corpus: A reference corpus for the German language of the 20th century. In C. Fellbaum (Ed.), *Idioms and collocations. Corpus-based linguistic and lexicographic studies* (pp. 23–41). London, UK: Continuum.

Gonzalez, J. E., Pollard-Durodola, S., Simmons, D. C., Taylor, A. B., Davis, M. J., Fogarty, M., & Simmons, L. (2014). Enhancing preschool children's vocabulary: Effects of teacher talk before, during and after shared reading. *Early Childhood Research Quarterly, 29,* 214–226. https://doi.org/10.1016/j.ecresq.2013.11.001

Grimm, H. (2010). *Sprachentwicklungstest für drei- bis fünfjährige Kinder (SETK 3–5)* [Speech Development Test For 3- to 5-Year-Old Children] (2nd ed.). Göttingen, Germany: Hogrefe.

Haden, C. A., Reese, E., & Fivush, R. (1996). Mothers' extratextual comments during storybook reading: Stylistic differences over time and across texts. *Discourse Processes, 21,* 135–169. https://doi.org/10.1080/01638539609544953

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Paul H. Brookes.

Hart, B., & Risley, T. R. (2003). The early catastrophe: The 30 million word gap. *American Educator, 27,* 4–9.

Hassinger-Das, B., Ridge, K., Parker, A., Golinkoff, R. M., Hirsh-Pasek, K., & Dickinson, D. K. (2016). Building vocabulary knowledge in preschoolers through shared book reading and gameplay. *Mind, Brain, and Education, 10*, 71–80. https://doi.org/10.1111/mbe.12103

Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., & Kliegl, R. (2011). dlexDB – eine lexikalische Datenbank für die psychologische und linguistische Forschung [dlexDB – A lexical database for the psychological and linguistic research]. *Psychologische Rundschau*, *62*, 10–20. https://doi.org/10.1026/0033-3042/a000029

Hoffman, J. L., Teale, W. H., & Paciga, K. A. (2014). Assessing vocabulary learning in early childhood. *Journal of Early Childhood Literacy*, *14*, 459–481. https://doi.org/10.1177/1468798413501184

Jimenez, S. R., & Saylor, M. M. (2017). Preschoolers' word learning and story comprehension during shared book reading. *Cognitive Development, 44*, 57–68.

Justice, L. M. (2002). Word exposure conditions and preschoolers' novel word learning during shared storybook reading. *Reading Psychology*, *23*, 87–106. https://doi.org/10.1080/027027102760351016

Lenhard, A., Lenhard, W., Segerer, R., & Suggate, S. (2015). *Peabody Picture Vocabulary Test – Fourth Edition (German Adaption)*. Frankfurt am Main, Germany: Pearson Assessment.

Lenhart, J., Lenhard, W., Vaahtoranta, E., & Suggate, S. (2019). The effects of questions during shared-reading: Do demand-level and placement really matter? *Early Childhood Research Quarterly, 47*, 49–61. https://doi.org/10.1016/j.ecresq.2018.10.006

Lenth, R. (2018). emmeans: estimated marginal means, aka least-squares means. R package version 1.2.4. Retrieved from https://CRAN.R-project.org/package=emmeans

Locke, A., Ginsborg, J., & Peers, I. (2002). Development and disadvantage: implications for the early years and beyond. *International Journal of Language & Communication Disorders*, *37*, 3–15. https://doi.org/10.1080/13682820110089911

Lüdecke, D. (2018). sjPlot: Data visualization for statistics in social science. https://doi.org/10.5281/zenodo.1308157. R package version 2.6.2. Retrieved from https://CRAN.R-project.org/package=sjPlot

Marulis, L. M., & Neuman, S. B. (2010). The effects of vocabulary intervention on young children's word learning: A meta-analysis. *Review of Educational Research*, *80*, 300–335. https://doi.org/10.3102/0034654310377087

Marulis, L. M., & Neuman, S. B. (2013). How vocabulary interventions affect young children at risk: A meta-analytic review. *Journal of Research on Educational Effectiveness*, *6*, 223–262. https://doi.org/10.1080/19345747.2012.755591

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Mol, S. E., Bus, A. G., & de Jong, M. T. (2009). Interactive book reading in early education: A tool to stimulate print knowledge as well as oral language. *Review of Educational Research*, *79*, 979–1007. https://doi.org/10.3102/0034654309332561

Mol, S. E., Bus, A. G., de Jong, M. T., & Smeets, D. J. H. (2008). Added value of dialogic parent–child book readings: A meta-analysis. *Early Education and Development*, *19*, 7–26. https://doi.org/10.1080/10409280701838603

Niklas, F. & Schneider, W. (2015). With a little help: improving kindergarten children's vocabulary by enhancing the home literacy environment. *Reading and Writing*, *28*, 491–508. https://doi.org/10.1007/s11145-014-9534-z

Ninio, A., & Bruner, J. (1978). The achievement and antecedents of labelling. *Journal of Child Language*, *5*, 1–15. https://doi.org/10.1017/S0305000900001896

Petermann, F. (2011). *Wechsler Preschool and Primary Scale of Intelligence-III: WPPSI – Third Edition – German Version* (2nd ed.). Frankfurt am Main, Germany: Pearson Assessment.

Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*, 413–425. https://doi.org/10.1016/j.jml.2008.02.002

R Core Team (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Reese, E., & Cox, A. (1999). Quality of adult book reading affects children's emergent literacy. *Developmental Psychology*, *35*, 20–28. https://doi.org/10.1037/0012-1649.35.1.20

Scarborough, H. S. (1998). Early identification of children at risk for reading disabilities. In B. K. Shapiro, P. J. Accardo, & A. J. Capute (Eds.), *Specific reading disability* (pp. 75–199). Timonium, MD: York Press.

Schroeder, S., Würzner, K.-M., Heister, J., Geyken, A., & Kliegl, R. (2015). childLex: A lexical database of German read by children. *Behavior Research Methods*, *47*, 1085–1094. https://doi.org/10.3758/s13428-014-0528-1

Sénéchal, M., Ouellette, G., & Rodney, D. (2006). The misunderstood giant: On the predictive role of early vocabulary to future reading. In D. K. Dickinson & S. B. Neuman (Eds.), *Handbook of early literacy research* (Vol. 2, pp. 173–182). New York, NY: The Guilford Press.

Silverman, R. (2007). A comparison of three methods of vocabulary instruction during read-alouds in kindergarten. *The Elementary School Journal*, *108*, 97–113

Stahl, S. A., & Fairbanks, M. M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research*, *56*, 72–110. https://doi.org/10.3102/00346543056001072

Swingley, D. (2010). Fast mapping and slow mapping in children's word learning. *Language Learning and Development. 6*, 179–183. https://doi.org/10.1080/15475441.2010.484412

Twisk, J., de Boer, M., de Vente, W., & Heymans, M. (2013). Multiple imputation of missing values was not necessary before performing a longitudinal mixed-model analysis. *Journal of clinical epidemiology*, *66*, 1022–1028. https://doi.org/10.1016/j.jclinepi.2013.03.017

Vaahtoranta, E., Suggate, S., Jachmann, C., Lenhart, J., & Lenhard, W. (2018). Can explaining less be more? Enhancing vocabulary through explicit versus elaborative storytelling. *First Language, 38*, 198–217. https://doi.org/10.1177/0142723717737452

Walsh, B. A., Sánchez, C., & Burnham, M. M. (2016). Shared storybook reading in Head Start: Impact of questioning styles on the vocabulary of Hispanic dual language learners. *Early Childhood Education Journal*, *44*, 263–273. https://doi.org/10.1007/s10643-015-0708-3

Walsh, R. L., & Hodge, K. A. (2018). Are we asking the right questions? An analysis of research on the effect of teachers' questioning on children's language during shared book reading with young children. *Journal of Early Childhood Literacy, 18*, 264–294. https://doi.org/10.1177/1468798416659124

Wasik, B. A., Hindman, A. H., & Snell, E. K. (2016). Book reading and vocabulary development: A systematic review. *Early Childhood Research Quarterly*, *37*, 39–57. https://doi.org/10.1016/j.ecresq.2016.04.003

Westfall, J. (2016). PANGEA: Power ANalysis for GEneral Anova designs. University of Texas at Austin: Working paper. Retrieved from: http://jakewestfall.org/publications/pangea.pdf

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, *17*, 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Highlights

- Effects of question placement and demand-level during small-group shared-reading were examined.
- Asking questions during storytelling enhanced vocabulary learning.
- Question placement and demand-level did not exert differential effects on word learning.
- Question placement or demand-level did not interact with children's vocabulary knowledge.
- Question placement or demand-level did not interact with children's phonological working memory.

8 General Discussion

The present thesis was conducted to extend our knowledge about story interventions designed to foster young children's vocabulary development. Although story interventions have been promoted for improving parents' and teachers' story sessions at home (e.g., Whitehurst et al., 1988) or in institutional settings such as kindergartens, preschools, and elementary schools (e.g., Biemiller & Boote, 2006), the contribution and importance of many individual story features need still to be determined (Flack et al., 2018; R. L. Walsh & Hodge, 2018; Wasik et al., 2016). Consequently, the present thesis examined two aspects of story sessions that present important avenues not only for future research but also for the further improvement of story interventions.

The first aspect concerned the method in which the narrator delivers the story. Although story interventions targeting young children's vocabulary typically use books, free-telling of stories has been proposed as an alternative approach (Collins, 1999; McCabe, 1997), which might be particularly promising for populations that are predominantly coined through oral traditions (Reese, 2012). Despite the fact that related topics such as joint reminiscing about past events have received some attention lately (e.g., Fivush et al., 2006; Reese et al., 2010; Sparks & Reese, 2013), little is known about the relative efficacy of free-telling approaches compared to more traditional read-aloud approaches or mechanisms that might possibly explain potential differences between both methods of story delivery. Thus, Study 1 and 2 of the present thesis focused on a comparison of both methods of story delivery, trying to link narrator behavior during story presentation to child engagement and learning (see Chapters 4 and 5).

The second aspect concerned features of questions that are an integral part of natural shared book reading situations (Ninio & Bruner, 1978). As questions can dramatically increase children's active participation during story interventions, they are viewed as playing a key role for successful story interventions (Zucker et al., 2010). Although the general benefit of asking questions has already been demonstrated (e.g., Ewers & Brownson, 1999; Sénéchal, 1997; Sénéchal et al., 1995; B.A. Walsh & Blewitt, 2006), our knowledge about effects of specific question features, such as demand level and placement, is still incomplete (R.L. Walsh & Hodge, 2018). Consequently, Study 3 and 4 of the present thesis focused on question demand level and placement as two question features that might play an important role for successful story interventions (see Chapters 6 and 7).

8.1 Free-Telling versus Reading Aloud as Alternative Methods of Story Delivery

In terms of story delivery, we compared reading aloud and free-telling of stories as two alternative methods for presenting stories to children. In contrast to the virtually omnipresent book reading, free-telling of stories has received relatively little attention so far as a means to foster young children's vocabulary development (Isbell et al., 2004). On the one hand, we were interested in effects of reading aloud versus free-telling of stories on vocabulary acquisition. On the other hand, we wanted to examine theoretically implicated mechanisms, namely narrator behavior, linguistic complexity, and child engagement that might differ between both methods of story delivery and might have an impact on word learning (Myers, 1990; Suggate et al., 2013; Zeece, 1997).

In Study 1, we adapted a within-subjects study design, which had been used in a previous study with elementary school children (Suggate et al., 2013). In both experiments of Study 1, we found no differences in narrators' frequency of eye contact, gesticulation, or voice modulation, nor in children's attentiveness, nor concerning word learning. Thus, our results stood in direct opposition to findings from previous studies for slightly older children, in which differences in children's learning gains (Suggate et al., 2013; Trostle & Hicks, 1998; Uchiyama, 2011) or narrator behavior (Myers, 1990) were observed.

However, in the general discussion of Study 1, we mentioned several methodological problems that may have had affected the results. Although telling six stories within one session might have demanded too much from preschool children as well as hearing each story only once might have reduced any learning gains in general, we assumed that main problem might have been the operationalization of the free-telling condition. First, telling six stories directly one after another required changing methods of story delivery several times within each subject, which might have impeded the narrator implementing the respective conditions of story delivery. Second, the very short length of the individual stories in combination with the number of repetitions might have drastically reduced the need for actually reading the text in the reading-aloud condition, thus threatening the integrity of the intervention. Both aspects, in consequence, might have resulted in very similar story presentations across both story-delivery conditions and thus in similar learning gains. As narrator behavior did not differ as a function of story delivery in both experiments of Study 1, this explanation seems at least plausible and merited further examination (for a more detailed analysis see the general discussion of Study 1 in Chapter 4).

Study 2 was designed to target the methodological problems of Study 1. In order to do so, each story was presented twice and the total number of stories was reduced from four to six. More importantly, however, story delivery was operationalized as a between-subjects factor, with each child receiving all stories in only one story-delivery condition, which in turn did not require the narrator to switch between methods of story delivery during one session. As hypothesized in the discussion of Study 1, this resulted in the narrators differing between story conditions by using more eye contact and gesticulation during free-telling of stories than during reading aloud, which better approximated differences in narrator behavior observed in naturalistic studies (e.g., Myers, 1990). This time, in line with previous studies (Suggate et al., 2013; Trostle & Hicks, 1998; Uchiyama, 2011) and probably as a result of differing narrator behavior, we also found that children learned more words during free-telling of stories and gained a better understanding of the story content.

Moreover, by including audiotape conditions of both methods of story delivery, Study 2 allowed a direct contrast of effects for narrator behavior versus those of language complexity. Corresponding with comparisons between oral and written language (e.g., Montag et al., 2015; Nation, 2006), analyses of language complexity indicated that free-telling of stories provided less complex linguistic input than reading aloud. However, as audiotape conditions of both methods of story delivery did not differ regarding word learning or story comprehension and were significantly inferior only to free-telling, but not to reading aloud, narrator behavior, but not linguistic complexity, seemed to matter more for children's learning.

This was somewhat surprising because numerous (longitudinal) studies found positive relationships between complex language input and children's language development (e.g., Dickinson & Porche, 2011; Hart & Risley, 1995). One explanation may be that language input in our study differed primarily regarding grammatical but not semantic complexity. As narrators were instructed to retell the story gist of the somewhat short stories and to employ each of the four target words once, there was not much scope for semantic differences to emerge. This was supported by high correlations found in the latent semantic analysis in Study 2, indicating large amounts of semantic overlap. Thus, due to the study design, differences might be only observable in grammatical measures but not in measures of word learning or story comprehension. Another explanation might be that differences in language complexity between both methods of story delivery were too small to matter. Although free-telling provided less complex linguistic input – effect sizes were $d = 0.53$ and $d = 0.74$ for LIX and FLESCH as indicators of language surface complexity – our narrators (i.e., student research assistants) possessed a high educational background and good language proficiency. Consequently, differences between their oral language and written material might be considerably smaller as in other less-well educated populations. Finally, linguistic complexity may predominantly have an incremental effect on vocabulary development.

Therefore, differences in linguistic input might have little short-term effects, but they might influence children's language development in the long run.

Notwithstanding the reflections on linguistic complexity, it seems that free-telling of stories offers an effective, supplemental method of story delivery to foster children's participation in stories, their vocabulary, and story comprehension. Therefore, in summary, free-telling can be seen as an opportunity to further enrich the repertoire of methods available to engage children in the interesting world of stories.

8.2 Question Demand Level and Placement

Regarding question demand level, three major hypotheses have been proposed in shared-reading literature. The aptitude-treatment-interaction hypothesis (Reese & Cox, 1999) assumes that children with good cognitive skills benefit from cognitively challenging high-demand questions, whereas lower-demand questions profit their less well-equipped peers. In contrast, the high-demand hypothesis (B.A. Walsh & et al., 2016) and the scaffolding-like hypothesis (Blewitt et al., 2009) proposed that there are no differential effects of question demand level on initial, primarily phonological word acquisition, but that high-demand questions or respectively a scaffolding-like gradual transition from low- to high-demand questions might be especially helpful to acquire deeper word knowledge. Concerning question placement, an interruption approach, favoring questions interspersed within stories (Brabham & Lynch-Brown, 2002), and a no-interruption approach, avoiding story interruptions by weaving the questions around the stories (Dickinson & Smith, 1994), as well as different kinds of aptitude-treatment-interaction hypotheses (Jimenez & Saylor, 2017; Reese & Cox, 1999) have been put forward.

In Study 3 and 4, we tested predictions made by different hypotheses on the effects of question demand level and placement. Study 3 used a one-to-one experimental design, that allowed focusing on each child individually, whereas Study 4 was designed as a small-group intervention (5-7 children), which represents a more naturalistic and feasible approach for kindergartens and preschools.

Study 3 was conceptually oriented at Blewitt et al.'s (2009) study design with its inclusion of a scaffolding-like demand level condition, its use of several measures for different facets of novel word acquisition, and its implementation as individual one-to-one reading sessions. We found that, in line with previous research (e.g., Blewitt et al., 2009; Ewers & Brownson, 1999; Sénéchal, 1997; Sénéchal et al., 1995; B.A. Walsh & Blewitt, 2006), using questions generally increased vocabulary learning compared to a just-reading condition, in which the narrators did not ask any questions. However, our results did not indicate meaningful differences as a function of different types of question demand level and placement. In addition, there was no interaction between these two question features and children's language skills. As this picture emerged for all three measures of vocabulary acquisition in the immediate as well as in the delayed posttest, the results of Study 3 did not provide evidence for any of the proposed hypotheses regarding question demand level or placement.

Although our results were not necessarily at odds with previous empirical findings – for example, Blewitt et al. (2009) or B. A. Walsh et al. (2016) also found no differences between different types of question placement –, we developed an attempt to explain the seemingly contradictory findings from experimental studies in the discussion of Study 3. Based on an inspection of the operationalization of different kinds of questioning styles in experimental studies, we identified two key points that merited further consideration. First, experimental studies typically gave little concrete indications about learning mechanisms by which different types of questions might produce different learning outcomes. Second, hypotheses were more or less the same for different kinds of age groups and reading settings. In particular, the latter might be problematic as it assumes that questions in one-to-one reading sessions, in which the adult narrator is not allowed to correct or elaborate on the child's answers (e.g., Blewitt et al., 2009; Justice, 2002; Reese & Cox, 1999), have the same effects as those in group settings, in which at least the children might correct and elaborate on each other's comments (e.g., B.A. Walsh et al., 2016), and those during natural occurring shared-reading, in which adult narrator is free to do as he or she pleases and typically corrects and extends the children's utterances (e.g., Dickinson & Smith, 1994; Gonzalez et al., 2014).

Consequently, we proposed that those one-to-one reading sessions, in which the adult narrator does not correct or elaborate on the child's answers, the benefit of questions might consist in guiding children's attention to important aspects of a story and in the cognitive processes that are initiated within each child through individual reflection about the questions. In contrast to older students, for whom individual cognitive processing of different types of questions has been shown to result in different learning gains (e.g., Cerdán et al., 2009; Jensen et al., 2014, Roelle & Berthold, 2017), this type of input may not be sufficient to produce differential learning gains in preschool children. Most studies using one-to-one reading sessions fit perfectly (e.g., Blewitt et al., 2009, Study 1; Justice, 2002; our own Study 3) or at least partially (Blewitt et al., 2009, Study 2) within our line of argument (see the discussion of Study 3 in Chapter 6 for an extended elaboration of the argument and a critical examination of individual studies).

In other settings, such as natural shared book readings or experimental small-group sessions, different types of questions might produce differential discussions, which in turn might result in differential learning gains. Supporting this idea, observational studies conducted in classroom settings and at home typically report differential relations between different types of question placement and/or demand level and children's vocabulary development (e.g., Dickinson & Smith, 1994; Haden et al., 1996; Pellegrini et al., 1985). However, as parents' or teachers' input might have differed not only regarding question features, but also concerning a wide range of other factors, such as the number of comments, explanations, affective behavior, or interaction length, the causal interpretation of question characteristics leading to different learning gains is not feasible.

In order to test the effects of question placement and demand level as stimulators of different kinds of discussion in a systematic and controlled way, the subsequent study (Study 4) used a small-group design with five to seven children per group. Comparable to previous studies (e.g., Blewitt et al., 2009; Reese & Cox, 1999; B. A. Walsh et al., 2016; our own Study 3), narrators were instructed to encourage the children to give input and discuss their peers' comments and answers, but they were not allowed to provide own content-related input. By doing so, one the one hand, we wanted to increase group discussion and therefore allow questions to unfold different patterns of interaction; on the other hand, we wanted to keep the focus on questions and to avoid mixing them with other types of narrator input such as additional explanations and comments. Comparable to Study 3, questions had a positive impact on word learning and generally resulted in larger gains than a just-reading control group. Again, there were no differences on any of the language measures between question placement and demand-level conditions, nor was there any meaningful interaction of those question features with children's general vocabulary knowledge or their phonological working memory. Therefore, Study 4 did also not support any of the proposed hypotheses on question placement and demand level.

Taken together, Studies 3 and 4 indicate that for young children differences in question placement or demand level seem not be sufficient to produce differential learning effects, at least when the adult narrator does not follow the questions with meaningful input such as corrective feedback, comments, or elaborations. Extended interaction between narrators and young children following questions – as typically observed in natural shared book reading (e.g., Dickinson & Smith, 1994; Ninio & Bruner, 1978) – may be truly crucial for vocabulary learning, with different kinds of question demand level and placement probably providing foundations for differential adult-child interaction to occur. Consequently, the claim that some types of questions might inherently result in better outcomes, which might possibly apply to older university students (e.g., Roelle & Berthold, 2017), seems unjustified for preschool children. This emphasizes the central role that adults play in story sessions with young children, not only for engaging children through questions, but also for providing corrective feedback and for elaborating on their utterances.

8.3 Limitations and Directions for Future Research

The empirical work conducted as a part of the present doctoral thesis adds to an increase of knowledge about two important aspects of story-based interventions designed to foster children's vocabulary development. More precisely, the current works helps to better understand effects and potential mechanisms of different types of question placement and demand level as well as of free-telling versus reading aloud as alternative methods of story delivery. However, there are some limitations that need to be addressed.

First, across studies narrator input was restricted and thus did not closely mirror real-world adult-child interactions during story sessions. Although this was a necessary constraint within the experimental designs in order to make sure that effects could be causally linked to differences in experimental conditions, this clearly reduces ecological validity of our studies. In particular, concerning our studies on question demand level and placement, it might be argued that instructing narrators not to provide corrective feedback or elaborate on children's answers and examining pure effects of questions might be artificial. However in this regard, our approach in Study 3 and 4 was comparable to that of most other relevant studies on question demand level and placement and the presented hypotheses were typically developed and proposed within those types of well-controlled studies (e.g., Blewitt et al., 2009; Justice, 2002; Reese & Cox, 1999; B. A. Walsh et al., 2016). Regarding question demand level and placement, future studies should systematically extend our study procedure. Following Study 4, a sensible next step would be to use different types of question placement and demand-level conditions as a starting point for (partially) scripted adult-child interactions. These scripts should ensure, for instance, that different types of question demand level receive similar amounts of feedback and elaborations, but that adult-child interaction stays on the targeted demand level and takes place within, before, or after the stories.

Second, our studies did not include video recordings of the reading sessions and did therefore not allow assessing narrator behavior, child engagement, and adult-child interaction during story presentation. Although we used observational protocols in our studies, some of the variables were marked by low interrater-reliability (see Study 1 and 2) and not all variables that could be of interest were observed. Concerning effects of question demand level and placement, for example, we simply do not know whether children gave correct answers and whether in Study 4 questions resulted in helpful, irrelevant, or even distracting group discussion.

Regarding our studies on reading aloud versus free-telling of stories as alternative methods of story delivery, this might have been even more problematic because narrator behavior and child engagement that are assumed to give free-telling an edge over reading aloud were not observed in a differentiated way. Here, effects of eye contact during story presentation should be separated from those of narrators' gesture usage, which in turn should be examined in a finer-grained way. As indicated by research on the effects of gestures on learning (e.g., Hostetter, 2011; Krahmer & Swerts, 2007; McNeill, 1992), it may be reasonable to distinguish representational forms (i.e., iconic and metaphoric gestures), deictic forms, and non-semantic beat forms of gesture.

Here, future studies should include video recordings of the story sessions. Moreover, it might be worthwhile to experimentally manipulate each of the different facets of narrator behavior to examine their specific effects.

Third, our studies represent only short-term interventions and they examined only short-term effects. Although Study 3 comprised a delayed posttest, which was conducted approximately four weeks after the end of the intervention phase, and Study 4 performed the posttest around one week after the last reading session, Studies 1 and 2 used only immediate posttests. Thus, we cannot say whether potential effects – for example, observed in Study 2 – may be preserved over a longer period. More problematically, however, our studies represent short-term interventions that do not reflect the incremental process that is typical for the vocabulary development (Biemiller & Slonim, 2001; Hart & Risley, 1995). For instance, as discussed for the potential influence of linguistic complexity during reading aloud versus free-telling of stories, differences might have little short-term effects, but they may influence children's language development in the long run. In consequence, to examine this type of incremental effects, there is a clear need for long-term intervention studies that also track potential effects over a prolonged span of time after the completion of the intervention.

Finally, the current work was restricted to selected aspects of vocabulary learning. This is particularly true for Studies 1 and 2, in which only shallow, receptive measures of target-word acquisition (picture selection) were used. Studies 3 and 4, in contrast, included additionally a target-word-definition task, which represents a depth measure of vocabulary knowledge (Hoffman, Teale, & Paciga, 2014). Following recommendations from recent reviews on vocabulary measurement (Hadley & Dickinson, 2018; Hoffman et al., 2014; Pearson, Hiebert, & Kamil, 2007), future studies should try to use a wider range of measures, which tap into different facets of vocabulary acquisition, in order to obtain a finer-grained picture of children's emerging word knowledge.

8.4 Conclusion and Practical Implications

Despite certain limitations, perhaps present in any line of scientific inquiry, the studies within this doctoral thesis contribute to the growing literature on means to foster young children's vocabulary development. The present doctoral thesis extends previous research and sheds light on the effects reading aloud versus free-telling of stories as methods of story delivery as well as on effects of question demand level and placement on children's word learning from listening to stories.

More precisely, we conclude: Firstly, that free-telling of stories seems to offer an interesting, additional method to involve children in the world of stories, which might be beneficial for children's vocabulary development, story comprehension, and engagement (Studies 1 and 2). Secondly, that adults' questions were demonstrated to increase vocabulary learning from stories, but that question demand level and placement may have little impact when decoupled from the adult's further contributions to the dialogic cycle that is typical for shared-reading (Studies 3 and 4).

In terms of educational practice, the results of the present doctoral thesis imply that a wide array of various activities – including both reading and telling stories to children – should be used to foster children's vocabulary development. As both types of story delivery have inherent advantages – reading provides more complex linguistic input and free-telling of stories enables the narrator to use more eye contact and gestures –, story delivery should not result in an either-or decision, but different methods should complement each other. Of course, this conclusion does not only apply to different methods of story delivery, but also to other types of activities such as conversations during toy play or mealtime (Hoff-Ginsberg, 1991) or reminiscing about pas real-life events (Fivush et al., 2006). In addition, as our world is ever changing and becoming increasingly complex, it would be simplistic to assume that a single method might fit it all. Instead, the variety of different methods available should be embraced in order to best foster each child's individual development.

Moreover, in line with other studies (e.g., Blewitt et al., 2009; Ewers & Brownson, 1999), the present doctoral thesis emphasizes the importance of meaningful interactivity. Thus, book and story-telling sessions as well as other activities should be marked by lively interaction between adult and child. Although we still do not know whether – under different circumstances – different types of question demand level and placement may result in differential adult-child interactions and thus in differential learning gains, questions may be particularly suited to stimulate adult-child interaction and active child engagement (Zucker et al., 2010). However, asking questions cannot be an end in itself. Asking questions, because particular guidelines for shared-reading mandate this, may have little value. Instead, questions in conjunction with a responsive adult should signal to the child that his or her thoughts matter and that an expression of his or her opinion is wanted and valued. Moreover, adults' subsequent elaborations on children's utterances may not only provide opportunities to correct false conceptions and to model more advanced language use, but they also acknowledge the children's contributions to the conversation.

Finally, it is necessary to draw attention to the fact that the so-called "vocabulary gap" (e.g., Biemiller, 2003; Christ & Wang, 2011) does not only mean that disadvantaged children know fewer words, but that they are deprived of world knowledge and experiences that are prerequisites for their development as well as their academic achievement. Thus, vocabulary interventions for disadvantaged at-risk children that are confined to the mere teaching of words do not do justice to the poverty of (social) experiences that probably underlie those children's restricted vocabulary breadth and depth and, consequently, they are condemned to fall short. Instead, interventions should additionally target the poverty of (social) experiences causing children's language impoverishment by providing meaningful interaction and discussion on precisely those missing concepts. This underlines again the important role that adults play – not only for conducting successful story interventions, which were explored through a rigorous program of research in the current thesis, but also for children's development in general.

References

Akinnaso, F. N. (1982). On the differences between spoken and written language. *Language and Speech, 25*, 97–125. https://doi.org/10.1177/002383098202500201

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., . . . Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York, NY: Longman.

Barton, B. (1986). *Tell me another: Storytelling and reading aloud at home, at school, and in the community*. Markham, CA: Heinemann.

Biau, E., & Fernández, L. M., Holle, H., Avila, C., & Soto-Faraco, S. (2016). Hand gestures as visual prosody: BOLD responses to audio–visual alignment are modulated by the communicative nature of the stimuli. *NeuroImage, 132*, 129–137. https://doi.org/10.1016/j.neuroimage.2016.02.018

Biemiller, A. (2003). Vocabulary: Needed if more children are to read well. *Reading Psychology*, *24*, 323–335. https://doi.org/10.1080/02702710390227297

Biemiller, A., & Boote, C. (2006). An effective method for building meaning vocabulary in primary grades. *Journal of Educational Psychology, 98,* 44–62. https://doi.org/10.1037/0022-0663.98.1.44

Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of Educational Psychology, 93,* 498–520. https://doi/10.1037/0022-0663.93.3.498

Blake, J., Macdonald, S., Bayrami, L, Agosta, V., & Milian, A. (2006). Book reading styles in dual-parent and single-mother families. *British Journal of Educational Psychology, 76,* 501–515. https://doi.org/10.1348/000709905X49719

Blewitt, P., Rump, K. M., Shealy, S. E., & Cook, S. A. (2009). Shared book reading: When and how questions affect young children's word learning. *Journal of Educational Psychology*, *101*, 294–304. https://doi.org/10.1037/a0013844

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W.H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. New York, NY: David McKay.

Bloom, K. (1974). Eye contact as a setting event for infant learning. *Journal of Experimental Child Psychology, 17,* 250–263. https://doi.org/10.1016/0022-0965(74)90070-8

Brabham, E. G., & Lynch-Brown, C. (2002). Effects of teachers' reading-aloud styles on vocabulary acquisition and comprehension of students in the early elementary grades. *Journal of Educational Psychology*, *94*, 465–473. https://doi.org/10.1037/0022-0663.94.3.465

Bus, A. G., van IJzendoorn, M. H., & Pellegrini, A. D. (1995). Joint book reading makes for success in learning to read: A meta-analysis on intergenerational transmission of literacy. *Review of Educational Research, 65,* 1–21. https://doi.org/10.3102%2F00346543065001001

Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan, & G. A. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 265–293). New York, NY: Cambridge University Press.

Carey, S. (2010). Beyond fast mapping. *Language Learning and Development, 6*, 184–205. https://doi.org/10.1080/15475441.2010.484379

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development*, *15*, 17–29.

Cerdán, R., Vidal-Abarca, E., Martínez, T., Gilabert, R., & Gil, L. (2009). Impact of question-answering tasks on search processes and reading comprehension. *Learning and Instruction*, *19*, 13–27. https://doi.org/10.1016/j.learninstruc.2007.12.003

Chall, J. S., Jacobs, V. A., & Baldwin, L. E. (1990). *The reading crisis: Why poor children fall behind*. Cambridge, MA: Harvard University Press.

Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1*, 73–105. https://doi.org/10.1111/j.1756-8765.2008.01005.x

Christ, T., & Wang, X. C. (2011). Closing the vocabulary gap?: A review of research on early childhood vocabulary practices. *Reading Psychology*, *32*, 426–458. https://doi.org/10.1080/02702711.2010.495638

Christian, K., Morrison, F. J., Frazier, J. A., & Massetti, G. (2000). Specificity in the nature and timing of cognitive growth in kindergarten and first grade. *Journal of Cognition and Development*, *1*, 429–448. https://doi.org/10.1207/S15327647JCD0104_04

Collins, F. (1999). The use of traditional storytelling in education to the learning of literacy skills. *Early Child Development and Care, 152*, 77–108. https://doi.org/10.1080/0300443991520106

Connor, C. M., Piasta, S. B., Fishman, B., Glasney, S., Schatschneider, C., Crowe, E., . . . Morrison, F. J. (2009). Individualizing student instruction precisely: Effects of child x instruction interactions on first graders' literacy development. *Child Development*, *80*, 77–100. https://doi.org/10.1111/j.1467-8624.2008.01247.x

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*, 268–294. https://doi.org/10.1037/0096-3445.104.3.268

Cromley, J. G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology*, *99*, 311–325. https://doi.org/10.1037/0022-0663.99.2.311

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology, 33*, 934–945. https://doi.org/10.1037/0012-1649.33.6.934

Curenton, S. M. (2006). Oral storytelling. A cultural art that promotes school readiness. *Young Children, 61*(5), 78–89.

Curenton, S. M., & Craig, M. J. (2011). Shared-reading versus oral storytelling: Associations with preschoolers' prosocial skills and problem behaviours. *Early Child Development and Care, 181*, 123–146. https://doi.org/10.1080/03004430903292208

Curenton, S. M., Craig, M. J., & Flanigan, N. (2008). Use of decontextualized talk across story contexts: How oral storytelling and emergent reading can scaffold children's development. *Early Education and Development, 19*, 161–187. https://doi.org/10.1080/10409280701839296

Dale, P. S., Crain-Thoreson, C., Notari-Syverson, A., & Cole, K. (1996). Parent-child book reading as an intervention technique for young children with language delays. *Topics in Early Childhood Special Education, 16*, 213–235. https://doi.org/10.1177%2F027112149601600206

Danis, A., Bernard, J.-M., & Leproux, C. (2000). Shared picture-book reading: A sequential analysis of adult-child verbal interactions. *British Journal of Developmental Psychology, 18*, 369–388. https://doi.org/10.1348/026151000165751

DeLoache, J. S., & DeMendoza, O. A. P. (1987). Joint picturebook interactions of mothers and 1-year-old children. *British Journal of Developmental Psychology, 5*, 111–123. https://doi.org/10.1111/j.2044-835X.1987.tb01047.x

De Temple, J. M. (2001). Parents and children reading books together. In D. K. Dickinson & P. O. Tabors (Eds.), *Beginning literacy with language* (pp. 31–51). Baltimore, MD: Paul H. Brookes.

De Temple, J. M., & Snow, C. E. (2003). Learning words from books. In A. van Kleeck, S. A. Stahl, & E. B. Bauer (Eds.), *On reading books to children: Parents and teachers* (pp. 16–36). Mahwah, NJ: Lawrence Erlbaum Associates.

Dickinson, D. K., & Keebler, R. (1989). Variation in preschool teachers' styles of reading books. *Discourse Processes, 12*, 353–375. https://doi.org/10.1080/01638538909544735

Dickinson, D. K., & Porche, M. V. (2011). Relation between language experiences in preschool classrooms and children's kindergarten and fourth-grade language and reading abilities. *Child Development, 82*, 870–886. https://doi.org/10.1111/j.1467-8624.2011.01576.x

Dickinson, D. K., & Smith, M. W. (1994). Long-term effects of preschool teachers' book readings on low-income children's vocabulary and story comprehension. *Reading Research Quarterly*, *29*, 104–122. https://doi.org/10.2307/747807

Dickinson, D. K., & Tabors, P. O. (Eds.). (2001). *Beginning literacy with language: Young children learning at home and school*. Baltimore, MD: Paul H. Brookes.

Elley, W. B. (1989). Vocabulary acquisition from listening to stories. *Reading Research Quarterly*, *24*, 174–187. https://doi.org/10.2307/747863

Ewers, C. A., & Brownson, S. M. (1999). Kindergarteners' vocabulary acquisition as a function of active vs. passive storybook reading, prior vocabulary, and working memory. *Reading Psychology*, *20*, 11–20. https://doi.org/ 10.1080/027027199278484

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science, 16*, 234–248. https://doi.org/10.1111/desc.12019

Fivush, R., Haden, C. A., & Reese, E. (2006). Elaborating on elaborations: Role of maternal reminiscing style in cognitive and socioemotional development. *Child Development, 77*, 1568–1588. https://doi.org/10.1111/j.1467-8624.2006.00960.x

Flack, Z. M., Field, A. P., & Horst, J. S. (2018). The effects of shared storybook reading on word learning: A meta-analysis. *Developmental Psychology, 54*, 1334–1346. https://doi.org/10.1037/dev0000512

Fletcher, K. L., & Reese, E. (2005). Picture book reading with young children: A conceptual framework. *Developmental Review, 25*, 64–103. https://doi.org/10.1016/j.dr.2004.08.009

Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends in Cognitive Sciences, 3*, 419–429. https://doi.org/10.1016/S1364-6613(99)01397-2

Gonzalez, J. E., Pollard-Durodola, S., Simmons, D. C., Taylor, A. B., Davis, M. J., Fogarty, M., & Simmons, L. (2014). Enhancing preschool children's vocabulary: Effects of teacher talk before, during and after shared reading. *Early Childhood Research Quarterly*, *29*, 214–226. https://doi.org/10.1016/j.ecresq.2013.11.001

Grugeon, E., & Gardner (2012). *The art of storytelling for teachers and pupils: Using stories to develop literacy in primary classrooms*. London, UK: Routledge.

Haden, C. A., Reese, E., & Fivush, R. (1996). Mothers' extratextual comments during storybook reading: Stylistic differences over time and across texts. *Discourse Processes*, *21*, 135–169. https://doi.org/10.1080/01638539609544953

Hadley, E. B., & Dickinson, D. K. (2018). Measuring young children's word knowledge: A conceptual review. *Journal of Early Childhood Literacy*. Advance online publication. https://doi.org/10.1177/1468798417753713

Hammett, L. A., van Kleeck, A., & Huberty, C. J. (2003). Patterns of parents' extratextual interactions during book sharing with preschool children: A cluster analysis study. *Reading Research Quarterly, 38*, 442–468. https://doi.org/10.1598/RRQ.38.4.2

Hargrave, A. C., & Sénéchal, M. (2000). A book reading intervention with preschool children who have limited vocabularies: The benefits of regular reading and dialogic reading.

Early Childhood Research Quarterly, 15, 75–90. https://doi.org/10.1016/S0885-2006(99)00038-1

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Paul H. Brookes.

Hart, B., & Risley, T. R. (2003). The early catastrophe: The 30 million word gap. *American Educator, 27*, 4–9.

Hayes, D. P. (1988). Speaking and writing: Distinct patterns of word choice. *Journal of Memory and Language, 27*, 572–585. https://doi.org/10.1016/0749-596X(88)90027-7

Hayes, D. P., & Ahrens, M. G. (1988). Vocabulary simplification for children: A special case of 'motherese'? *Journal of Child Language, 15*, 395–410. https://doi.org/10.1017/S0305000900012411

Heath, S. B. (1982). What no bedtime story means: Narrative skills at home and school. *Language in Society, 11*, 49–76. https://doi.org/10.1017/S0047404500009039

Hemphill, L., & Tivnan, T. (2008). The importance of early vocabulary for literacy achievement in high-poverty schools. *Journal of Education for Students Placed at Risk, 13*, 426–451. https://doi.org/10.1080/10824660802427710

Hindman, A. H., Connor, C. M., Jewkes, A. M., & Morrison, F. J. (2008). Untangling the effects of shared book reading: Multiple factors and their associations with preschool literacy outcomes. *Early Childhood Research Quarterly, 23*, 330–350. https://doi/10.1016/j.ecresq.2008.01.005

Hindman, A. H., Wasik, B. A., & Erhart, A. C. (2012). Shared book reading and Head Start preschoolers' vocabulary learning: The role of book-related discussion and curricular connections. *Early Education and Development, 23*, 451–474. https://doi.org/10.1080/10409289.2010.537250

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development, 74,* 1368–1378. https://doi.org/10.1111/1467-8624.00612

Hoff-Ginsberg, E. (1991). Mother-child conversation in different social classes and communicative settings. *Child Development, 62*, 782–796. https://doi.org/10.1111/j.1467-8624.1991.tb01569.x

Hoffman, J. L., Teale, W. H., & Paciga, K. A. (2014). Assessing vocabulary learning in early childhood. *Journal of Early Childhood Literacy, 14*, 459–481. https://doi.org/10.1177/1468798413501184

Holler, J., Schubotz, L., Kelly, S., Hagoort, P., Schuetze, M., & Özyürek, A. (2014). Social eye gaze modulates processing of speech and co-speech gesture. *Cognition, 133,* 692–697. https://doi.org/10.1016/j.cognition.2014.08.008

Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological Bulletin, 137*, 297–315. https://doi.org/10.1037/a0022128

Howe, A., & Johnson, J. (Eds.) (1992). *Common bonds: Storytelling in the classroom*. London, UK: Hodder & Stoughton.

Hubbard, A. L., Wilson, S. M., Callan, D. E., & Dapretto, M. (2009). Giving speech a hand: Gesture modulates activity in auditory cortex during speech perception. *Human Brain Mapping, 30*, 1028–1037. https://doi.org/10.1002/hbm.20565

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive Psychology, 61*, 343–365. https://doi.org/10.1016/j.cogpsych.2010.08.002

Isbell, R., Sobol, J., Lindauer, L., & Lowrance, A. (2004). The effects of storytelling and story reading on the oral language complexity and story comprehension of young children. *Early Childhood Education Journal, 32*, 157–163. doi:10.1023/B:ECEJ.0000048967.94189.a3

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the test…or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review, 26*, 307–329. https://doi.org/10.1007/s10648-013-9248-9

Jimenez, S. R., & Saylor, M. M. (2017). Preschoolers' word learning and story comprehension during shared book reading. *Cognitive Development, 44*, 57–68. https://doi.org/10.1016/j.cogdev.2017.08.011

Justice, L. M. (2002). Word exposure conditions and preschoolers' novel word learning during shared storybook reading. *Reading Psychology, 23*, 87–106. https://doi.org/10.1080/027027102760351016

Justice, L. M., & Pullen, P. C. (2003). Promising interventions for promoting emergent literacy skills: Three evidence-based approaches. *Topics in Early Childhood Special Education, 23*, 99–113. https://doi.org/10.1177%2F02711214030230030101

Kostyrka-Allchorne, K., Cooper, N. R., & Simpson, A. (2017). The relationship between television exposure and children's cognition and behaviour: A systematic review. *Developmental Review, 44*, 19–58. https://doi.org/10.1016/j.dr.2016.12.002

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language, 57*, 396–414. https://doi.org/10.1016/j.jml.2007.06.005

Leseman, P. P. M., & de Jong, P. F. (1998). Home literacy: Opportunity, instruction, cooperation and social-emotional quality predicting early reading achievement. *Reading Research Quarterly, 33*, 294–318. https://doi.org/10.1598/RRQ.33.3.3

Linebarger, D. L., & Vaala, S. E. (2010). Screen media and language development in infants and toddlers: An ecological perspective. *Developmental Review, 30*, 176–202. https://doi.org/10.1016/j.dr.2010.03.006

Lockhart, R. S., & Craik, F. I. M. (1990). Levels of processing: A retrospective commentary on a framework for memory research. *Canadian Journal of Psychology, 44*, 87–112. https://doi.org/10.1037/h0084237

Martinez, M. G., & Teale, W. H. (1993). Teacher storybook reading style: A comparison of six teachers. *Research in the Teaching of English, 27*, 175–199.

Marulis, L. M., & Neuman, S. B. (2010). The effects of vocabulary intervention on young children's word learning: A meta-analysis. *Review of Educational Research, 80,* 300–335. https://doi.org/10.3102/0034654310377087

Marulis, L. M., & Neuman, S. B. (2013). How vocabulary interventions affect young children at risk: A meta-analytic review. *Journal of Research on Educational Effectiveness*, *6*, 223–262. https://doi.org/10.1080/19345747.2012.755591

McCabe, A. (1997). Cultural background and storytelling: A review and implications for schooling. *The Elementary School Journal, 97*, 453–473. https://doi.org/10.1086/461876

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago, IL: University of Chicago Press.

Mol, S. E., & Bus, A. G. (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin*, *137*, 267–296. https://doi.org/10.1037/a0021890

Mol, S. E., Bus, A. G., & de Jong, M. T. (2009). Interactive book reading in early education: A tool to stimulate print knowledge as well as oral language. *Review of Educational Research*, *79*, 979–1007. https://doi.org/10.3102/0034654309332561

Mol, S. E., Bus, A. G., de Jong, M. T., & Smeets, D. J. H. (2008). Added value of dialogic parent–child book readings: A meta-analysis. *Early Education and Development*, *19*, 7–26. https://doi.org/10.1080/10409280701838603

Monk, A. F., & Gale, C. (2002). A look is worth a thousand words: Full gaze awareness in video-mediated conversation. *Discourse Processes, 33*, 257–278. https://doi.org/10.1207/S15326950DP3303_4

Montag, J. L., Jones, M. N., & Smith, L. B. (2015). The words children hear: Picture books and the statistics for language learning. *Psychological Science*, *26*, 1489–1496. https://doi.org/10.1177/0956797615594361

Morgan, P. L., Farkas, G., Hillemeier, M. M., Hammer, C. S., & Maczuga, S. (2015). 24-month-old children with larger oral vocabularies display greater academic and behavioral functioning at kindergarten entry. *Child Development, 86*, 1351–1370. https://doi.org/ 10.1111/cdev.12398

Muter, V., Hulme, C., Stevenson, J., & Snowling, M. J. (2004). Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading development: Evidence from a longitudinal study. *Developmental Psychology, 40*, 665–681. https://doi.org/10.1037/0012-1649.40.5.665

Myers, P. (1990). Stories from print. *Language Arts*, *67*, 824–831.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*, 59–82. https://doi.org/10.3138/cmlr.63.1.59

Ninio, A. (1980). Picture-book reading in mother-infant dyads belonging to two subgroups in Israel. *Child Development*, *51*, 587–590. https://doi.org/10.2307/1129299

Ninio, A., & Bruner, J. (1978). The achievement and antecedents of labelling. *Journal of Child Language*, *5*, 1–15. https://doi.org/10.1017/S0305000900001896

O’Donnell, R. C. (1974). Syntactic differences between speech and writing. *American Speech, 49*, 102–110. https://doi.org/10.2307/3087922

Pearson, P. D., Hiebert, E. H., & Kamil, M. L. (2007). Vocabulary assessment: What we know and what we need to learn. *Reading Research Quarterly*, *42*, 282–296. https://doi.org/10.1598/RRQ.42.2.4

Pellegrini, A. D., Brody, G. H., & Sigel, I. E. (1985). Parents’ book-reading habits with their children. *Journal of Educational Psychology, 77*, 332–340. https://doi.org/10.1037/0022-0663.77.3.332

Pellegrini, A. D., Perlmutter, J. C., Galda, L., & Brody, G. H. (1990). Joint reading between black Head Start children and their mothers. *Child Development, 61*, 443–453. https://doi.org/10.1111/j.1467-8624.1990.tb02791.x

Purpura, D. J., Hume, L. E., Sims, D. M., & Lonigan, C. J. (2011). Early literacy and early numeracy: The value of including early literacy skills in the prediction of numeracy development. *Journal of Experimental Child Psychology*, *110*, 647–658. https://doi.org/10.1016/j.jecp.2011.07.004

Raines, S. C., & Isbell, R. T. (1994). *Stories: Children's literature in early education*. Albany, NY: Delmar.

Reese, E. (2012). The tyranny of shared book-reading. In S. Suggate & E. Reese (Eds.), *Contemporary debates in childhood education and development* (pp. 59–68). London, UK: Routledge.

Reese, E., & Cox, A. (1999). Quality of adult book reading affects children’s emergent literacy. *Developmental Psychology*, *35*, 20–28. https://doi.org/10.1037/0012-1649.35.1.20

Reese, E., Leyva, D., Sparks, A., & Grolnick, W. (2010). Maternal elaborative reminiscing increases low-income children's narrative skills relative to dialogic reading. *Early Education and Development, 21*, 318–342. https://doi.org/10.1080/104092892010481552

Roelle, J., & Berthold, K. (2017). Effects of incorporating retrieval into learning tasks: The complexity of the tasks matters. *Learning and Instruction, 49*, 142–156. https://doi.org/10.1016/j.learninstruc.2017.01.008

Roelle, J., Roelle, D., & Berthold, K. (2019). Test-based learning: Inconsistent effects between higher- and lower-level test questions. *The Journal of Experimental Education, 87*, 299–313. https://doi.org/10.1080/00220973.2018.1434756

Roney, R. C. (1996). Storytelling in the classroom: Some theoretical thoughts. *Storytelling World*, *9*, 7–9.

Scarborough, H. S., & Dobrich, W. (1994). On the efficacy of reading to preschoolers. *Developmental Review, 14*, 245–302. https://doi.org/10.1006/drev.1994.1010

Scott, J. A., Nagy, W. E., & Flinspach, S. L. (2008). More than merely words: Redefining vocabulary learning in a culturally and linguistically diverse society. In A. E. Farstrup & S. J. Samuels (Eds.), *What research has to say about vocabulary instruction* (pp. 182–209). Newark, DE: International Reading Association.

Sénéchal, M. (1997). The differential effect of storybook reading on preschoolers' acquisition of expressive and receptive vocabulary. *Journal of Child Language, 24*, 123–138. htpps://doi.org/10.1017/S0305000996003005

Sénéchal, M., Ouellette, G., & Rodney, D. (2006). The misunderstood giant: On the predictive role of early vocabulary to future reading. In D. K. Dickinson & S. B. Neuman (Eds.), *Handbook of early literacy research* (Vol. 2, pp. 173–182). New York, NY: The Guilford Press.

Sénéchal, M., Thomas, E., & Monker, J.-A. (1995). Individual differences in 4-year-old children's acquisition of vocabulary during storybook reading. *Journal of Educational Psychology*, *87*, 218–229. https://doi.org/10.1037/0022-0663.87.2.218

Sigel, I. E., & McGillicuddy-Delisi (1984). Parents as teachers of their children: A distancing behavior model. In A. D. Pellegrini & T. D. Yawkey (Eds.), *The development of oral and written language in social contexts* (pp. 71–92). Norwood, NJ: Ablex.

Smith, M. W., & Dickinson, D. K. (1994). Describing oral language opportunities and environments in Head Start and other preschool classrooms. *Early Childhood Research Quarterly*, *9*, 345–366. https://doi.org/10.1016/0885-2006(94)90014-0

Snow, C. E., & Goldfield, B. A. (1983). Turn the page please: Situation-specific language acquisition. *Journal of Child Language*, *10*, 551–569. https://doi.org/10.1017/S0305000900005365

Sparks, A., & Reese, E. (2013). From reminiscing to reading: Home contributions to children's developing language and literacy in low-income families. *First Language, 33*, 89–109. https://doi.org/10.1177/0142723711433583

Stahl, S. A., & Fairbanks, M. M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research*, *56*, 72–110. https://doi.org/10.3102/00346543056001072

Storch, S. A., & Whitehurst, G. J. (2002). Oral language and code-related precursors to reading: Evidence from a longitudinal structural model. *Developmental Psychology, 38*, 934–947. https://doi.org/10.1037/0012-1649.38.6.934

Strasser, K., Larraín, A., & Lissi, M. R. (2013) Effects of storybook reading style on comprehension: The role of word elaboration and coherence questions. *Early Education and Development, 24*, 616–639. https://doi.org/10.1080/10409289.2012.715570

Suggate, S., Lenhard, W., Neudecker, E., & Schneider, W. (2013). Incidental vocabulary acquisition from stories: Second and fourth graders learn more from listening than reading. *First Language, 33*, 551–571. https://doi.org/10.1177/0142723713503144

Swingley, D. (2010). Fast mapping and slow mapping in children's word learning. *Language Learning and Development, 6*, 179–183. https://doi.org/10.1080/15475441.2010.484412

Taboada, A. (2012). Relationships of general vocabulary, science vocabulary, and student questioning with science comprehension in students with varying levels of English proficiency. *Instructional Science, 40*, 901–923. https://doi.org/10.1007/s11251-011-9196-z

Tompkins, V., Zucker, T. A., Justice, L. M., & Binici, S. (2013). Inferential talk during teacher–child interactions in small-group play. *Early Childhood Research Quarterly, 28*, 424–436. https://doi.org/10.1016/j.ecresq.2012.11.001

Trostle, S., & Hicks, S. J. (1998). The effects of storytelling versus story reading on comprehension and vocabulary knowledge of British primary school children. *Reading Improvement, 35*, 127–136.

Uchiyama, T. (2011). Reading versus telling of stories in the development of English vocabulary and comprehension in young second language learners. *Reading Improvement, 48*, 168–178.

Valdez-Menchaca, M. C., & Whitehurst, G. J. (1992). Accelerating language development through picture book reading: A systematic extension to Mexican day care. *Developmental Psychology, 28,* 1106–1114. https://doi.org/10.1037/0012-1649.28.6.1106

van Kleeck, A., Gillam, R. B., Hamilton, L., & McGrath, C. (1997). The relationship between middle-class parents' book-sharing discussion and their preschoolers' abstract language development. *Journal of Speech Language and Hearing Research*, *40*, 1261–1271. https://doi.org/10.1044/jslhr.4006.1261

van Kleeck, A., Vander Woude, J., & Hammett, L. (2006). Fostering literal and inferential language skills in Head Start preschoolers with language impairment using scripted book-sharing discussions. *American Journal of Speech-Language Pathology, 15*, 85–95. https://doi.org/10.1044/1058-0360(2006/009)

Walker, D., Greenwood, C., Hart, B., & Carta, J. (1994). Prediction of school outcomes based on early language production and socioeconomic factors. *Child Development, 65,* 606–621. https://doi.org/10.1111/j.1467-8624.1994.tb00771.x

Walsh, B. A., & Blewitt, P. (2006). The effect of questioning style during storybook reading on novel vocabulary acquisition of preschoolers. *Early Childhood Education Journal, 33,* 273–278. https://doi.org/10.1007/s10643-005-0052-0

Walsh, B. A., Sánchez, C., & Burnham, M. M. (2016). Shared storybook reading in Head Start: Impact of questioning styles on the vocabulary of Hispanic dual language learners. *Early Childhood Education Journal*, *44*, 263–273. https://doi.org/10.1007/s10643-015-0708-3

Walsh, R. L., & Hodge, K. A. (2018). Are we asking the right questions? An analysis of research on the effect of teachers' questioning on children's language during shared book reading with young children. *Journal of Early Childhood Literacy*, *18*, 264–294. https://doi.org/10.1177/1468798416659124

Wasik, B. A., Hindman, A. H., & Snell, E. K. (2016). Book reading and vocabulary development: A systematic review. *Early Childhood Research Quarterly*, *37*, 39–57. https://doi.org/10.1016/j.ecresq.2016.04.003

Wheeler, M. P. (1983). Context-related age changes in mothers' speech: joint book reading. *Journal of Child Language*, *10,* 259–263. https://doi.org/10.1017/S0305000900005304

Whitehurst, G. J., Arnold, D. S., Epstein, J. N., Angell, A. L., Smith, M., & Fischel, J. M. (1994). A picture book reading intervention in day care and home for children from low-income families. *Developmental Psychology*, *30*, 679–689. https://doi.org/10.1037/0012-1649.30.5.679

Whitehurst, G. J., Falco, F. L., Lonigan, C. J., Fischel, J. E., DeBaryshe, B. D., Valdez-Menchaca, M. C., & Caulfield, M. (1988). Accelerating language development through picture book reading. *Developmental Psychology*, *24*, 552–559. https://doi.org/10.1037/0012-1649.24.4.552

Wittrock, M. C. (2010). Learning as a generative process. *Educational Psychologist, 45*, 40–45. https://doi.org/10.1080/00461520903433554

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, *17*, 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Woolbert, C. H. (1922). Speaking and writing—A study of differences. *Quarterly Journal of Speech, 8*, 271–285. https://doi.org/10.1080/00335632209379390

Wright, A. (1995). *Storytelling with children*. Oxford, UK: Oxford University Press.

Zeece, P. D. (1997). Bringing books to life: Literature-based storytelling. *Early Childhood Education Journal, 25,* 39–43. https://doi.org/10.1023/A:1025633914983

Zucker, T. A., Justice, L. M., Piasta, S. B., & Kaderavek, J. N. (2010). Preschool teachers' literal and inferential questions and children's responses during whole-class shared reading. *Early Childhood Research Quarterly, 25*, 65–83. https://doi.org/10.1016/j.ecresq.2009.07.001

List of Figures

List of Tables

Study 4

Appendix

Appendix A: Material Study 1

Receptive target-word task (word recognition)

Als erstes habe ich dir wieder viele schöne Bilder mitgebracht. Willst du die Bilder mal anschauen?

Zum Einüben wird ein Übungsitem vorgegeben. – *Zeige mit deinem Finger auf* ***Boot****.*

Richtig gezeigt: *Das hast du toll gemacht.*

Nicht gezeigt oder falsch gezeigt: *Willst du es noch einmal probieren?* – Falls immer noch Probleme bestehen: Richtige Lösung zeigen und das Kind dazu bringen, auf das Boot zu zeigen.

Bei den folgenden Zielwörtern darf **nicht** mehr geholfen werden.

Table A1

Words used in the receptive target-word task of Experiment 1 (Study 1)

Word	Approximate English translation
Kanapee	settee
Zuber	tub
Balg	brat
Remise	remise
Trosse	hawser
Kleinod	bijou
Kardätsche	horse brush
Kumme	bowl
Kladde	waste book
Kofel	mound
Klampfe	guitar
Pfuhl	puddle

Table A2

Words used in the receptive target-word task of Experiment 2 (Study 1)

Word	**Approximate English translation**
Kodex	codex
Fackel	torch
Zähre	tear
Suhle	wallow
Gestrüpp	thicket
Illusionist	illusionist
Barke	skiff
Oheim	uncle
Gewand	garment
Huf	hoof
Reisig	brushwood
Behausung	dwelling

Observer questionnaire

Beobachterin______________________________ **Erzählerin**_____________________

Versuchsnummer (d.h. Set) _______________ **Kind**________________________

Dauer der Erzählung (Min:Sek)______:_______

Art der Erzählung: ○ Vorlesen aus Text ○ Mündliche Erzählung

Geschichte 1	Anzahl	**Geschichte 2**	Anzahl	**Geschichte 3**	Anzahl
Kofel		Kleinod		Kanapee	
Trosse		Kumme		Klampfe	

Geschichte 4	Anzahl	**Geschichte 5**	Anzahl	**Geschichte 6**	Anzahl
Kardätsche		Balg		Remise	
Zuber		Pfuhl		Kladde	

Verhalten des Erzählers/der Erzählerin

Wie oft wurden Gesten verwendet?

nie	selten	manchmal	oft	immer

Wie oft wurde Stimmmodulation verwendet?

nie	selten	manchmal	oft	immer

Wie lang waren die Sätze?

sehr kurz	kurz	mittel	lang	sehr lang

Wie schnell wurde gesprochen?

sehr langsam	langsam	mittel	schnell	sehr schnell

Wie viel Augenkontakt wurde hergestellt?

gar nicht	wenig	mittel	oft	ständig

Verhalten des Kindes

Wie viele Fragen hat das Kind gestellt bzw. wie oft hat es die Erzählung unterbrochen?

gar nicht	einmal	zwei- oder dreimal	vier- oder fünfmal	mehr als fünfmal

Welche Wortnachfragen hat das Kind gestellt (Wörter notieren)?

Hat das Kind eines der beiden unbekannten Wörter laut wiederholt (auch Häufigkeit notieren)?

Wie hat sich das Kind verhalten?

sehr ruhig	ruhig	mittel	unruhig	sehr unruhig

Appendix B: Material Study 2

Receptive target-word task (word recognition)

Als erstes habe ich dir wieder viele schöne Bilder mitgebracht. Willst du die Bilder mal anschauen?

Zum Einüben wird ein Übungsitem (Ü1) vorgegeben. – *Zeige mit deinem Finger auf* ***Junge****.*

Richtig gezeigt: *Das hast du toll gemacht.*

Nicht gezeigt oder falsch gezeigt: *Willst du es noch einmal probieren?* – Falls immer noch Probleme bestehen: Richtige Lösung zeigen und das Kind dazu bringen, auf den Jungen zu zeigen.

Table B1

Words used in the receptive target-word task (Study 2)

Word	**Approximate English translation**
Target words	
Kodex	codex
Fackel	torch
Zähre	tear
Suhle	wallow
Gestrüpp	thicket
Flagge	flag
Barke	skiff
Remise	remise
Bühne	stage
Huf	hoof
Beifall	applause
Gewand	garment
Bräutigam	groom
Robe	gown
Klampfe	guitar
Zuber	tub
Control words	
Mädchen	girl
Junge	boy
Fahrrad	bicycle
Stuhl	chair

Story comprehension task

Instruktion: Frage vorlesen und auf Antwort des Kindes warten. Wenn Kind etwas antwortet, nochmal nachhaken *(Fällt dir noch etwas dazu ein? Willst du noch etwas dazu sagen?)*. Wenn ein Kind nichts sagt, nochmal nachhaken *(Fällt dir irgendetwas dazu ein? Hast du irgendeine Idee? Trau dich ruhig!)*.

Wichtig:

- Diese Schritte befolgen
- Empathisch zuhören
- Keine Rückmeldung über Richtigkeit
- Nicht zu lange auf eine Antwort warten, Kinder nicht frustrieren

Geschichte 1: *Du erinnerst dich doch bestimmt noch an die Geschichte mit dem Elefanten, dem Affen, dem Hasen und der Schnecke? Ich hab da ein paar Sachen vergessen, kannst du mir helfen?*

1. Was kann denn der Hase besonders gut?
2. Was kann denn der Affe besonders gut?
3. Warum reißt der Elefant den Baum, auf dem der Affe klettert, raus?

Geschichte 2: *Du erinnerst dich doch bestimmt noch an die Geschichte mit dem kleinen Simsalino? Ich hab da ein paar Sachen vergessen, kannst du mir helfen?*

1. Wo muss der kleine Simsalino stehen, wenn der große Simsalo vor dem Publikum zaubert?
2. Was macht der kleine Simsalino mit dem großen Simsalo?
3. Warum darf der kleine Simsalino nicht zusammen mit dem großen Simsalo zaubern?

Geschichte 3: *Du erinnerst dich doch bestimmt noch an die Geschichte mit den Piraten. Ich hab da ein paar Sachen vergessen, kannst du mir helfen?*

1. Vor wem verstecken sich die Piraten denn?
2. Wovor haben die Piraten solche Angst?
3. Warum müssen sich die Piraten baden?

Geschichte 4: *Du erinnerst dich doch bestimmt noch an die Geschichte mit den Schweinen Ringelschwänzchen und Schwarte. Ich hab da ein paar Sachen vergessen, kannst du mir helfen?*

1. Warum spritzt Schwarte die Gäste mit Wasser ab?
2. Was macht Schwarte mit der Farbe?
3. Was macht der Regen mit der Kleidung der Hochzeitsgäste?

Observer qestionnaire

BeobachterIn ______________________	**ErzählerIn** ______________________	
Datum __________	**Bedingung: ___vorgelesen**	**___ frei erzählt**

Verhalten des Erzählers/der Erzählerin

Wie oft wurden Gesten verwendet?

nie	selten	manchmal	oft	sehr oft

Wie variierend war die Stimme (Dynamik, Intonation, Intensität)?

sehr monoton	monoton	mittel	variierend	sehr variierend

Wie verständlich wurde gesprochen? (Aussprache, Lautstärke)

sehr unverständlich	unverständlich	mittel	verständlich	sehr verständlich

Wie viel Augenkontakt hat der/die ErzählerIn zu den Kindern hergestellt?

nie	selten	manchmal	oft	sehr oft

Verhalten der Kinder

Wie **motorisch** unruhig waren die Kinder während der Geschichte?

sehr unruhig	unruhig	mittel	ruhig	sehr ruhig

Wie **aufmerksam/interessiert** waren die Kinder bei der Geschichte?

sehr unaufmerksam	unaufmerksam	mittel	aufmerksam	sehr aufmerksam

Wie sehr haben sich die Kinder **aktiv** an der Geschichte (durch Fragen, Kommentare) beteiligt?

gar nicht	kaum	manchmal	oft	sehr oft

Appendix C: Material Study 3

Expressive target-word task (picture naming)

Ich zeige dir jetzt ein paar Bilder. Ich möchte, dass du mir sagst, welchen Gegenstand du darauf siehst oder was der Mensch oder das Tier auf dem Bild macht.

Besonderheit bei Verben: Hier fragen: *Was macht/machen XY auf dem Bild?*

Kind benennt ein anderes Objekt auf dem Bild: Benennt das Kind ein anderes Objekt, das ebenfalls auf dem Bild abgebildet ist (z.B. bei Atoll das Meer anstelle des Atolls), dann auf den Ziel-Gegenstand (Atoll) deuten und noch einmal nachfragen.

Kind benutzt eine andere Bezeichnung für den Ziel-Gegenstand: Benutzt das Kind eine andere Bezeichnung für den Ziel-Gegenstand, dann noch einmal unter Verweis auf das in der Geschichte dafür verwendete Wort nachfragen. – *Ja das stimmt, aber erinnerst du dich noch an das Wort, das ich in der Geschichte dafür benutzt habe?*

Expressive target-word task (word definition)

Als erstes habe ich mit dir ein Erklär-Spiel vor. Ich frage dich immer etwas und du versuchst mir das zu erklären. Ok? Das ist ziemlich schwierig, daher ist es auch gar nicht schlimm, wenn du mir ein paar Dinge nicht erklären kannst. Fangen wir an.

Receptive target-word task (word recognition)

Als nächstes habe ich dir viele schöne Bilder mitgebracht. Willst du die Bilder mal anschauen?

Übungsitem:
Zum Einüben wird ein Übungsitem vorgegeben. – *Zeige mit deinem Finger auf* ***Mädchen****.*
Nicht gezeigt oder falsch gezeigt: *Willst du es noch einmal probieren?* – Falls immer noch Probleme bestehen: Richtige Lösung zeigen und das Kind dazu bringen, auf das Mädchen zu zeigen.

Testitems:
Bei den folgenden Zielwörtern darf **nicht** mehr geholfen werden. Es soll auch **nicht** gesagt werden, ob die Antwort des Kindes richtig oder falsch war. Auch bei Nachfrage mit einem **neutralen** Ausspruch reagieren, z.B. „Das machst du gut.“. Korrigiert sich das Kind selbst, gilt die letzte Bildauswahl des Kindes.

Jetzt hab ich noch ein paar Bilder mitgebracht. Wir machen das wieder genauso wie gerade eben. Ich sage dir ein Wort und du deutest mit deinem Finger auf das passende Bild.

Table C1

Words used in the target-word tasks (Study 3)

Word	**Approximate English translation**
Target words	
anketten	to chain
Atoll	atoll
aushändigen	to hand (something)
Barke	skiff
Besserwisser	know-all
Buddel	bottle
Eisbrecher	icebreaker
Geäst	branches
Hauer	fang
Heimstätte	homestead
Kanapee	settee
Katapult	catapult
Klaue	claw
Korkenzieher	corkscrew
Leu	lion
Lurch	amphibian
Ödland	wasteland
Pfuhl	puddle
Reisig	brushwood
Schemen	spectre
schlummern	to doze
skandieren	to chant
(sich) suhlen	to wallow

zanken	to squabble

Control words

(sich) amüsieren	to enjoy oneself
Beistand	support
hechten	to dive headlong
Robe	gown
säubern	to cleanse
Schwätzchen	chin wag
sinnieren	to muse
Sud	brew
trällern	to warble
Tümmler	porpoise
Warze	wart
zusammenbrauen	to concoct

Appendix D: Material Study 4

Expressive target-word task (word definition)

<u>Vorgehen:</u> Jeweils die angegebene Frage stellen. Falls das Kind nichts sagt, nachfragen und ermuntern: *„Hast du irgendeine Idee, was das sein/bedeuten könnte? Trau dich ruhig!“* Wenn das Kind etwas gesagt hat, dann in einem zweiten Schritt nachfragen: *„Fällt dir noch etwas dazu ein?“*

Als erstes habe ich mit dir ein Erklär-Spiel vor. Ich frage dich immer etwas und du versuchst mir das zu erklären. Ok? Das ist ziemlich schwierig, daher ist es auch gar nicht schlimm, wenn du mir ein paar Dinge nicht erklären kannst.

Receptive target-word task (word recognition)

Als nächstes habe ich dir viele schöne Bilder mitgebracht. Willst du die Bilder mal anschauen?

<u>Übungsitem:</u>
Zum Einüben wird ein Übungsitem vorgegeben. – *Zeige mit deinem Finger auf* ***Mädchen****.*

Nicht gezeigt oder falsch gezeigt: *Willst du es noch einmal probieren?* – Falls immer noch Probleme bestehen: Richtige Lösung zeigen und das Kind dazu bringen, auf das Mädchen zu zeigen.

<u>Testitems:</u>

- Bei den folgenden Zielwörtern darf **nicht** mehr geholfen werden.
- Es soll auch **nicht** gesagt werden, ob die Antwort des Kindes richtig oder falsch war. Auch bei Nachfrage mit einem **neutralen** Ausspruch reagieren, z.B. „Das machst du gut.“
- Korrigiert sich das Kind selbst, gilt die letzte Bildauswahl des Kindes.
- Das Zielwort soll **ohne Artikel** dargeboten werden („Zeig mir mal Mädchen.“).

Jetzt hab ich noch ein paar Bilder mitgebracht. Wir machen das wieder genauso wie gerade eben. Ich sage dir ein Wort und du deutest mit deinem Finger auf das passende Bild.

Table D1

Words used in the target-word tasks (Study 4)

Word	**Approximate English translation**
Target words	
aushändigen	to hand (something)
Besserwisser	know-all
Hauer	fang
Lamäng	hand
Stoppelzieher	corkscrew
Leu	lion
Lurch	amphibian
Ödland	wasteland
Pfuhl	puddle
Reisig	brushwood
schlummern	to doze
(sich) suhlen	to wallow
Control words	
Robe	gown
sinnieren	to muse
Sud	brew
trällern	to warble
Zinken	(big) nose
zusammenbrauen	to concoct