Reading and Writing
https://doi.org/10.1007/s11145-023-10463-x

Stimulating inference-making in second grade children when reading and listening to narrative texts

Brechtje E. J. van Zeijts[1] · Lesya Y. Ganushchak[1] · Bjorn B. de Koning[1] · Huib K. Tabbers[1]

Accepted: 13 June 2023

Abstract

Inference-making is a central element of successful reading comprehension, yet provides a challenge for beginning readers. Text decoding takes up cognitive resources which prevents beginning readers from successful inference-making and compromises reading comprehension. Listening does not require any decoding and could therefore offer a less demanding context to practice inference-making. The present study examined whether stimulating inference-making in a listening context is more effective and less cognitively demanding for beginning readers than a reading context. In three experiments, Dutch second grade children read two narratives and listened to two narratives. Inference-making was stimulated by asking them inferential questions during reading or listening and we compared this to a no-questioning control condition. After each narrative, we measured cognitive load and comprehension. It was expected that inferential questioning would increase cognitive load and negatively affect reading comprehension, but positively affect listening comprehension. The results indeed showed that inferential questioning increased cognitive load, but did not lead to differences in performance on open-ended comprehension questions (Experiment 1 & 2). When measuring comprehension with a free recall protocol (Experiment 3), we found a negative effect on total recall in both the reading and listening conditions. Taken together, we found no support for the hypothesized interaction. This raises questions about the effectiveness of inferential questioning for reading and listening comprehension of beginning readers, and whether listening is a good modality for improving inference-making.

Keywords Reading comprehension · Listening comprehension · Inference-making · Narrative comprehension · Inferential questioning · Primary school

✉ Brechtje E. J. van Zeijts
vanzeijts@essb.eur.nl

[1] Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, Burgemeester Oudlaan 50, Rotterdam 3062 PA, The Netherlands

Published online: 28 July 2023

Introduction

Many children struggle with reading comprehension and fail to reach a sufficient level of comprehension by the end of primary school (Dutch Inspectorate of Education, 2020; National Assessment of Educational Progress, 2019). A central element of successful reading comprehension is the ability to draw inferences (Kendeou et al., 2016). Inference-making involves identifying implicit relations between parts of the text, and between the text and background knowledge (Cain & Oakhill, 1999). Research indicates that inference-making is especially challenging for younger children, which is illustrated by the fact that older children generate more, and a greater variety of inferences (van den Broek et al., 2013). One of the reasons why younger children experience difficulties with inference-making is related to the availability of cognitive resources. Due to automatization of lower-level reading skills (e.g., decoding skills), older children and adult readers have more cognitive resources available for higher-level comprehension processes, such as inference-making (Kendeou et al., 2014). Beginning readers, on the other hand, need their cognitive resources for decoding individual words in addition to comprehension-related processes. This may hinder inference-making and can seriously compromise text comprehension. The present study therefore investigates whether listening to a narrative may provide beginning readers with a suitable context for inference-making that is less cognitively demanding than reading a narrative.

The importance of inference-making for comprehension

Cognitive theories of reading comprehension state that to comprehend a text, readers must construct a coherent mental representation of the text (McNamara & Magliano, 2009). The construction of a mental representation involves identifying relations between various pieces of information in the text, and between information in the text and the reader's background knowledge (Kintsch & van Dijk, 1978). These relations are often inferential, meaning that readers must fill in information that is not explicitly stated in the text (Oakhill et al., 2003). The ability to draw inferences is considered the cornerstone of language comprehension (Kendeou et al., 2016), and has been shown to be highly predictive of reading comprehension (Cain & Oakhill, 1999).

Inference-making is not only relevant for comprehension of written text, but also for comprehension in non-reading contexts, such as films and audio-materials (e.g., Florit et al., 2011; Magliano et al., 2013; Tibus et al., 2013). Several researchers advocated that higher-level comprehension processes are modality-independent and therefore similar across different media (Gernsbacher et al., 1990; Kendeou et al., 2014). For example, the inferential Language Comprehension (iLC) framework states that language comprehension depends on a general inference skill that can transfer across different contexts and media (Kendeou et al., 2020). This view is empirically supported by longitudinal research indicating that comprehension skills in non-reading contexts in preschool and early grades are predictive of later reading

comprehension (Catts et al., 2015; de Jong and van der Leij, 2002; Kendeou et al., 2005; Storch and Whitehurst, 2002). Specifically for inference-making, it was found that the inferential processes of 4-year-olds listening to spoken narratives or watching televised stories are very similar to those of older children when reading (Kendeou et al., 2008). These findings suggest that the development of inference-making skills can be supported at a young age using non-written media, like audio materials, and thereby prepare children for reading comprehension (Kendeou et al., 2014, 2020).

Despite the similarities in comprehension processes across different media, an important difference between reading contexts and non-reading contexts is that reading comprehension not only requires higher-level comprehension skills such as inference-making, but also basic reading skills (Simple View of Reading, Hoover and Gough, 1990). Basic reading skills such as text decoding are crucial for converting written letters or words into sounds. Hence, reading comprehension imposes an additional challenge on beginning readers. Indeed, research indicates that in the first grades of primary school the level of listening comprehension is often higher than the level of reading comprehension (Diakidoy et al., 2005; Verlaan et al., 2017). Once decoding becomes more automatized, the level of reading comprehension becomes comparable to the level of listening comprehension. This indicates that for beginning readers, listening comprehension is less cognitively demanding, as understanding spoken text does not require any text decoding and this allows children to focus more on comprehension.

Stimulating children's inference-making

While beginning readers are able to make inferences (Kendeou et al., 2008; van den Broek et al., 1996), they are less likely to do so spontaneously compared to older readers (Oakhill et al., 2003). To stimulate children to make inferences during reading, researchers have evaluated various interventions and strategies that, in general, appear to be effective in improving inferencing skills and reading comprehension (for review see Elleman, 2017; Hall, 2016). However, a meta-analysis on the effectiveness of reading comprehension strategies reported lower effect sizes for elementary school children compared to middle school and high school students (on researcher-developed tests, Berkeley et al., 2010). Similarly, some studies that stimulated inference-making via in-text prompting also showed negative effects in beginning readers. For example, van den Broek et al. (2001) looked at the effect of answering inferential questions during reading in different age groups (i.e., fourth, seventh, and tenth graders and college students). They found that asking inferential questions during reading resulted in lower comprehension scores for fourth graders, whereas for proficient readers (i.e., college students), this questioning technique improved comprehension scores. Van den Broek and colleagues suggested that the questioning strategy was ineffective for beginning readers because it had interfered with their inference-making due to additional cognitive demands. For example, the questioning strategy requires attention switching between tasks. This suggests that

particularly beginning readers might benefit from stimulating inference-making in a non-reading context, because this does not require decoding skills.

While most research on stimulating inference-making has focused on comprehension of written text (Elleman, 2017), there is an increasing number of studies exploring the idea of practicing comprehension using other media than text, which have shown promising results. For example, some intervention studies reported a positive effect of audio-based training programs on reading comprehension outcomes (Aarnoutse et al., 1998; Brand-Gruwel et al., 1998; Carretti et al., 2014; Clarke et al., 2010). Other studies investigated the effect of asking inferential questions while listening to a spoken narrative among kindergartners (Butterfuss et al., 2021), toddlers and third graders (van den Broek et al., 2011), and third and fifth grade children (Freed & Cain, 2017). According to the iLC framework, asking questions can prompt inference-making because it activates the information needed to draw an inference and facilitates integration. Indeed, these studies found that asking inferential questions *during* listening supported narrative comprehension and resulted in higher comprehension outcomes than asking the same questions *after* listening. These studies, however, did not include a no-questioning control condition and therefore do not provide information about whether stimulating inference-making during listening is more effective than not asking inference stimulating questions. Neither did these studies compare the listening condition to a reading condition, to check whether the inferential questions benefitted comprehension only with spoken text, and not with written text. Therefore, to obtain stronger evidence on whether inference-making in young children can indeed be stimulated by using spoken text, it is important to evaluate the effect of inferential questions in both a reading and a listening context. The direct comparison of the effect of inference-stimulation in a reading context and a listening context, and inclusion of a no-questioning control group for both modalities were addressed in the present study.

The present study

The goal of this study was to examine whether stimulating inference-making in a listening context is more effective and less cognitively demanding for beginning readers than in a reading context. After reading and listening to narrative texts, we measured children's comprehension and their cognitive load. Cognitive load refers to the load that a certain task imposes on a child's cognitive system (Cognitive Load Theory, Paas et al., 2019). This is measured using a mental effort rating scale (Paas, 1992). Mental effort is the amount of cognitive processing that a child invests in a task. The participants in this study were Dutch second grade children; second graders' decoding skills are still developing (Verhoeven & van Leeuwe, 2009), so they can be considered a group of beginning readers. Inference-making was stimulated by asking children inferential questions during reading and during listening, specifically focused on important causal connections (Butterfuss et al., 2021; van den Broek et al., 2001; van den Broek et al., 2011). Causal inferences, which identify how events or facts lead to or depend on each other, are especially important for narrative comprehension (Trabasso & van den Broek, 1985). Narratives often are structured as a sequence of causally related events

(Best et al., 2008). Asking questions during comprehension can direct children's attention to important causal connections in the text and hereby stimulate inference-making. This way, the questioning technique can support the construction of a coherent mental representation and thus improve narrative comprehension.

We report the results of three experiments. In Experiment 1, we evaluated the effect of asking inferential questions during reading and during listening with second grade children. In Experiment 2, we replicated Experiment 1 with a slightly younger sample that had lower decoding skills. In Experiment 3, we again built on Experiment 1 using a different design and a different comprehension measure.

Experiment 1

Experiment 1 investigated the effect of modality (reading vs. listening) and stimulating inference-making (inferential questioning vs. no inferential questioning) on comprehension outcomes and cognitive load. Children read two narratives and listened to two narratives. For half of the narratives, inference-making was stimulated by asking questions that targeted important causal relations. After each narrative, we measured comprehension using questions targeting the causal relations in the text, and we measured cognitive load using a mental effort rating scale.

Previous studies found that in second grade, children's level of listening comprehension is higher than their level of reading comprehension (Diakidoy et al., 2005; Verlaan et al., 2017). Therefore, we expected higher comprehension scores in the listening conditions than in the reading conditions (Hypothesis 1). We expected the effect of stimulating inference-making on comprehension to depend on modality. Simultaneously decoding text, attempting to understand the text, and answering questions during reading might be too demanding for second grade children. Hence, we expected that with written text, questioning would negatively affect comprehension (van den Broek et al., 2001; Hypothesis 2a). Alternatively, listening does not require decoding, which allows children to focus on comprehension. Asking inferential questioning during listening might direct attention to causal connections and stimulate inference-making resulting in better comprehension (Butterfuss et al., 2021; Freed & Cain, 2017; van den Broek et al., 2011). Therefore, we expected that with spoken text, questioning would positively affect comprehension (Hypothesis 2b). Regarding cognitive load, we hypothesized that reading results in higher cognitive load than listening (Hypothesis 3). Reading is expected to be more demanding because it requires text decoding. This might be extra challenging as our participants were beginning readers. Finally, we expected that stimulating inference-making through inferential questioning leads to higher cognitive load (Hypothesis 4), because of the additional task of answering questions.

Method

Participants and design

Participants were 43 second grade children from two elementary schools in the Netherlands. Data from three children were excluded, because of missing answers on the comprehension questions, resulting in a sample of 40 children (19 boys and 21 girls), aged 7 to 8 years old ($M = 8.11$, $SD = 0.36$). Before testing, children's parents or caretakers signed an informed consent letter. A 2×2 within-subjects design was used, with modality (reading vs. listening) and inference-stimulation (questioning vs. no-questioning) as independent variables, and comprehension and cognitive load as dependent variables. Thus, children participated in all four experimental conditions. In each of the resulting within-subjects conditions, a different narrative was presented. A Latin square design was used to counterbalance both condition order and assignment of narratives to conditions (Zeelenberg & Pecher, 2015), resulting in eight counterbalance conditions to which children were randomly assigned.

Materials

Narrative texts and audio-recordings

Four narrative texts from the studies of Kraal et al. (2018; 2019) were used (see Supplementary Material). The texts were about a little mouse saving his village from a cat (text 1), elephant Okke wanting to lose weight (text 2), two brothers encountering a problem with their sister's iPad (text 3), and children playing hide-and-seek in the schoolyard (text 4). Because these texts were developed for the Kraal et al. studies they were (1) unfamiliar to our participants and (2) tested before in a sample of Dutch second grade children. Using P-CLIB software (Evers, 2008), each text's level of decoding and comprehension difficulty was determined at second grade level.

In the reading conditions, the narratives were presented as written text on a computer screen, in Arial font size 12. The texts contained 117 to 169 words and were 19 to 25 sentences long. In the listening conditions, the narratives were presented as audio recordings. A control panel was presented on the computer screen, with a pause button and a slider to go back and forward in the audio-recording to mimic the reread possibilities that children have during reading. Each text was narrated and recorded by the first author (female native speaker of Dutch), resulting in audio-recordings of 50 to 69 s long. In both the reading and listening conditions, the narratives were divided into four or five fragments. Fragment length varied between one and ten sentences. Each fragment required at least one inference for understanding the story.

For the questioning conditions, two additional narratives were used to allow children to practice the procedure of answering questions during reading or listening. One practice narrative, based on a chapter in the book 'Hare and toad are moving' (in Dutch: *Haas en pad gaan verhuizen*, Bergsma, 2009), was presented as written

text. This narrative was divided into two fragments of both three sentences long. The other practice narrative was presented as an audio-recording, again narrated by the first author. This narrative was based on the book 'Bite the thief in his bum' (in Dutch: *Bijt de dief in zijn billen!*, Noort, 2015). This narrative was 17 s long and divided into three fragments, each two sentences long. Both practice narratives were written at second grade level.

Inference-stimulating questions

In the questioning conditions, inferential questions were asked during reading and listening (see Supplementary Material). Each narrative included four open-ended questions prompting both text-connecting and gap-filling inferences. The inference-stimulating questions were meant to direct children's attention to these important implicit connections in the text (Butterfuss et al., 2021; van den Broek et al., 2001; van den Broek et al., 2011). For example, when reading or listening to the narrative Hide-and-seek, children were asked 'Why do you think he climbed that high tree?'. This should stimulate them to infer that the tree is a good hiding spot. After a fragment, children clicked to the next page and were presented with an alarm clock and the text 'It's time for a question!'. They were then not able to go back in the narrative (van den Broek et al., 2001; van den Broek et al., 2011). The inferential questions addressed the same implicit relations as our comprehension test questions, but the questions were worded differently to prevent a practice effect. For example, in the iPad narrative, two brothers are playing with their sister's iPad when suddenly the screen turns black. The inferential question 'What do you think is wrong with the iPad?' is targeting the same causal connection as the comprehension question 'Why do Simon and Tom think the iPad is broken?'.

Children's answers to the inferential questions were recorded and afterwards transcribed. Answers were scored as correct (1 point), half correct (0.5 points), or incorrect (0 points) using an answer model by two independent raters. One point was given when a child explicitly mentioned the full inference in their answer. Half a point was given when a child only provided part of the inference, or provided relevant information without making an explicit inference. Children's scores were summed for each narrative, resulting in an overall inference score (0–4) per narrative. Both raters scored the answers of all forty children. As for all tasks involving multiple raters reported in this manuscript, the inter-rater reliability was computed using a two-way mixed, absolute, single-measures intra-class correlation (ICC; Hallgren, 2012), which showed high inter-rater agreement (ICC = 0.95; range separate narratives = 0.92-0.95).

Comprehension test

Comprehension of our four narratives was measured with six open-ended questions per narrative taken from the studies of Kraal et al. (2018; 2019), see Supplementary

Material. The questions were slightly adapted to ensure that each narrative had two factual questions[1], three text-connecting inference questions, and one gap-filling inference question (Cain & Oakhill, 1999). Factual questions asked for information literally stated in the text. Text-connecting inference questions required children to combine parts of information in the text. Gap-filling inference questions required children to combine information in the text with background knowledge.

The questions were asked by the test leader and children's verbal responses were recorded and afterwards transcribed. Children did not receive feedback on their answers, only neutral prompts and encouragements were given. When a participant did not respond or did not understand a question, the question was repeated once. Similar to the inference-stimulating questions, all responses were scored by two independent raters as correct (1 point), half correct (0.5 points), or incorrect (0 points) using an answer model. For example, one text-connecting inference question about the narrative Hide-and-seek was 'How come Stef saw Luuk?'. The correct answer was that Stef saw a girl looking up to Luuk, who was hiding in a tree. A half-correct answer could be that the girl betrayed Luuk's hiding place, without explaining how exactly. An incorrect answer could be that Stef just looked up and saw Luuk in the tree. Per narrative separate scores were calculated for factual questions (0–2) and comprehension questions (0–4). One rater scored the answers of all forty children, a second rater scored the answers of ten randomly selected children. Inter-rater reliability was high (ICC = 0.89; range separate narratives = 0.82-0.98).

Cognitive load

Cognitive load was measured with the question "How much effort did you invest in understanding the narrative?". Children indicated their invested mental effort on a rating scale (Fig. 1), ranging from 1 (i.e., very little) to 9 (i.e., very much), based on Paas (1992). This mental effort rating has been shown to provide a reliable and valid measure of cognitive load during a task (Paas et al., 2003). Based on Laurie-Rose et al. (2014), we made the scale more age-appropriate. That is, the left endpoint of this scale showed a picture of a girl smiling and seeming to invest very little mental effort, while the right endpoint showed a picture of a girl looking very concentrated and seeming to invest much mental effort. Moreover, the rating scale was presented on the computer and offered a movable pointer, because providing a numerical response to the mental effort question is too complicated for second grade children (Laurie-Rose et al., 2014).

[1] The factual questions were part of the comprehension test that we adopted from the studies of Kraal et al. (2017) and Kraal et al. (2018). We will report mean scores and standard deviations for factual questions and inference-questions in the Results section. But in the analyses, we only used the four inference questions, as children's memory of factual information is not the aim of this study. We will refer to the four inference-questions as 'comprehension questions' to avoid confusion with the inference-stimulating questions asked during the narrative.

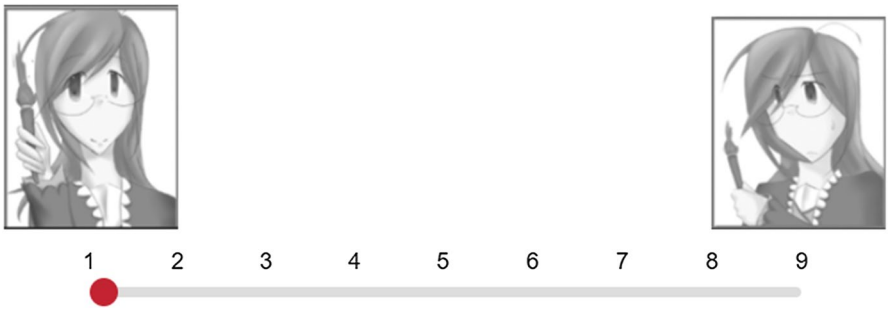

Fig. 1 The 9-point mental effort rating scale with an anime drawing at both endpoints

Decoding skills

Decoding skills were assessed using a Dutch standardized test called the Three Minutes Test (in Dutch: *Drie Minuten Toets*, van Til et al., 2018). The Three Minutes Test measures the speed and accuracy of reading individual words. Children read aloud as many unrelated words as possible in three minutes, focusing on both speed and accuracy. The more words children read correctly within the allotted time, the better their decoding skills. This standardized test was administered by the participating schools. The two participating schools had used different versions of the test (i.e., 2009 and 2018 version). The reliability of both versions is high, with a reliability coefficient of 0.97 for second grade children (Krom et al., 2010; Til et al., 2018). To enable comparisons of both versions, children's DLE-scores were used (i.e., didactic age equivalent, in Dutch: *didactische leeftijdsequivalent*, Melis et al., 2012). A DLE-score represents a child's level on a certain skill in months and provides information on how she/he performs compared to the norm. Each 10-month schoolyear stands for 10 DLE points (1 point per month), so when completing second grade, the average student should have a DLE-score of 20.

Procedure

The standardized decoding test was administered by the participating schools in May or June, which was in the same period as our experiment, which took place in May. For our experiment, all children were tested individually during school hours by the first author, in a separate room in the participating school. Children were seated behind a laptop on which the texts and audio-recordings were presented. They were informed that the test would take approximately 25 min and that the session would be audio-recorded. First, children were presented with two practice narratives. The first practice narrative was presented as written text and children were instructed to read the text aloud at their own pace. The second practice narrative was presented in audio-form and the test leader explained how to pause and go back- and forward. After each fragment, an inference-stimulating question was asked by the test leader. Children were instructed to answer out-loud, without time limits. After a narrative,

Table 1 Mean scores (and SD) on factual questions, comprehension questions, and cognitive load in each condition

	Reading		Listening	
	With questions	Without questions	With questions	Without questions
Factual (0–2)	1.85 (0.36)	1.95 (0.22)	1.68 (0.53)	1.74 (0.54)
Comprehension (0–4)	2.68 (0.81)	2.80 (0.85)	2.69 (0.79)	2.54 (0.94)
Cognitive Load (1–9)	2.18 (1.17)	1.98 (1.31)	2.53 (1.72)	1.80 (1.22)

children filled in the mental effort scale and the test leader checked whether the child had understood the rating scale.

Next, children were consecutively presented with the four experimental narratives. The order of narratives and conditions depended on the counterbalance condition. After each narrative, children filled in the mental effort rating scale and the test leader asked the open-ended comprehension questions for that narrative. Finally, children were thanked for their participation and informed not to talk with their classmates to avoid that they shared relevant information about the narratives.

Results and discussion

Before testing our hypotheses, children's DLE-scores on the standardized decoding test were analyzed. On average, children's DLE scores were considerably higher ($M = 22.83$, $SD = 14.75$; range: 7 to 55) than expected based on the months of education received (18 months) and thus were more representative for the beginning of third grade. Next, we analyzed children's performance on the inferential questions in the questioning conditions. The average score on the inferential questions was 2.91 (out of 4), and there was no significant difference between questions asked during reading ($M = 3.08$, $SD = 0.90$) and during listening ($M = 2.74$, $SD = 1.28$), $t(39) = 1.18$, $p = .247$, $d = 0.19$. So, in both conditions, children answered most inferential questions correctly.

Table 1 presents the mean scores and standard deviations of the factual questions, comprehension questions, and cognitive load in each condition. To test our hypotheses, a repeated measures ANOVA was performed on the comprehension scores with modality and inference-stimulation as within-subject factors. Contrary to expectations, comprehension scores in the reading conditions ($M = 2.74$, $SD = 0.10$) and the listening conditions ($M = 2.61$, $SD = 0.12$) did not differ significantly, $F(1, 39) = 0.81$, $p = .374$, $n_p^2 = 0.02$. There was also no significant main effect of stimulating inference-making, $F(1, 39) = 0.01$, $p = .914$, $n_p^2 < 0.01$, which means that comprehension scores in the questioning conditions ($M = 2.68$, $SD = 0.10$) and the no-questioning conditions ($M = 2.67$, $SD = 0.11$) were comparable. There was no significant modality × questioning interaction, $F(1, 39) = 1.84$, $p = .182$, $n_p^2 = 0.05$. Regarding cognitive load, a repeated measures ANOVA was performed with modality and inference-stimulation as within-subject factors. As expected, there was

a significant effect of stimulating inference-making, $F(1, 39) = 11.84$, $p = .001$, $n_p^2 = 0.23$, indicating that inferential questioning ($M = 2.35$, $SD = 0.19$) resulted in higher mental effort ratings than no inferential questioning ($M = 1.89$, $SD = 0.17$). Cognitive load in the reading conditions ($M = 2.08$, $SD = 0.18$) and the listening conditions ($M = 2.16$, $SD = 0.21$) did not significantly differ, $F(1, 39) = 0.22$, $p = .645$, $n_p^2 < 0.01$. There was no significant modality × questioning interaction, $F(1, 39) = 3.36$, $p = .074$, $n_p^2 = 0.08$.

Together, contrary to our expectations, the results showed no difference between the reading and listening conditions in terms of both comprehension outcomes and cognitive load (Hypothesis 1 & 3). As expected, inferential questioning increased cognitive load (Hypothesis 4), but contrary to our expectations, this increased cognitive load did not affect reading or listening comprehension (Hypothesis 2a & 2b). Our expectations were based on the assumption that the sample would consist of beginning readers whose decoding skills are not yet automated. However, children's decoding skills appeared to be at a level that is normally obtained early in Grade 3, and we suspect that this is because they were tested at the end of second grade. This means that the narratives, which were written at second grade level, were relatively easy for the children. Therefore, the absence of a difference between modalities on comprehension might be due to children's relatively well-developed decoding skills which made that reading the texts did not require much effort. This is confirmed by the relatively low mental effort ratings in all four experimental conditions (Table 1). The absence of a negative effect of inference stimulation on reading comprehension can similarly be explained with this decoding skill-mental effort relation. Based on this, a logical next step was to test our hypotheses on a sample of younger children, whose decoding skills were at a lower level.

Experiment 2

Experiment 2 exactly replicated Experiment 1 with a younger sample (i.e., children tested early in Grade 2), to ensure children had relatively low decoding skills.

Method

Participants and design

Participants were 22 second grade children, all from the same elementary school in the Netherlands. Data from four children were excluded, because of missing answers on the comprehension questions, resulting in a sample of 18 children (15 boys and 3 girls). Children were between 7 and 8 years old ($M = 7.78$, $SD = 0.43$). Parents or caretakers of the children signed an informed consent letter. The same within-subjects design was used as in Experiment 1, with modality (reading vs. listening) and questioning (questioning vs. no-questioning) as independent variables, and comprehension outcomes and cognitive load as dependent variables. Again, we

counterbalanced condition order and assignment of narratives to conditions (Zeelenberg & Pecher, 2015).

Materials

All materials were the same as in Experiment 1, but we added children's scores on a standardized assessment of reading comprehension skills (administered by the schools) to validate our comprehension test.

Inference-stimulating questions and comprehension test

For scoring the inference-stimulating questions and the comprehension test, one rater scored the answers of all children and a second rater scored the answers of ten randomly selected children. Inter-rater reliability was high for both the inference-stimulating questions (ICC = 0.89; range separate narratives = 0.77 − 1.00) and the comprehension test (ICC = 0.94; range separate narratives = 0.90-0.97). Moreover, we tested the validity of scores on the comprehension test using a standardized assessment of reading comprehension (see Reading Comprehension Skills). Children's standardized scores were correlated to their comprehension scores in the reading conditions, $r = .67$, $p = .003$, and the listening conditions, $r = .51$, $p = .036$. These strong and significant correlations provide evidence for the validity of scores on our comprehension test questions, which were used in both Experiment 1 and 2.

Decoding skills

The participating school had administered the 2009 version of the Three Minutes Test (Krom et al., 2010). The reliability of this version is high, with a Cronbach's alpha of 0.97 (Krom et al., 2010).

Reading comprehension skills

Reading comprehension skills were assessed with a Dutch standardized test of reading comprehension (Cito test), showing good reliability (reliability coefficient = 0.86; Jolink et al., 2015). The participating school had administered the 2018 version of this test. Children received a booklet with texts and a booklet with multiple choice questions that they had to answer after each text. Children worked individually and without a time limit.

Procedure

The standardized decoding test was administered by the school in June, at the end of first grade. The rest of the experiment was conducted during the start of second grade. The procedure was the same as Experiment 1.

Table 2 Mean scores (and SD) on factual questions, comprehension questions, and cognitive load in each condition

	Reading		Listening	
	With questions	Without questions	With questions	Without questions
Factual (0–2)	1.50 (0.51)	1.72 (0.46)	1.56 (0.62)	1.31 (0.75)
Comprehension (0–4)	1.92 (1.09)	2.11 (0.76)	2.25 (0.86)	1.86 (0.90)
Cognitive load (1–9)	3.91 (2.58)	3.27 (2.53)	2.55 (1.41)	2.23 (1.69)

Results and discussion

Before testing our hypotheses, children's DLE-scores on the standardized decoding test (administered at the end of first grade) were analyzed. The mean DLE-scores ($M = 9.90$, $SD = 7.02$) indicated that the children's decoding skills were at the expected level, given the months of education received (9–10 months). Next, we checked children's performance on the inferential questions asked during reading and listening. The average score (2.56 out of 4) was lower than in Experiment 1, but there was no significant difference between questions asked during reading ($M = 2.57$, $SD = 1.05$) and during listening ($M = 2.55$, $SD = 1.03$), $t(21) = 0.09$, $p = .932$, $d = 0.02$.

Table 2 presents the mean scores on the factual questions, comprehension questions, and cognitive load in each condition. Repeated measures ANOVAs with modality and inference-stimulation as within-subjects factors were performed on comprehension scores and mental effort ratings. Contrary to our expectations, comprehension scores in the reading conditions ($M = 2.01$, $SD = 0.18$) and the listening conditions ($M = 2.06$, $SD = 0.16$) did not significantly differ, $F(1, 17) = 0.03$, $p = .855$, $n_p^2 < 0.01$. There was no significant main effect of stimulating inference-making, $F(1, 17) = 0.32$, $p = .579$, $n_p^2 = 0.02$, indicating that comprehension scores in the questioning conditions ($M = 2.08$, $SD = 0.16$) and the no-questioning conditions ($M = 1.99$, $SD = 0.15$) did not significantly differ. There was no significant modality × questioning interaction on comprehension scores, $F(1, 17) = 2.23$, $p = .153$, $n_p^2 = 0.12$. The analysis of cognitive load shows that, in line with our hypotheses, mental effort ratings in the reading conditions ($M = 3.59$, $SD = 0.51$) were significantly higher than ratings in the listening conditions ($M = 2.39$, $SD = 0.31$), $F(1, 21) = 5.99$, $p = .023$, $n_p^2 = 0.22$. As expected, inferential questioning during reading and listening ($M = 3.23$, $SD = 2.75$) resulted in higher cognitive load compared to no inference-stimulation ($M = 2.75$, $SD = 0.35$), $F(1, 21) = 4.84$, $p = .039$, $n_p^2 = 0.19$. There was no significant modality × questioning interaction for cognitive load, $F(1, 21) = 0.40$, $p = .53$, $n_p^2 = 0.02$.

Together, in contrast to Experiment 1, reading a narrative imposed higher cognitive load compared to listening to a narrative (Hypothesis 3). This can be explained by the fact that in Experiment 2 children's decoding skills were less advanced and therefore more cognitive resources were needed for text decoding.

Listening to a narrative did not impose additional cognitive demands as listening comprehension does not require decoding skills. However, contrary to our expectations, this higher cognitive load did not result in lower comprehension outcomes in the reading condition (Hypothesis 1). Instead, children's comprehension level was comparable across the two modalities. Apparently, reading required the children to invest more effort in understanding, but not that much to cause cognitive overload and impair comprehension. Regarding the effect of inference-stimulation, the results indicated that answering inferential questions during reading and listening increased cognitive load, as expected (Hypothesis 4). However, contrary to expectations, but in line with Experiment 1, the results showed that stimulating inference-making did not affect comprehension outcomes (Hypotheses 2a & 2b).

So why did we not find the benefits of inferential questioning that were reported by van den Broek and colleagues (2001; 2011)? One notable difference with their studies is how we measured comprehension. Based on the work of Kraal and colleagues (2018; 2019), we used open-ended comprehension questions directly targeting the same causal connections as the inference-stimuling questions that were asked during reading and listening. Consequently, children in the condition without questioning were prompted to generate these inferences during the comprehension test, and may have done so successfully based on their memory of the narrative. So, our comprehension test may not have distinguished very well between inferences made during reading or listening and inferences made during the test. In the van den Broek et al. (2001; 2011) studies, however, a free recall protocol was used to measure children's comprehension after reading or listening. An advantage of this is that children are not directly prompted to draw inferences, which may better reflect differences in inferences made during reading or listening.

To conclude, in this experiment we tested younger children with lower decoding skills and found that inferential questioning increased cognitive load. However, still no effect on comprehension outcomes was found. Therefore, in Experiment 3 we tested whether the way comprehension was measured might explain why no effect on comprehension was obtained.

Experiment 3

Experiment 3 builds on the previous two experiments, but now using van den Broek et al.'s (2001; 2011) free recall task. Again, we compared a listening condition to a reading condition, and included a no-questioning control group. However, instead of using open-ended comprehension questions (Experiment 1 and 2), we used free recall to measure children's comprehension. Additionally, we looked at children's recall of specific inferences in the narrative which enables a comparison with the open-ended comprehension questions used in our previous two experiments. Based on our findings in Experiment 1 and 2, we expected no comprehension differences between reading and listening conditions. Data were

collected at the end of second grade, so based on the results of Experiment 1, we expected no cognitive load differences between reading and listening conditions. Using van den Broek et al.'s free recall task, we expected a negative effect of inference-stimulation on children's comprehension in the reading condition (van den Broek et al., 2001). Moreover, we expected a positive effect of inference-stimulation in the listening condition (Butterfuss et al., 2021; Freed & Cain, 2017; van den Broek et al., 2011). Finally, as in Experiment 1 and 2, we expected children in the questioning condition to experience higher cognitive load compared to children in the no-questioning condition.

Method

Participants and design

Participants were 60 second grade children from four elementary schools in the Netherlands. Data from two children were excluded because of failed audio-recordings, resulting in a sample of 58 children (29 boys, 29 girls; age range 7–8 years). Children's parents or caretakers signed an informed consent letter. We used a 2×2 design with modality (reading vs. listening) as within-subjects variable and inference-stimulation (questioning vs. no-questioning) as between-subjects variable, and comprehension and cognitive load as dependent variables. Inference-stimulation was, in constrast to Experiment 1 and 2, a between-subjects variable to better align with the van den Broek et al. (2001) study. Children were randomly assigned to either the questioning condition ($N = 30$) or the no-questioning condition ($N = 28$). To check whether the inference-stimulation groups were comparable, we looked at children's standardized decoding skills and standardized reading comprehension skills. The same four narratives were used as in the previous experiments. In each condition, children received two narratives that were matched on difficulty (based on scores in the previous two experiments). One pair consisted of the hide-and-seek and elephant Okke narratives and the other pair consisted of the little mouse and iPad narratives. For each participant, each pair was randomly assigned to the reading or listening condition. The four narratives were presented in a fixed order (hide-and-seek, little mouse, elephant Okke, and iPad), so half of the children started with reading and the other half with listening. Presentation of texts and audio-recordings was alternated.

Materials

Same as Experiment 1 and 2, except for the measurements described below.

Comprehension test. Comprehension was measured using a free recall task (van den Broek et al., 2001; van den Broek et al., 2011). After each narrative, children saw an image of a light bulb and the text "What do you recall from the narrative?". The test leader instructed them to tell as much as possible about what they

Fig. 2 The 9-point mental effort rating scale with a bear at both endpoints

remembered from the story. Neutral prompts and encouragements were used to stimulate children to continue talking. When children indicated to not remember anything, one question was asked to stimulate talking (e.g., who were the characters in the story?). No feedback was given on the free recall responses.

Free recall responses were recorded and afterwards transcribed. Four independent raters divided the responses into subject-verb clauses (or idea units; Trabasso & van den Broek, 1985). The original text of the narratives was also divided into idea units. Following van den Broek et al. (2001; 2011), children's overall recall scores were calculated by dividing the number of correctly recalled idea units by the total number of idea units in the narrative. Additionally, we considered 'question-related' recall to evaluate the effect of inferential questioning more specifically. We labeled the idea units in the narratives relevant for answering the inferential questions and calculated question-related recall scores by dividing the number of recalled question-related idea units by the total number of question-related idea units in the narrative. Two of the four independent raters scored the free recall responses of 20% of the children ($N = 12$). Inter-rater reliability was high for the total recall scores (ICC = 0.90) and question-related recall scores (ICC = 0.80). After establishing good reliability, the four raters scored the responses of the remaining children.

We again validated the scores on our comprehension measure using a standardized assessment of reading comprehension (see Reading Comprehension Skills). This resulted in medium and significant correlations between children's standardized scores and their overall recall in the reading condition, $r = .43$, $p = .001$, and listening condition, $r = .36$, $p = .008$. The same was found for question-related recall in the reading condition, $r = .47$, $p < .001$, and listening condition, $r = .31$, $p = .025$.

Cognitive load. The same instrument was used as in Experiment 1 and 2, but based on our experiences in these experiments, we replaced the anime drawings with pictures of bears from The Bear Cards as these better assist children in recognizing and talking about their feelings and invested effort (Fig. 2, Qcards, 2010).

Table 3 Mean (and SD) decoding and reading comprehension skills of children in the questioning and no-questioning group

	Questioning	No-questioning
Decoding skill (0–7)	2.98 (1.08)	3.39 (1.26)
Reading comprehension skill (1–5)	3.59 (1.58)	3.64 (1.41)

Decoding skills. All four schools had administered the 2018 version of the Three Minutes Test. The reliability of this version is high (reliability coefficient = 0.97; van Til et al., 2018). Schools provided children's level of functioning, indicating whether a child performs at a level representative for the middle/end of a certain grade level (e.g., M4 for the middle of grade 2). These levels were transformed into numbers (i.e., <M3 = 0, M3 = 1, E3 = 2, M4 = 3 etc.).

Reading comprehension skills. All four schools had administered the 2018 version of the standardized Cito reading comprehension test (see Experiment 2). The participating schools provided us with children's norm scores, based on a national norming sample of Cito. Children were given an A (highest scoring 25% of children in the Netherlands), B (25% scoring above average), C, (25% scoring below average), D, (15% scoring well below average) or E (lowest scoring 10%).

Procedure

The standardized decoding and reading comprehension tests were administered by the participating schools in January. The rest of the experiment was conducted in May. The procedure was identical to the procedure in Experiment 1 and 2, with a few exceptions. In the no-questioning condition, the practice and experimental narratives were presented without inferential questions in-between the fragments. After reading or listening to a fragment, children clicked 'next', and were immediately presented with the next fragment. After each narrative, children indicated their mental effort and then completed the free recall test.

Results and discussion

Before testing our hypotheses, children's decoding test scores were analyzed. Children's mean scores ($M = 3.18$, $SD = 1.18$) suggest that our sample has a level of decoding that is representative for the middle of second grade (matching the time of testing). Moreover, we examined children's performance on the inferential questions asked during reading and listening. Average performance on the inferential questions was 3.10 (out of 4), which is comparable to Experiment 1 but slightly higher than in Experiment 2. Again, scores on the inferential questions asked during reading ($M = 3.01$, $SD = 0.71$) and listening ($M = 3.19$, $SD = 0.68$) did not differ significantly, $t(28) = 1.33$, $p = .195$, $d = 0.25$. Finally, comparing the questioning and the no-questioning groups on reading comprehension and decoding skills

Table 4 Mean scores (and SD) on total recall, question-related recall, and cognitive load in each condition

	Reading		Listening	
	With questions	Without questions	With questions	Without questions
Total recall	0.30 (0.13)	0.36 (0.15)	0.29 (0.11)	0.38 (0.14)
Question-related recall	0.36 (0.15)	0.39 (0.17)	0.38 (0.12)	0.43 (0.15)
Cognitive load	2.57 (1.10)	2.14 (1.22)	2.10 (1.09)	2.00 (1.43)

(Table 3) showed that the randomization of participants to conditions was successful, as none of the differences were significant, all $ps > 0.05$. Additionally, we tested whether children's decoding skills and reading comprehension skills correlated with our dependent measures. Only reading comprehension skills were significantly correlated with children's total recall and question-related recall, as reported before (see Comprehension Test). Therefore, we added reading comprehension skill as a covariate when analyzing total recall and question-related recall scores.

Table 4 presents the total recall, question-related recall, and cognitive load in each condition. We expected no difference between reading and listening in terms of comprehension and cognitive load. Moreover, we expected inferential questioning to increase cognitive load, and to have a positive effect on comprehension in the listening condition, but a negative effect on comprehension in the reading condition. To test the hypotheses, separate mixed ANCOVAs with modality as within-subjects factor and inference-stimulation as between-subjects factor were performed on total recall scores, question-related recall scores, and cognitive load. When analyzing total recall and question-related recall, standardized reading comprehension was added as covariate.

For total recall, in line with our expectations, we found no difference between reading ($M = 0.33$, $SD = 0.02$) and listening ($M = 0.33$, $SD = 0.02$), $F(1, 56) = 1.34$, $p = .253$, $n_p^2 = 0.03$. Contrary to our expectations, we found a significant effect of inference-stimulation, $F(1, 56) = 5.48$, $p = .023$, $n_p^2 = 0.10$, with children in the questioning condition ($M = 0.29$, $SD = 0.02$) recalling fewer idea units than children in the no-questioning condition ($M = 0.37$, $SD = 0.02$). Finally, contrary to our expectations, there was no modality × inference-stimulation interaction, $F(1, 56) = 0.85$, $p = .363$, $n_p^2 = 0.02$. For question-related recall, we found a significant main effect of modality on comprehension, $F(1, 56) = 5.16$, $p = .028$, $n_p^2 = 0.10$. Children recalled more question-related idea units in the listening condition ($M = 0.41$, $SD = 0.02$) compared to the reading condition ($M = 0.37$, $SD = 0.02$). However, there was no significant main effect of inference-stimulation, $F(1, 56) = 1.29$, $p = .262$, $n_p^2 = 0.03$, and also no significant modality × inference-stimulation interaction, $F(1, 56) = 0.23$, $p = .637$, $n_p^2 < 0.01$. Regarding cognitive load, we again found no significant difference between reading a narrative ($M = 2.36$, $SD = 0.15$) and listening to a narrative ($M = 2.01$, $SD = 0.17$), $F(1, 56) = 3.57$, $p = .064$, $n_p^2 = 0.06$. Also, contrary to our expectations, there was no significant cognitive load difference between children in the questioning condition ($M = 2.33$, $SD = 0.19$) and children in the condition

without questioning ($M = 2.07$, $SD = 2.33$), $F(1, 56) = 0.90$, $p = .346$, $n_p^2 = 0.02$. This lack of a difference in cognitive load may be related to questioning being a between-subjects factor, so children did not have a standard for comparison. There was no significant modality × inference-stimulation interaction, $F(1, 56) = 1.01$, $p = .320$, $n_p^2 = 0.02$.

The finding that stimulating inference-making negatively affects children's narrative comprehension corresponds with the results of van den Broek et al. (2001). We replicated their result that for beginning readers, answering inferential questions during reading negatively affects comprehension outcomes. However, our idea was that listening to narratives may provide children with a cognitively less demanding context in which inference-stimulation could effectively improve comprehension. Instead, our results suggest that asking inferential questions is not beneficial for beginning readers' comprehension, independent from modality. Our findings add to the literature by showing that inferential questioning during listening was also not effective in improving comprehension, when compared to a control group without inferential questioning. This raises questions about the suitability of inferential questioning as a strategy to improve comprehension in beginning readers.

A possible explanation for the overall negative effect of inferential questioning on total recall is that children in the inferential questioning conditions were less inclined to share everything they remembered. They had already answered the inferential questions during reading and listening, so possibly they thought it was not necessary to repeat this information during the free recall task. This explanation is not supported by our data, as there was no effect of inferential questioning on recall of question-related idea units. Given that the questioning condition showed lower scores on total recall of idea units, a more likely explanation is that children in this condition had focused only on recalling information that was relevant for answering the inferential questions and left out other parts of the narrative.

General discussion

In three experiments, we investigated whether stimulating inference-making among beginning readers is more effective and less cognitively demanding in a listening context than in a reading context. We expected that stimulating inference-making through inferential questioning would increase cognitive load, and that it would negatively affect comprehension in a reading context (due to additional cognitive demands related to decoding), but positively affect comprehension in a listening context (because questions direct children's attention to important inferences in the narrative). The results indicated that, overall, inferential questioning indeed increased children's cognitive load (Experiment 1 & 2). However, inferential questioning did not affect comprehension in the way we expected, because in all three experiments, we did not find the hypothesized interaction effect between inferential questioning and modality (i.e., reading vs. listening). Instead, we found that asking inferential questions did not improve narrative comprehension: not in a reading context, nor in a listening context. This indicates that stimulating inference-making

in a listening context is not necessarily more effective or less cognitively demanding than in a reading context.

Effect of stimulating inference-making with questioning

The lack of a positive effect of stimulating inference-making during listening extends prior research as follows. Previous studies reported that inferential questioning during listening is more effective than questioning after listening (Butterfuss et al., 2021; Freed & Cain, 2017; van den Broek et al., 2011). But our results indicate that asking inferential questions during listening does not improve comprehension compared to no-questioning. Importantly, this result was obtained both when measuring comprehension using open-ended comprehension questions (Experiment 1 & 2), and when using a free recall protocol (Experiment 3). For the latter, it was even found that inferential questioning negatively affected children's total recall in both the reading and listening condition. These findings raise the question whether inferential questioning should be used for improving comprehension of beginning readers.

This lack of a positive effect on listening (or reading) comprehension might simply indicate that our participants were well able to understand the narratives without the questions. However, children's comprehension scores in the no-questioning condition show that there is still room for improvement. On average, children answered 2.67 (Experiment 1) or 1.99 (Experiment 2) out of 4 questions correctly, and recalled only 37% of all idea units (Experiment 3). Another explanation could be that the questioning technique did not work as planned. We exploratively tested this explanation by correlating performance on the inferential questions with comprehension scores[2]. The correlations indicate that overall, children's ability to answer the inferential questions was not related to their comprehension after reading or listening. So, children who were well able to answer the inferential questions during reading or listening did not necessarily perform better on the comprehension test, and vice versa. This is surprising, given that our inferential questions were aligned with the comprehension test. That is, they were targeting the same implicit relations in the narrative.

A reason for why the questioning technique did not work as expected could be that children failed to recognize the link between the inference-stimulating questions and the comprehension test. Even though the questions were aimed at the same causal connections, they still require some sort of transfer. Similarly, the free recall task is a different task than answering the inferential questions during the narrative, so performance on these tasks may be unrelated as well. Another difficulty

[2] For Experiment 1, we found a medium and significant correlation between children's scores on the inferential questions and their comprehension scores, $r = .44$, $p = .004$. However, for Experiment 2, scores on the inferential questions and comprehension scores were not significantly correlated, $r = .24$, $p = .330$. Similarly, in Experiment 3, we found no significant correlation between children's performance on the inferential questions and their total recall scores, $r = .02$, $p = .920$, or their question-related recall, $r = .15$, $p = .448$.

might have been that children were not given the option to go back in the narrative when they were unable to answer the inference-stimulating questions. These options would allow children to actually search for the causal connections that the questions were aimed at and this way facilitate inference-making. They also did not receive feedback on their answers, just like in previous studies (Freed & Cain, 2017; van den Broek et al., 2001; 2011). According to the iLC framework (Kendeou et al., 2020), inferencing can be supported via inferential questioning when including scaffolding and feedback. Indeed, the results of Butterfuss et al. (2021) suggest that immediate feedback explaining why the inference was correct or incorrect may be a valuable addition. Therefore, it would be interesting for future research to compare the effect of inferential questioning during reading and during listening in a setting where children can go back in the narrative and/or receive feedback on their answers to the inference-stimulating questions.

Finally, in this study we used online questioning, which involves prompting inference-making at points when they are needed for comprehension, so called coherence breaks (Butterfuss et al., 2021). This online approach may have affected the results, for example by interrupting the comprehension process and hereby increasing cognitive load, as was shown by our results. It is uncertain whether questioning after reading or listening would yield similar results. Especially given that offline questioning does not interrupt reading or listening, which possibly allows young children to focus their limited resources on comprehension. Moreover, the duration of the questioning intervention may have affected the results. Previous studies showed positive effects when beginning readers follow multiple lessons with inferential questioning (e.g., McMaster et al., 2012). Possibly, young children need more time to get used to answering questions during reading or listening, before it can positively affect their comprehension. Future research could also examine the effect of teaching a comprehension strategy in both a reading and listening context. Using a strategy does not require assistance of a teacher or researcher, which allows children to work independently. For example, we could also train them to formulate their own questions (i.e., a self-questioning strategy, Joseph et al., 2016). It would be interesting to examine whether children can be trained in using a strategy with audio materials, and whether this also results in improved inference-making and comprehension when reading. Practicing with a strategy in a listening context could be less cognitively demanding and more motivating for children than reading.

Effect of modality

Another unexpected finding from our study is that there were hardly any differences between reading and listening in terms of comprehension and cognitive load. The only difference we found was that reading imposed slightly higher cognitive load for less advanced readers (Experiment 2), and that listening resulted in better recall of question-related idea units (Experiment 3). This suggests that listening was slightly easier for beginning readers. These findings do not align with our expectations based on previous studies that compared the level of reading and listening comprehension in primary school and reported differences for this age group (Diakidoy et al.,

2005; Sticht & James, 1984; Verlaan et al., 2017; Wannagat et al., 2017). Possibly, instructing children to read the narratives aloud resulted in a reading condition that was more similar to listening than expected. On the other hand, the reading condition still required children to decode written text, so this cannot fully explain the lack of a difference. Another potential explanation is related to differences between languages (Florit & Cain, 2011). Most studies reporting differences in the level of reading and listening comprehension were conducted in languages with a deep orthography such as English (except for Wannagat et al., 2017). In such languages the mapping between graphemes (i.e., letters or clusters of letters) and phonemes (i.e., sounds) is less consistent. In transparent orthographies, such as the Dutch language, the grapheme-phoneme mapping is more consistent, which makes it easier to learn and master basic reading skills. So, in our sample of Dutch second graders, reading comprehension may be less affected by children's decoding skills and consequently the level of reading and listening comprehension is already comparable at a younger age. This also aligns with our cognitive load results indicating that understanding spoken narratives is not necessarily more cognitively demanding than understanding written narratives. It is of course possible that listening and reading were cognitively demanding in a different way. For example, while written text imposes demands due to the requirement to decode the text, spoken text is transient requiring children to keep earlier presented information active in working memory to connect it to later presented information. So possibly, local or text-based inferences are easier to make when reading than when listening (Freed & Cain, 2021). Future research is needed to investigate to what extent differences in cognitive load relate to the effectiveness of inferential questioning. This would extend prior research comparing listening and reading comprehension in a novel direction as cognitive load has so far not been measured in these studies.

Limitations

A limitation of our experiments is that we used relatively small samples, especially in Experiment 2. However, due to the within-subjects designs in Experiment 1 and 2, there were still 43 and 18 children per condition, respectively. This is comparable with previous studies (e.g., van den Broek et al., 2001), showing significant effects of inferential questioning. Another limitation might be that we used a subjective one-item measure of cognitive load (Paas, 1992). Although this scale has been shown to be a reliable and valid assessment of cognitive load during a task (Paas et al., 2003), more elaborate (e.g., Leppink et al., 2015) or physiological (e.g., Skulmowski and Rey, 2017) measures of cognitive load may provide different outcomes. As described in the method, we aimed to make the scale more age-appropriate (Laurie-Rose et al., 2014). However, it remains uncertain whether the young children in our sample were capable of giving a good estimate of their mental effort during reading or listening. Another point is that the alarm clock and message that were presented on the computer screen, prior to asking the inference-stimulating questions, potentially disrupted children's situation model construction. This may have affected children's comprehension processes and outcomes, so future studies

should prevent these additional distractions. Finally, our results may be restricted to our specific materials, that is, the narratives and comprehension questions that were used. For example, we used narratives with a causal story structure, but a different genre with other inference types may yield different results. Additionally, the internal consistency of our comprehension assessment appeared to be rather low, which is not surprising, but relevant to note for researchers wanting to use these materials.

Conclusion

To conclude, the present study was the first to evaluate the effect of inferential questioning on both reading and listening comprehension, including a no-questioning control group. Adding this control group led us to conclude that inferential questioning was actually not beneficial for children's reading comprehension or listening comprehension, which is a different from what earlier studies suggested (Butterfuss et al., 2021; Freed & Cain, 2017; van den Broek et al., 2011). Our direct comparison between stimulating inference-making in a reading and a listening context demonstrated that inferential questioning did not improve children's narrative comprehension, not in a reading context, nor in a listening context. So, a practical implication is that listening is not necessarily an easier context to practice inference-making skills for beginning readers than reading. These findings call for more research on how we can support inference-making in beginning readers, and whether we can use other media than written text to improve reading comprehension.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11145-023-10463-x.

Acknowledgements This study was partly funded by a grant from the Netherlands Initiative for Education Research (NRO), grant nr. 40.5.18300.038 (NRO-PROO Samenhangende Onderzoeksprojecten 2018). We would like to thank Astrid Kraal for sharing her texts and corresponding comprehension questions with us. Data collection and data processing of the third experiment were conducted as part of the master thesis of Renske Doornbos. We thank her, Renske Beeren, Evita Moerlie, Naomi Verstoep, and Sanne van der Ent for their contribution to collecting, transcribing, and scoring the data.

Declarations

Conflict of interest We have no conflicts of interest to disclose. An earlier version of this paper was presented at the Junior Researchers (JURE) 2021 conference of the European Association for Research on Learning and Instruction (EARLI). Correspondence concerning this article should be addressed to Brechtje E. J. van Zeijts. Email: vanzeijts@essb.eur.nl.

References

Aarnoutse, C. A. J., van den Bos, K. P., & Brand-Gruwel, S. (1998). Effects of listening comprehension training on listening and reading. *The Journal of Special Education*, *32*(2), 115–126. https://doi.org/10.1177/002246699803200206.

Bergsma, A. (2009). *Haas en pad gaan verhuizen*. Maretak.

Berkeley, S., Scruggs, T. E., & Mastropieri, M. A. (2010). Reading comprehension instruction for students with learning disabilities, 1995–2006: A meta-analysis. *Remedial and Special Education*, *31*(6), 423–436. https://doi.org/10.1177/0741932509355988.

Best, R. M., Floyd, R. G., & McNamara, D. S. (2008). Differential competencies contributing to children's comprehension of narrative and expository texts. *Reading Psychology*, *29*(2), 137–164. https://doi.org/10.1080/02702710801963951.

Brand-Gruwel, S., Aarnoutse, C. A. J., & van den Bos, K. P. (1998). Improving text comprehension strategies in reading and listening settings. *Learning and Instruction*, *8*(1), 63–81. https://doi.org/10.1016/S0959-4752(97)00002-9.

Butterfuss, R., Kendeou, P., McMaster, K. L., Orcutt, E., & Bulut, O. (2021). Question timing, language comprehension, and executive function in inferencing. *Scientific Studies of Reading*, *52*(1), 1–18. https://doi.org/10.1080/10888438.2021.1901903.

Cain, K., & Oakhill, J. V. (1999). Inference making ability and its relation to comprehension failure in young children. *Reading and Writing*, *11*(5), 489–503. https://doi.org/10.1023/A:1008084120205.

Carretti, B., Caldarola, N., Tencati, C., & Cornoldi, C. (2014). Improving reading comprehension in reading and listening settings: The effect of two training programmes focusing on metacognition and working memory. *British Journal of Educational Psychology*, *84*(2), 194–210. https://doi.org/10.1111/bjep.12022.

Catts, H. W., Herrera, S., Nielsen, D. C., & Bridges, M. S. (2015). Early prediction of reading comprehension within the simple view framework. *Reading and Writing*, *28*(9), 1407–1425. https://doi.org/10.1007/s11145-015-9576-x.

Clarke, P. J., Snowling, M. J., Truelove, E., & Hulme, C. (2010). Ameliorating children's reading-comprehension difficulties: A randomized controlled trial. *Psychological Science*, *21*(8), 1106–1116. https://doi.org/10.1177/0956797610375449.

De Jong, P. F., & van der Leij, A. (2002). Effects of phonological abilities and linguistic comprehension on the development of reading. *Scientific Studies of Reading*, *6*(1), 51–77. https://doi.org/10.1207/S1532799XSSR0601_03.

Diakidoy, I. A. N., Stylianou, P., Karefillidou, C., & Papageorgiou, P. (2005). The relationship between listening and reading comprehension of different types of text at increasing grade levels. *Reading Psychology*, *26*(1), 55–80. https://doi.org/10.1080/02702710590910584.

Dutch Inspectorate of Education (2020). *Peil.Taal en rekenen. Einde basisonderwijs 2018–2019*. https://www.onderwijsinspectie.nl/documenten/themarapporten/2020/04/22/peil.taal-en-rekenen-2018-2019.

Elleman, A. M. (2017). Examining the impact of inference instruction on the literal and inferential comprehension of skilled and less skilled readers: A meta-analytic review. *Journal of Educational Psychology*, *109*(6), 761–781. https://doi.org/10.1037/edu0000180.

Evers, G. (2008). Programma voor berekening Cito LeesIndex voor het Basisonderwijs. P-CLIB versie 3.0. Cito.

Florit, E., & Cain, K. (2011). The simple view of reading: Is it valid for different types of alphabetic orthographies? *Educational Psychology Review*, *23*(4), 553–576. https://doi.org/10.1007/s10648-011-9175-6.

Florit, E., Roch, M., & Levorato, M. C. (2011). Listening text comprehension of explicit and implicit information in preschoolers: The role of verbal and inferential skills. *Discourse Processes*, *48*(2), 119–138. https://doi.org/10.1080/0163853X.2010.494244.

Freed, J., & Cain, K. (2017). Assessing school-aged children's inference-making: The effect of story test format in listening comprehension. *International Journal of Language & Communication Disorders*, *52*(1), 95–105. https://doi.org/10.1111/1460-6984.12260.

Freed, J., & Cain, K. (2021). Assessment of inference-making in children using comprehension questions and story retelling: Effect of text modality and a story presentation format. *International Journal of Language & Communication Disorders*, *56*(3), 637–652. https://doi.org/10.1111/1460-6984.12620.

Gernsbacher, M. A., Varner, K. R., & Faust, M. E. (1990). Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning Memory and Cognition*, *16*(3), 430–445. https://doi.org/10.1037/0278-7393.16.3.430.

Hall, C. S. (2016). Inference instruction for struggling readers: A synthesis of intervention research. *Educational Psychology Review*, *28*(1), 1–22. https://doi.org/10.1007/s10648-014-9295-x.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, *2*(2), 127–160. https://doi.org/10.1007/BF00401799.

Jolink, A., Tomesen, M., Hilte, M., Weekers, A., & Engelen, R. (2015). *Wetenschappelijke verantwoording begrijpend lezen 3.0 voor groep 4 [Scientific justification of reading comprehension tests for second grade]* Cito. https://www.cito.nl/kennis-en-innovatie/kennisbank/038-wetenschappelijke-verantwoording-lvs-toetsen-begrijpend-lezen-3-0-voor-groep-4.

Joseph, L. M., Alber-Morgan, S., Cullen, J., & Rouse, C. (2016). The effects of self-questioning on reading comprehension: A literature review. *Reading & Writing Quarterly*, *32*(2), 152–173. https://doi.org/10.1080/10573569.2014.891449.

Kendeou, P., Lynch, J. S., van den Broek, P., Espin, C. A., White, M. J., & Kremer, K. E. (2005). Developing successful readers: Building early comprehension skills through television viewing and listening. *Early Childhood Education Journal*, *33*(3), 91–98. https://doi.org/10.1007/s10643-005-0030-6.

Kendeou, P., Bohn-Gettler, C., White, M. J., & van den Broek, P. (2008). Children's inference generation across different media. *Journal of Research in Reading*, *31*(3), 259–272. https://doi.org/10.1111/j.1467-9817.2008.00370.x.

Kendeou, P., van den Broek, P., Helder, A., & Karlsson, J. (2014). A cognitive view of reading comprehension: Implications for reading difficulties. *Learning Disabilities Research & Practice*, *29*(1), 10–16. https://doi.org/10.1111/ldrp.12025.

Kendeou, P., McMaster, K. L., & Christ, T. J. (2016). Reading comprehension: Core components and processes. *Policy Insights from the Behavioral and Brain Sciences*, *3*(1), 62–69. https://doi.org/10.1177/2372732215624707.

Kendeou, P., McMaster, K. L., Butterfuss, R., Kim, J., Bresina, B., & Wagner, K. (2020). The inferential language comprehension (iLC) framework: Supporting children's comprehension of visual narratives. *Topics in Cognitive Science*, *12*(1), 256–273. https://doi.org/10.1111/tops.12457.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363–394. https://doi.org/10.1037/0033-295X.85.5.363.

Kraal, A., Koornneef, A. W., Saab, N., & van den Broek, P. W. (2018). Processing of expository and narrative texts by low-and high-comprehending children. *Reading and Writing*, *31*(9), 2017–2040. https://doi.org/10.1007/s11145-017-9789-2.

Kraal, A., van den Broek, P. W., Koornneef, A. W., Ganushchak, L. Y., & Saab, N. (2019). Differences in text processing by low-and high-comprehending beginning readers of expository and narrative texts: Evidence from eye movements. *Learning and Individual Differences*, *74*(2019), 101752. https://doi.org/10.1016/j.lindif.2019.101752.

Krom, R., Jongen, I., Verhelst, N., Kamphuis, F., & Kleintjes, F. (2010). *Wetenschappelijke verantwoording DMT and AVI [Scientific justificiation of decoding skills tests]*. Cito.

Laurie-Rose, C., Frey, M., Ennis, A., & Zamary, A. (2014). Measuring perceived mental workload in children. *The American Journal of Psychology*, *127*(1), 107–125. https://doi.org/10.5406/amerjpsyc.127.1.0107.

Leppink, J., Gog van, T., Paas, F., & Sweller, J. (2015). Cognitive load theory: Researching and planning teaching to maximise learning. In J. Cleland, & S. J. Durning (Eds.), *Researching Medical Education* (pp. 207–218). John Wiley & Sons.

Magliano, J. P., Loschky, L. C., Clinton, J. A., & Larson, A. M. (2013). Is reading the same as viewing? In B. Miller, L. E. Cutting, & P. McCardle (Eds.), *Unraveling the behavioral, neurobiological and genetic components of reading comprehension* (pp. 78–90). Brookes Publishing Co.

McMaster, K. L., van den Broek, P., Espin, C. A., White, M. J., Rapp, D. N., Kendeou, P., & Carlson, S. (2012). Making the right connections: Differential effects of reading intervention for subgroups of comprehenders. *Learning and Individual Differences*, *22*(1), 100–111. https://doi.org/10.1016/j.lindif.2011.11.017.

McNamara, D. S., & Maglino, J. (2009). Toward a comprehensive model of comprehension. In B. H. Ross (Ed.), *Psychology of learning and motivation*, (pp. 297–384). Elsevier. https://doi.org/10.1016/S0079-7421(09)51009-2.

Melis, G., Oosterveld, P., & Schokker, J. (2012). *Het verantwoord gebruik van didactische leeftijdsequivalenten: De kritiek weerlegd* Boom test uitgevers. https://www.boomtestonderwijs.nl/media/14/dle_boek_artikel_verantwoord_gebruik_van_dles.pdf.

National Assessment of Educational Progress (2019). The nation's report card. Retrieved from https://www.nationsreportcard.gov/highlights/reading/2019/.

Noort, N. (2015). *Bijt de dief in zijn billen!* Zwijsen.

Oakhill, J. V., Cain, K., & Bryant, P. E. (2003). The dissociation of word reading and text comprehension: Evidence from component skills. *Language and Cognitive Processes*, *18*(4), 443–468. https://doi.org/10.1080/01690960344000008.

Paas, F. G. W. C. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive load approach. *Journal of Educational Psychology*, *84*(2), 429–434. https://doi.org/10.1037/0022-0663.84.4.429.

Paas, F., Tuovinen, J. E., Tabbers, H., & van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, *38*(1), 63–71. https://doi.org/10.1207/S15326985EP3801_8.

Qcards (2010). *The Bear Cards*. http://www.qcards.com.au/the-bear-cards.

Skulmowski, A., & Rey, G. D. (2017). Measuring cognitive load in embodied learning settings. *Frontiers in Psychology*, *8*(1191), 1–6. https://doi.org/10.3389/fpsyg.2017.01191.

Sticht, T. G., & James, J. H. (1984). Listening and reading. In P. D. Pearson, R. Barr, M. L. Kamil, & P. Mosenthal (Eds.), *Handbook of reading research* (pp. 293–317). White Plains.

Storch, S. A., & Whitehurst, G. J. (2002). Oral language and code-related precursors of reading: Evidence from a longitudinal structural model. *Developmental Psychology*, *38*(6), 934–947. https://doi.org/10.1037/0012-1649.38.6.934.

Sweller, J., van Merriënboer, J. J., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, *31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5.

Tibus, M., Heier, A., & Schwan, S. (2013). Do films make you learn? Inference processes in expository film comprehension. *Journal of Educational Psychology*, *105*(2), 329–340. https://doi.org/10.1037/a0030818.

Trabasso, T., & van den Broek, P. (1985). Causal thinking and the representation of narrative events. *Journal of Memory and Language*, *24*(5), 612–630. https://doi.org/10.1016/0749-596X(85)90049-X.

van den Broek, P., Lorch, E. P., & Thurlow, R. (1996). Children's and adults' memory for television stories: The role of causal factors, story-grammar categories, and hierarchical level. *Child Development*, *67*(6), 3010–3028. https://doi.org/10.1111/j.1467-8624.1996.tb01900.x.

van den Broek, P., Risden, K., Tzeng, Y., Trabasso, T., & Basche, P. (2001). Inferential questioning: Effects on comprehension of narrative texts as a function of grade and timing. *Journal of Educational Psychology*, *93*(3), 521–529. https://doi.org/10.1037/0022-0663.93.3.521.

van den Broek, P., Kendeou, P., Lousberg, S., & Visser, G. (2011). Preparing for reading comprehension: Fostering text comprehension skills in preschool and early elementary school children. *International Electronic Journal of Elementary Education*, *4*(1), 259–268.

van den Broek, P. W., Helder, A., & van Leijenhorst, L. (2013). Sensitivity to structural centrality: Developmental and individual differences in reading comprehension skills. In M. A. Britt, S. R. Goldman, J.-F. Rouet. (Eds.), *Reading: From words to multiple texts* (pp. 132–146). Routledge. https://doi.org/10.4324/9780203131268.

van Til, A., Kamphuis, F., Keuning, J., Gijsel, M., Vloedgraven, J., & de Wijs, A. (2018). *Wetenschappelijke verantwoording LVS-toetsen DMT [Scientific justification of the Three Minutes Test]*. Cito. https://www.cito.nl/kennis-en-innovatie/kennisbank/106-wetenschappelijke-verantwoording-lvs-toetsen-dmt-voor-gr3-tm-halverwege-gr8.

Verhoeven, L., & van Leeuwe, J. (2009). Modeling the growth of word-decoding skills: Evidence from dutch. *Scientific Studies of Reading*, *13*(3), 205–223. https://doi.org/10.1080/10888430902851356.

Verlaan, W., Pearce, D. L., & Zeng, G. (2017). Revisiting sticht: The changing nature of the relationship between listening comprehension and reading comprehension among upper elementary and middle school students over the last 50 years. *Literacy Research and Instruction*, *56*(2), 176–197. https://doi.org/10.1080/19388071.2016.1275070.

Wannagat, W., Waizenegger, G., & Nieding, G. (2017). Multi-level mental representations of written, auditory, and audiovisual text in children and adults. *Cognitive Processing*, *18*(4), 491–504. https://doi.org/10.1007/s10339-017-0820-y.

Zeelenberg, R., & Pecher, D. (2015). A method for simultaneously counterbalancing condition order and assignment of stimulus materials to conditions. *Behavior Research Methods*, *47*(1), 127–133. https://doi.org/10.3758/s13428-014-0476-9.