Developmental differences in children's generation of knowledge-based inferences.

Inference making is essential for adequate understanding of text as it involves the addition of information that has not been explicitly stated into the mental representation of the text (Johnson-Laird, 1983; Kintsch, 1988; Van Dijk & Kintsch, 1983). Children make the inferences necessary to understand short (aurally presented) narratives from as young as 4 years (Kendeou et al., 2008; Lepola et al., 2012; Tompkins et al., 2013) and inference skills predict children's and adolescents' reading comprehension, even after controlling for factors such as IQ, word reading, vocabulary and relevant background knowledge (Ahmed et al., 2016; Barnes et al.,1996). In this study we focus on inferences that require integration of information in text with background knowledge (knowledge-based inferences: Graesser et al., 1994). The ability to generate knowledge-based inferences that are necessary for good comprehension improves with age (Barnes et al., 1996; Chrysochoou & Bablekou, 2010; Currie & Cain, 2015; Schmidt & Paris, 1983), but it is unclear which factors contribute to these established developmental improvements. Empirical studies are necessary to inform our theoretical models of inference making, its development and breakdown, so that we can provide evidence-based targeted instruction and intervention to foster better inference making skills (Hall & Barnes, 2017).

Knowledge-based inferences involve integrating information in the text with background knowledge and are considered necessary to construct a coherent and accurate mental representation of a text's meaning (Kintsch, 1988)[1]. The ability to generate knowledge-based inferences improves between 5 to 15 years (Barnes et al., 1996; Currie & Cain, 2015; Schmidt & Paris, 1983) and is related to gains in memory capacity and vocabulary knowledge in that period of development, both of which are established

[1] We note that all inferences rely to some extent on vocabulary and/or background knowledge, however, we focus on those that rely more heavily on the integration of relevant knowledge, do not have an explicit cue such as a synonym to signal integration, and which are necessary for maintaining coherence. We do not include elaborative inferences in this definition as, although they may enrich the mental representation, they are not essential for coherence.

predictors of inference making performance (Cain & Oakhill, 2014; Currie & Cain, 2015; Language and Reading Research Consortium (LARRC), Currie & Muijselaar, 2019). However, although we know these factors predict the product of inference generation (e.g., answering an inferential question), we know less about how children of different ages utilise the information provided in a text in order to generate knowledge-based inferences.

In this study we assessed children's ability to use a series of converging clues embedded into short texts to generate a knowledge-based inference (e.g., 'sand', 'pier', 'water' which, when integrated, support the generation of the inference 'seaside'). The process of generating this type of knowledge-based inference relies on two mechanisms considered in theories of skilled text comprehension: Enhancement of relevant concepts and suppression of irrelevant concepts (Gernsbacher, 1990; Kintsch, 1988). These mechanisms rely not only on the activation of the associates of a given word and knowledge of its interrelations, but also on the ability to inhibit irrelevant inferences that might be generated, both of which are considered crucial components of the construction and refinement of a mental representation of a text (Gernsbacher 1990; Gernsbacher & Faust, 1991; Kintsch, 1988). There is a large body of literature assessing how skilled readers (i.e., adults) process text (e.g., see Albrecht & O'Brien, 1993; Graesser et al., 1994). In contrast, studies of developmental differences in children's ability to use information in text to generate knowledge-based inferences are lacking. To address this gap in our knowledge we examined how children aged 6, 8 and 10 years of age make use of converging evidence to generate target knowledge-based inferences and suppress competing inferences.

Schmidt and Paris (1983) demonstrated that even young children are sensitive to the amount of evidence a text provides for target inferences. They assessed 5- to 10-year-old children's ability to use a series of converging clues presented in short narratives, to generate knowledge-based inferences. In one experiment children listened to stories with either zero,

one, or three converging clues for the target inference. Developmental improvements in performance were found, with all age groups performing better when there were more clues to support the target inference. In a second experiment, designed to examine how children monitor incoming clues and update inferences, children were asked to sort a set of picture cards representing possible inferences into plausible and implausible sets, after hearing sentences containing either relevant clues or irrelevant filler information. The 5-year-olds were less accurate in their final selection and made a greater number of errors when sorting their cards. They were also more likely than older children to use irrelevant filler sentences to inform their decisions.

These experiments demonstrate clear developmental improvements in the use of converging evidence when generating knowledge-based inferences; however, the tasks prompted the children to make an inference either through the inference-tapping question (experiment 1) or by reflecting on what was plausible (experiment 2). In this study, we examined children's ability to make necessary knowledge-based inferences without an explicit prompt, by recording their response to four single-word probes that differed in their relation to the passage and the target inference. This paradigm differs to those that have used an inference-tapping question, which explicitly directs the reader to produce the inference (e.g., Question: Where were the children? Answer: At the seaside). The probe words were: (1) the target inference word, (e.g., 'seaside' in a text that mentions 'sand', 'pier', 'water'); (2) a competing inference (e.g., riverside); (3) an unrelated concept (e.g., ladder); and (4) a literal probe (a word that appeared explicitly in the text). Thus, three of the probes did not feature explicitly in the text: The target inference, the competing inference, and the unrelated concept, and the literal probe appeared once. The four probes were presented after their associated passage. By analysing both accuracy and speed of response to the probes, this

paradigm measures the accessibility of the inferable concept after passage presentation without using a question that explicitly requires inference generation.

In addition to using converging evidence for a target inference, it is also important to suppress competing inferences once they are deemed irrelevant. Schmidt and Paris (1983) demonstrated that younger children are more susceptible to interference from competing inferences that they generate, even in the presence of converging evidence. When asked to explicitly reflect on competing inferences they found that younger children had greater difficulty producing the correct target inference. Further empirical evidence (using a more implicit design to assess suppression) is provided by Lorsbach et al (1998). In their study, 9-year-olds, 12-year-olds and adults listened to stories that included a garden path, where the opening supported one interpretation (inference) that had to be suppressed from the participant's mental representation when subsequent text strongly supported an alternative interpretation. Children were more likely than adults to wrongly accept the original 'competing' interpretation at the end of the story rather than the new alternative interpretation, which was supported by the information in the text. This finding indicates that children are less able than adults to suppress an irrelevant competing inference.

Congruent with this finding, Perez et al. (2015) used ERP data to demonstrate that, under certain conditions, adults fail to suppress initial inferences that are not supported by subsequent text. Specifically, they found that adults with low working memory had difficulty revising their situation model and integrating new information that supported an alternative inference, despite being able to detect whether or not incoming information was consistent with their current situation model. In the current study we assess if the ability to monitor and update the choice of inference is weaker in younger children (Schmidt & Paris, 1983). We manipulate the amount of converging evidence for the inference and examine how this

impacts the acceptance/rejection of target and competing inferences, as well as the time taken to make this decision.

Response times were included in our study to help shed light on when the target inference was activated (Rapp et al., 2007). Response time techniques are well-established in studies of adults' inference making and demonstrate that adults engage in inferential processing during the reading of texts (Albrecht & O'Brien, 1993; Casteel, 1993; Long & Chong, 2001). However, to date few studies have made use of response time paradigms to specifically assess inference generation in young children. Response times to single-word probes at the end of sentences and passages have been used to provide insight into how adults and children enhance the correct meaning (and suppress the incorrect meaning) of ambiguous words (Barnes et al., 2004; Gernsbacher & Faust, 1991; McNamara & McDaniel, 2004). In a series of experiments, Barnes et al. (2004) used response times to single-word probes to assess different components of comprehension in 12-year-old typically developing children and children with hydrocephalus. One experiment examined the suppression of irrelevant meanings for ambiguous homographs (e.g., 'spade'). Children with hydrocephalus took longer than typically developing children to decide if a probe word was related to the preceding sentence in an ambiguous condition even when the word was presented after a delay of 1000ms, the point at which suppression of the irrelevant meaning should have occurred. We adopted a similar method: Children responded to the set of four word probes at the end of short passages to assess when the inferable concept was activated (i.e. while listening to the text or upon hearing the probe word) and the efficiency of the suppression of the competing inference. For the target inference, we compared response times to an inference probe with those for a literal probe explicitly stated only once at the beginning of the passage. Longer response times for the inference probe would suggest that the inferable concept was not activated during text presentation but at the point of testing. For the

competing inference, we compared the speed of response to an unrelated probe. We are not aware of any studies to date utilising response time measures to specifically assess when knowledge-based inferences are activated and the suppression of competing knowledge-based inferences in different age groups of children.

The current study

Materials similar to Schmidt and Paris (1983) were used in the current study. Children aged 6, 8 and 10 years of age listened to short six-sentence stories, each containing clues for a knowledge-based inference (See Table 1). The task was designed to span a wide age range of children, so we used a listening paradigm to ensure that performance was not influenced by word reading skills. In half of the stories there was an opening (but at this stage ambiguous) clue and one additional specific clue (referred to as the one clue condition), each embedded into a sentence in the text; in the other half of the stories there was one opening clue and three specific clues (referred to as the three clue condition). We designed our texts to ensure that clues were known to all ages of children in this study and we aimed to use the most relevant clues for each concept (see method section for further detail). Children responded yes/no to four probe words after each text, according to whether they thought they were related to the text they had just listened to. The four probe words included the target inference, a literal word from the beginning of the text, a competing inference and an unrelated filler concept (See Table 1). We measured the accuracy of responses to these four probe words and, to give new insight into the efficiency of activation and suppression in children, we also measured response times for each probe. In order to account for the influence of individual differences in participants and the influence of specific texts we used mixed effects modelling in our analyses. This enables generalisation of our findings to other populations of participants and materials.

In relation to age we expected developmental improvements in accuracy for the target inference probe (Barnes et al., 1996; Currie & Cain, 2015; Schmidt & Paris, 1983). With regards to the number of clues available we predicted higher levels of accuracy to the target inference in the three clue condition, in line with the previous literature (Schmidt & Paris, 1983). In general, we expected older children to respond more quickly to the probes than younger children and we predicted faster responses to the inference word in the three clue condition. However, we were particularly interested in the response to the inference versus the literal probes as this comparison provides unique insight into potential developmental differences in the processing of this type of inference. If the target inference is activated while listening to the text, accuracy to the inference probe should be at least equivalent, if not greater, than accuracy to the literal probe (a word that appeared only once in the text). Critically, in the case of the response times, we expected equivalent or faster responses to the inference than the literal probe, particularly given that the clues for the inference concept occurred more often in the text and therefore should benefit from enhanced activation. If this effect is more pronounced in older children, this may suggest that they are more likely than younger children to activate the inferable concept while listening to the passage.

Turning to the competing inference, we expected accuracy to be higher in the three clue condition where the additional clues should make rejection of the competing inference more likely. In relation to age we predicted that the younger children would accept the competing inference as being related to the story more often than older children (Schmidt & Paris, 1983). With regards to the response times and the efficiency of the suppression of the competing inference, if the competing inference had been activated and not fully suppressed we would expect to find longer response times in general to the competing inference than the unrelated filler. There may also be evidence of increased delay in response to the competing

inference for younger relative to older children, given the evidence for weaker executive function skills (Cain, 2006; Caretti et al., 2005).

Method

Participants

Three age groups participated in this study, comprising 26 children aged 5-6 years (15 boys; *M* = 6,04, *SD* = 4 months, range = 5,10 – 6,09), 27 children aged 7-8 years (13 boys; *M* = 8,04, *SD* = 3 months, range = 7,11 – 8,10), and 25 children aged 9-10 years (13 boys; *M* = 10,04, *SD* = 3 months, range = 9,10 – 10,09). Consent was obtained from parents and headteachers, and children gave their assent prior to testing. All participating children had English as a first language and children with a statement of special educational needs did not take part.

Materials

Sixteen six-sentence stories were created, some adapted from the materials used by Lorsbach et al. (1998) and Perez et al. (2015) with the addition of some new stories. There were two versions of each story: a one clue version and a three clue version; the clues supported a knowledge based inference, which was the identity of the object (see Table 1 for an example). The number of clues was a within-subjects factor. The one and three clue versions were counterbalanced across two presentation lists so that each child saw only one version of each text. In both versions, sentence one introduced the protagonist of the story and sentence two provided an opening clue that contained an overarching category that the item might be an exemplar of (but at this stage ambiguous). There were then two versions of sentences three and four: In the three clue version, two specific clues to the identity of the exemplar were provided, one in each sentence; in the one clue version, two filler sentences appeared that did not provide any additional information relating to the target inference. In both versions, sentence five contained the same final specific clue and sentence six provided

an ending to the story that did not refer back to the inference item. The name of the inference item was not stated in the stories and was referred to as 'it' or 'thing' throughout.

The clues for each target inference item were selected using the University of South Florida Free Association Norms (Nelson et al., 1998). Wherever possible clues were selected from the top 15 attributes for that item, excluding any other exemplars of the same category (e.g., for the item 'apple' any other examples of fruit were excluded). This database was constructed from American adult norms, so some associates were not relevant for children and could not be used (e.g., 'Newton' and 'gravity' for apple). The clue words were checked for frequency of occurrence in books for children aged 5 to 9 years using the Children's Printed Word Database (CPWD: Masterson et al., 2010). This ensured that the children would be familiar with the clue words selected from the database. All clue words (n=64) were listed in the CPWD apart from three: '*safari'*, '*living room'* and '*program'* (4.68% of all clues). Piloting of the task with children in each age group indicated that children of this age range were familiar with these words. Of the remaining clues, there were no overall differences in frequency between any of the clues (including the opening clue, all ts < 1.8, all ps > .09). To ensure consistency in reference to the target inference by pronouns in the two versions, the one and three clue stories were matched for the number of direct references by pronouns in sentences three and four. See Table 1 for an example. The stories were recorded and edited using Audacity software and presented to the children via a laptop through headphones. The texts were presented in a random order via Eprime (Schneider et al., 2002) for each participant. Response accuracy and time to respond to the probe words were recorded using Eprime software.

Immediately after each story, the child heard four words individually in a random order (assigned by Eprime for each story to avoid order effects) and was asked to judge whether each was related to the story by responding yes/no on a keypad. After responding to

a probe word there was a 1000ms delay before the next word was presented. Participants were able to respond while the word was being pronounced and there was no maximum time window for responding. The words comprised: The name of the target inference item; a literal word from sentence one; a word that was not related to the story; and a word that was a competing inference. The literal word was included to check for memory of events at the beginning of the story. Examples are provided in Table 1. The unrelated filler probe was included as a check for a bias to 'yes' responses. The competing inference was a member of the category clue stated in the opening and was included to assess whether or not the competing inference was active in their mental representation of the story.

A pilot study with adults informed the choices for the competing inferences. Ten adults were given two opening sentences of the story and asked to name three things that the story could be about. Their only guidance was the opening clue in sentence two. Where possible the top two items produced by the adults were used as the inference and competing inference. Five of the stories had to be changed after this pilot, so that the vocabulary used was in line with children's knowledge. The probe words were also checked for frequency of occurrence in children's stories using the CPWD (Masterson et al., 2010). The mean frequency for the four word probe types did not differ (all ts < 1.0, all ps > .10) [2].

Procedure

Children completed the task individually in a quiet room. The task took approximately fifteen minutes including practice items and took place in a single session lasting around twenty minutes. There were three practice stories at the beginning of the session. The first practice story was completed with the assessor. Children listened to the following text (opening ambiguous clue is shown in italics, specific clues are underlined): 'Emma did not eat any

[2] One competing inference word *'badminton'* was not in the frequency database, however, piloting indicated children were familiar with this word.

sweets today. Emma had an *appointment* after school. Before Emma went she used her toothbrush. Emma sat in the waiting room until it was her turn. Emma had her teeth polished. Then it was time to go home.' The researcher would then respond in the following way after hearing each of the probe words: For the inference probe the researcher would say, 'dentist' – the story talked about the character going for an 'appointment' and then the story talked about a 'toothbrush' 'waiting room' and 'teeth' so I think this word is related to the story. You can press 'YES.' For the literal word: 'sweets' the story said that 'Emma did not eat any sweets today' so I think this word is related to the story. You can press 'YES.' For the unrelated filler: 'cushion' I don't think this word is related to the story. You can press 'NO'. For the competing inference: 'hospital' I don't think this word is related to the story. You can press 'NO'. The child then completed two further practice items - they responded on the keypad themselves but any wrong answers or questions about how to respond to a probe word were explained in the same manner as above. The child completed the rest of the task individually, with the assessor sat next to the child throughout the task. The test stories and probe words were played over headphones to minimise any external distractions.

Data Analysis

The accuracy and response time data were analysed using mixed effects modelling with the lme4 package (Bates et al., 2015). The models included the fixed effects of condition (one clue, three clue), word type (inference or literal; unrelated filler or competing inference), age (6 years, 8 years, 10 years) and their interactions. These variables were contrast coded[3] with the contrasts for age specified so that the beta coefficient for the 6- to 8-year-olds represents the difference between the beta coefficients for the 6- and 8-year-olds around the overall mean and the beta coefficient for the 8- to 10-year-olds represents the difference

[3] The contrasts of condition and word type were coded in the following way: for condition one clue =+1, three clue = -1; for the word type 'yes' analysis inference = +1, literal =-1; for the word type 'no' analysis competing = +1, filler =-1.

between the beta coefficients for the 8- and 10-year-olds around the overall mean. Word probe comparisons (inference or literal; unrelated filler or competing inference) were analysed in separate analyses because the correct response for inference and literal probe words required a 'yes' response and the correct response for competing inference and filler probe words required a 'no' response and these response types may involve different search and response strategies.

Maximal models containing fixed and random effects with both intercept and slopes terms (Barr et al., 2013) did not converge, so we followed recommendations to simplify the specification of random effects for each model (Brauer & Kurtin, 2017; Matuschek et al., 2017). All significant interactions were examined visually with interaction plots and also by sub-setting the data and providing the beta coefficient as an indication of the average difference between conditions.

Results

The results are reported in two sections, first the accuracy data for the responses to the probe words and then the response time data to these words. The data and code for the following analyses are available on OSF (LINK HERE).

Accuracy

Inference and literal probe words

A correct response to these two categories of word was 'yes.' Table 2 shows the mean proportion correct for all four probe words. A 'yes' response bias (showing no discrimination between the prompts) would be indicated by very high accuracy for the inference and literal prompts and very low accuracy for the competing inference and filler prompts, because a correct response to the latter two is 'no'. The full model summary for the inference and literal analysis is reported in Table 3. The best-fitting model (here and for all main analyses, unless otherwise stated) included the fixed effects of clue condition (one clue, three clues), word

type (inference, literal) and age (6-, 8-, 10-year-olds); random intercepts for subjects and items and random slopes for word type in the participant random term.

There was a main effect of clue condition because children were more likely to respond correctly in the three clue condition than in the one clue condition (See Table 3 for model summary). In general, there was a main effect of probe word type because inference probes were more likely to be answered correctly than the literal probes. There was also a main effect of age, but only for the comparison of the 8- and 10-year-olds to the overall mean and this was due to better performance by the 10-year-old children. The 6- and 8-year-olds did not differ in response accuracy. These effects were further qualified by two interactions, one between clue condition and type of probe word, the other between age and type of probe word. The clue condition x probe word interaction was examined via an interaction plot and by sub-setting the data by type of probe word and running models excluding probe word type as a fixed effect. For inference probe words, as expected, children were more likely to respond correctly in the three clue condition than in the one clue condition (B = -0.45, SE = 0.08, z = -5.43, p < .001; M_{one} = .68(.47); M_{three} = .81(.40). In contrast, there was no difference between clue conditions for the literal probe words (B = 0.13, SE = 0.07, z = 1.92, p = .06; M_{one} = .76 (.43); M_{three} = .72 (.45)), which was in line with our expectations because the literal word appeared only once in each text.

The age x type of probe interaction was examined visually and by sub-setting the data by age group and running the models excluding age group as a fixed effect. The slope for type of probe word was also excluded because the models for the 6- and 10-year-olds did not converge.[4] A different pattern of responses was found for each age group. In general the 6-year-olds were more likely to answer literal probe words correctly than inference probe

[4] We report the model excluding the word type slope for the 8-year-olds. A model including the word type slope did converge but did not change the pattern of results.

words (B = -0.29, SE = 0.08, z = -3.60, p < .001; $M_{inference}$ = .61 (.49); $M_{literal}$ = .73 (.45), the 8-year-olds did not differ in response accuracy for inference and literal probe words (B = 0.01, SE = 0.08, z = 0.08, p = .94; $M_{inference}$ = .72 (.45); $M_{literal}$ = .72 (.45)), whereas the 10-year-olds were more likely to correctly respond to the inference than the literal probe words (B = 0.58, SE = 0.11, z = 5.31, p < .001; $M_{inference}$ = .90 (.30); $M_{literal}$ = .77 (.42)) (see Figure 1).

Competing and unrelated filler probe words.

A correct response to these words was 'no'. See Table 2 for the means for each word type and Table 4 for the model summary. The best-fitting model for this analysis was the same as above but also included a slope for year in the items random term. There were two significant main effects. First, the type of probe word influenced performance: participants were more likely to respond correctly to filler probes than competing inference probes Second, age influenced performance: The 8-year-olds were more likely to respond correctly than the 6-year-olds but the 8- and 10-year-olds did not differ in accuracy. Contrary to predictions the effect of clue condition was not significant.

Inference and literal probe words: Dprime (d') analysis.

As noted, the means reported in Table 2 do not indicate a response bias. For an additional check, we conducted a supplementary analysis for the inference and literal probes to check for sensitivity to the correct 'target' responses, which was similar to the dprime (d′) analysis conducted by Lorsbach et al. (1998). To do this, we calculated 'hits' (proportion of correct 'yes' responses to the inference and literal probes) and 'false alarms' (proportion of incorrect 'yes' responses to the competing inference and unrelated filler probes) per condition for each child. The proportionate scores were converted into z-scores. The inference d′ was calculated as Zhits (inference) – Zfalse alarms (competing inference); the literal d′ score was calculated as Zhits (literal) – Zfalse alarms (unrelated filler).

An ANOVA with d′ scores as the dependent variable, year group as a between-subjects factors and clue condition (one vs three) and probe (inferential dprime vs literal dprime) as within subjects factors revealed no significant main effects or interactions for the number of clues or probe type (all Fs <2.00 and all ps > .14). There was one significant main effect of age $F(2, 75) = 10.04$, $p < .001$, because, overall, the 8- and 10-year-olds were more accurate than the 6-year-olds, whose d′ scores indicated that they were less likely to discriminate between the target items (inference, literal) and foils (competing inference, unrelated filler). However, we note that our study was not designed to use dprime as the primary outcome measure because the competing inference was expected to have a higher acceptance rate (lower accuracy) for younger children and/or when fewer cues were present. A full table of means and ANOVA output are reported in our online materials [LINK HERE].

Response times

The probe words were short (one word) stimuli and so the response times were measured from the onset of the word until the child pressed a button on the response box. We report the response times for probe words that were responded to correctly (See Table 5). Given the lower levels of accuracy for the 6-year-olds, the mean response times for all responses (correct and incorrect) is available on OSF for comparison (LINK HERE). The percentage of data points above 3sd of the mean response time per word type were: inference 1.5%; literal 2.24%; competing inference 1.36% and filler 2.16% (however, we had a large age range and there were no datapoints outside of 3sd of the mean for each word type for each individual subject in both clue conditions). There was some evidence of positive skew for all word probes, so the analyses presented here were conducted on log transformed correct only data. We report the model summaries for all responses (not log transformed) and response times for correct items (not log transformed) on OSF (LINK HERE).

Inference and literal probe words: Correct responses

See Table 5 for mean response times to each of the probe words and Table 6 for the model summary. There was a significant main effect of age: The 6-year-olds were slower than the 8-year-olds, but the 8- and 10-year-olds did not differ in response time, in general. The effect of condition (one vs three clues) was not significant, but type of probe word (inference vs literal) was found to be significant. This effect was qualified by a significant interaction between type of probe word and age for the 8- and 10-year-olds only. The interaction was not significant for the comparison of 6- and 8-year-olds. The interaction was examined by sub-setting the data by age group and running the models excluding age group as a fixed effect. The slope for word type was excluded as the model for the 8-year-olds would not converge. For the 8-year-olds there was no overall effect of probe word type (B = -0.013, SE = 0.007, t = -1.81, p = .07. However, the 10-year-olds were quicker to respond to the inference word than the literal word (B = -0.044, SE = 0.0061, t = -7.31, $p < .001$).[5]

Competing inference and unrelated filler probe words: Correct responses

See Table 5 for the means for each word type and for model output see Table 7. There were two significant main effects. First, the type of probe influenced the speed of response: children took longer to respond to the competing inference than the unrelated filler word. Second, the 8-year-olds were quicker to respond in general than the 6-year-olds but the 8- and 10-year-olds did not differ. The effect of clue condition was not significant.

Discussion

We assessed accuracy and speed of children's responses to probe words to examine knowledge-based inference making in children aged 6, 8 and 10 years. The number of clues for the target inference was manipulated within participants. As predicted, children were

[5] We report the model excluding the probe word type slope for the 10-year-olds. A model including the probe word type slope did converge but did not change the pattern of results.

more likely to respond correctly to the inference probe when more clues were presented (three vs one clue) (Schmidt & Paris, 1983), and the 10-year-olds were more likely to answer correctly than the younger children, in general (Barnes et al., 1996; Currie & Cain, 2015; Schmidt & Paris, 1983). For the inference and literal probes, the age groups showed different patterns of accuracy: The 10-year-olds were more likely to respond accurately to the inference; the 8-year-olds did not differ and the 6-year-olds were more likely to respond accurately to the literal probe. Response times for the inference and literal probe showed an effect of age because the two oldest age groups were quicker to respond in general. The age groups also had different response time patterns to the inference and literal probes: The 10-year-olds were quicker to respond to the inference than the literal probe, however, this effect was not evident in the two youngest age groups. In general, there was no effect of clue condition on response times.

The youngest age group performed at a lower level in general on the competing and unrelated filler probes. However, all age groups were slower and less accurate in response to the competing inference than the filler probe suggesting some instances of inhibition difficulty. Contrary to predictions, the number of clues did not influence accuracy or speed of responses to the competing inference for any of the age groups.

The patterns of performance for accuracy to the target inference are in line with the findings of Schmidt and Paris (1983): Older children were better at identifying the target inference than the younger age groups, even in the most supportive clue condition when the evidence strongly converged towards the target inference. Other empirical studies of knowledge-based inference making also find developmental differences, with increasing accuracy in successive age groups (Barnes et al., 1986; Chrysochoou & Bablekou, 2010; Currie & Cain, 2015). Novel to our study, we assessed responses to an inferable concept versus memory for other explicitly stated information in the same text. There was a clear

developmental progression with qualitative differences in the patterns of performance: The 6-year-olds were most accurate for the literal probes, the 10-year-olds were most accurate for the inference probes, whereas the 8-year-olds did not differ between types of probe. Therefore by 10 years of age the inferential concept was more salient than the literal, even though the latter was explicitly presented in the text. However, at 6 years of age the literal concept was more salient.

Part of the explanation for these findings could be that the target inference referred to a concept that was more central to each story (despite not being explicitly stated) than the literal word. Older children may be better at assessing what the central and peripheral components of a text are and have more cognitive resources available to direct their attention accordingly (see Miller & Keenan, 2009, for work with good and poor readers). Future work should explore these ideas in relation to inference generation. Related to this, the use of pronouns was a necessary part of the design in order to refer to the target inference without explicitly stating the inference. Pronouns such as 'it' may have assisted in keeping the inferential concept (and its related clues) in focus and older children may have benefited from this cue to a greater extent than younger children (e.g., Engelen et al., 2014). Although repetition of words has been found to improve performance in lexical decision tasks (Scarborough et al., 1977), only the 6-year-olds appear to have found the literal condition easier, perhaps indicating some benefit from explicit repetition of the literal word at this age.

The developmental pattern of response times to the inference and literal probe words was similar to the accuracy scores. Overall the 6-year-olds were slower to respond than the 8- and 10-year-olds. However, the 10-year-olds' response times were faster for the inference than the literal word. Because the oldest children responded more accurately and more quickly to the inference word than the literal word (which explicitly appeared in the text) we propose that they most likely experienced enhanced activation of the inferable concept as

they were listening to the text. This, in turn, enabled them to respond to the inference probe not only more accurately, but also more quickly than the literal probe. The additional clues provided in the three clue condition were not necessarily required by the 10-year-olds to respond more quickly to the inference probe word; one specific clue appears to be sufficient to enhance activation of the target inference. There was some indication of subtle influences of condition on response times for the younger children. First, the 8-year-olds showed a tendency towards faster response times for the inference probe in the three clue condition. Further, the mean response times show that the 6- and 8-year-olds took longer to respond to the literal word in the three clue condition, although this effect did not reach statistical significance. These findings might reflect that the text was more clearly focussed on the inference in the three clue condition. There was considerable variability in the response times, particularly for the youngest age group, which may (given their weaker performance) have been influenced by interference from other probe words, additional processing at the point of presentation of the probe and by confidence in their responses.

Why were the older children able to respond more accurately and quickly to the inference word? Although we controlled for both the associative strength and familiarity of the clue words, one reason could be that older children may be more likely to have rich and inter-connected vocabulary knowledge containing more detailed knowledge of semantic associates. Therefore when they hear a clue, such as '*kennel*' for the concept of *'dog'*, other related words are automatically activated, including the inference and related concepts (Kintsch, 1988; Perfetti, 2007). Related to this, one interpretation of the supplementary dprime analysis is that the younger children had less precise or complete mental models of the text, hence their greater acceptance of the foils and lower levels of accuracy in general. Older children and better readers have greater opportunities to broaden their knowledge from more complex and varied texts, which is critical for providing exposure to words in different

contexts and helping children to build a rich repertoire of words and their semantic associates (Nation, 2017). This additional knowledge would, in turn, enable the initial mental representation of a text to be richer, more flexible and, in the case of this study, perhaps more likely to activate and enhance the target inferable concept (Gernsbacher & Faust, 1991; Kintsch, 1988). In future work it would be beneficial to assess vocabulary knowledge alongside this type of task, particularly depth of vocabulary knowledge given its association with inference making (Cain & Oakhill, 2014; LARRC et al., 2019). Future research could also consider the semantic diversity of critical words in the texts to assess how this impacts inference generation (Hsiao & Nation, 2019; Nation, 2017).

Although we designed our texts to be appropriate for the age range in our study, assessment of the child's background knowledge for the specific concepts in the texts would offer insight into whether this knowledge was available. However, even when knowledge is available, it is not always easily accessible (Barnes et al., 1996), which will affect the likelihood that it is integrated with information provided in text to generate inferences. For example, Barnes et al. (1996) found age differences in the role of knowledge accessibility with it playing a particularly important role for essential coherence inference making in 6- to 9-year-olds but having a more important role for non-essential, elaborative inference making in older age groups. Therefore, in addition to differences in the richness of knowledge in different age groups, there may also be differences in the organisation and access to knowledge that could impact the likelihood the information would be integrated and/or the time taken to do so (Barnes et al., 1996). Future work is needed to explore these ideas further to help us to understand how knowledge availability and accessibility influence key components of the inference process.

Turning to performance on the competing inference and unrelated filler probe words, accuracy was higher for the filler than the competing inference, regardless of clue condition.

The youngest children performed more poorly in general. Developmental improvements in executive function, which includes the ability to regulate the contents of working memory and suppress (or inhibit) no longer relevant content, may in part explain the general age-related changes in performance seen here for the competing inference (Cain, 2006; Caretti et al., 2005; Butterfuss & Kendeou, 2018). In addition, a comprehension strategy focussing more closely on literal content (Karlsson et al., 2018; McMaster et al., 2012) could explain both the higher levels of incorrect rejection of the target inference probe and relatively high levels of correct rejection of the competing inference, if the younger children were simply making fewer inferences in general.

All age groups were slower to respond to the competing inference than the unrelated filler, suggesting that this concept may have led to some interference or perhaps (if activated), had not been fully inhibited (Lorsbach et al., 1998). Adults with weak comprehension skills demonstrate this type of interference and are less able to suppress the incorrect meaning of ambiguous words (Gernsbacher & Faust, 1991). Contrary to our predictions, having more clues for the target inference did not improve the accuracy or speed of rejection of the competing inference. In fact, the only instance where the number of clues influenced performance was in the accuracy (but not speed) of response to the target inference. In other words, increasing the likelihood that the target inference will be generated did not increase the efficiency of suppression of the competing inference.

Our findings have important implications for classroom practice and intervention. The developmental differences we observe indicate that age and inferential processing ability should be taken into account when planning instruction in inference making and in the remediation of reading comprehension difficulties. For example, although texts with a large number of clues for a knowledge-based inference can help to improve performance, for younger children around a third of the time this was still not sufficient. We should point out,

however, that the fact the 6-year-olds made gains on this task in the three clue condition demonstrates that this task was not beyond their attainment: They simply needed more clues to assist them in making the inference. Younger children, who may focus on literal content, could also be less likely to benefit from targeted inferential questions presented while listening or reading to text (see van den Broek et al., 2012, and McMaster et al., 2012), even in relatively short narratives like those we assessed here. However, older children who are more likely to be capable of activating relevant inferences while listening to text, may benefit from this type of instruction. In general, rich discussion of key words but also critically their associations to other words and concepts and encouragement to link the text with this knowledge are important for knowledge-based inference at all ages. With regards to the suppression of competing inferences, classroom materials designed to converge towards target inferences may not necessarily always lead to suppression of other competing inferences, even in older children. Making this process more explicit by modelling how the use of converging evidence can be used to rule out different knowledge-based inferences could be helpful.

There are limitations to this study, and we discuss the most pertinent here and how they might be addressed in future research. First, although we propose that the older age group's enhanced accuracy and speed of responding to the inference probe indicates activation during presentation of the text, evidence from online measures such as eye tracking and ERPs is needed to establish precisely when activation and encoding of an inference takes place (e.g., see Perez et al., 2015; Perez et al., 2016 for similar work with adults). Evidence from think-aloud tasks where participants are asked to reflect on a text at specific points during reading (or listening) could also shed light on the timing and quality of inferential processing (Karlsson et al., 2018; McMaster et al., 2012). In this study we focus on knowledge-based inferences but future work might usefully consider different types of

inference, to obtain a more comprehensive picture of how children process text to generate inferences at different ages. It may also be helpful to include filler texts that did not require an inference or required a different response to prevent any meta-comprehension of the study aims, particularly by older children.

Second, we did not take an independent measure of memory to assess how it related specifically to this type of inference. Schmidt and Paris (1983) suggest that younger children may be more likely to process texts sentence by sentence in a piecemeal manner and are less able to integrate their meanings. Lower memory capacity could have made it more difficult for the younger children to keep all of the clues in mind especially given that they may not have the same vocabulary resources to support this (Currie & Cain, 2015). However, the 6-year-olds were able to correctly respond to the literal word (which appeared at the beginning of each text) around 70% of the time, which strongly suggests that their memory for the texts in general was sufficient. Given our findings for the competing inference, a measure of memory designed to test the ability to suppress irrelevant information could be included in future work to shed light on the mechanisms at play (Caretti et al., 2005).

All four probe words were presented in a random order at the end of each text, therefore an additional consideration (with regards to the competing inference in particular) relates to when each probe was presented. If the competing inference probe was heard after the inference probe this could have led to activation of the competing inference, even if it had initially been suppressed (due to the competing and target inferences belonging to the same semantic category). Alternatively, hearing the competing inference probe before the inference probe may have led to the increased likelihood of an incorrect 'yes' response. The probe words were, however, randomised per item for each participant to minimise any order effects related to this and all texts were written to converge strongly towards the target inference. In future work we could focus on contrasting activation of target and competing inferences,

using a task that does not require a binary response to several probe words at the end of each text and instead assesses either the inference or the competing inference for each text.

Some may argue that our design does not necessarily tap into the mechanisms of knowledge-based inference generation and could be considered to be assessing lexical/memory retrieval; if so, the term inference could be replaced by 'associations', or 'context checking.' However, the type of knowledge-based inference that was a focus of our study does rely upon associations between words and, therefore, we would argue that our task does assess inference making ability. To activate and enhance the inference probe word goes beyond association between the clue words or between clue words and other associates – the child has to infer the target inference that is common to all of the clues. The oldest age group were also faster to respond to the inference probe than a word that explicitly appeared in the text. Therefore we would also argue for enhanced activation of the inference word that goes beyond context checking.

In summary our findings indicate developmental changes in children's ability to use converging evidence to activate and enhance knowledge-based inferences and suppress competing inferences. For the target inference there were clear qualitative differences in the patterns of accuracy and response times compared to the literal probe with a developmental shift from better performance for a literal concept in younger children through to more accurate and faster performance for an inferential concept in older children. This latter result also suggests that older children may have activated the inference while listening to the text, however, studies using online methodologies are needed to confirm this. In terms of suppression of the competing inference, although the youngest age group were more likely to wrongly accept the competing inference, all age groups made some errors and showed evidence of interference in responding to the competing inference. Taken together, these

findings are crucial for helping us to develop classroom practice to support the development of knowledge-based inference making in children of different ages.

Acknowledgments

This research was funded by an Economic and Social Research Council (ESRC) PhD studentship awarded to the first author.

References

Ahmed, Y., Francis, D. J., York, M., Fletcher, J. M., Barnes, M., & Kulesz, P. (2016). Validation of the direct and inferential mediation (DIME) model of reading comprehension in grades 7 through 12. *Contemporary Educational Psychology, 44*, 68-82. https://doi.org/10.1016/j.cedpsych.2016.02.002

Albrecht, J. E., & O'Brien, E. J. (1993). Updating a mental model: maintaining both local and global coherence. *Journal of Experimental Psychology*, *19*(5), 1061-1070. https://doi.org /10.1037/0278-7393.19.5.1061

Barnes, M.A., Dennis, M., & Haefele- Kalvatis, J. (1996). The effects of knowledge availability and knowledge accessibility on coherence and elaborative inferencing in children from six to fifteen years of age. *Journal of Experimental Child Psychology, 61*(3). https://doi.org /216-241. 10.1006/jecp.1996.0015

Barnes, M. A., Faulkner, H., Wilkinson, M., & Dennis, M. (2004). Meaning construction and integration in children with hydrocephalus. *Brain and Language*, *89*(1), 47-56. https://doi.org /10.1016/S0093-934X(03)00295-5

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

Brauer, M., & Curtin, J. J. (2017). Linear mixed-effects models and the analysis of nonindependent data: a unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, *23*(3), 389-411. https://doi.org/10.1037/met0000159

Butterfuss, R., & Kendeou, P. (2018). The role of executive functions in reading comprehension. *Educational Psychology Review*, *30*(3), 801-826. https://doi.org/10.1007/s10648-017-9422-6.

Cain, K., & Oakhill, J. (2014). Reading comprehension and vocabulary: Is vocabulary more important for some aspects of comprehension? *L'Année Psychologique*, *114*(4), 647-662. https://doi.org/10.4074/S0003503314004035

Carretti, B., Cornoldi, C., De Beni, R., & Romanò, M. (2005). Updating in working memory: A comparison of good and poor comprehenders. *Journal of Experimental Child Psychology*, *91*(1), 45-66. https://doi.org/10.1016/j.jecp.2005.01.005.

Casteel, M. A. (1993). Effects of inference necessity and reading goal on children's inferential generation. *Developmental Psychology*, *29*(2), 346-357. https://doi.org/10.1037/0012-1649.29.2.346

Chrysochoou, E., & Bablekou, Z. (2010). Phonlogical loop and central executive contributions to oral comprehension skills of 5.5 to 9.5 years old children. *Applied Cognitive Psychology*, *25*(4), 576-583. https://doi.org/10.1002/acp.1723.

Currie, N. K., & Cain, K. (2015). Children's inference generation: the role of vocabulary and working memory. *Journal of Experimental Child Psychology, 137*, 57-75. https://doi.org/10.1016/j.jecp.2015.03.005

Engelen, J. A. A., Bouwmeester, S., de Bruin, A. B.H., & Zwaan, R.A . (2014). Eye movements reveal differences in children's referential processing during narrative comprehension. *Journal of Experimental Child Psychology, 118*(1), 57-77. https://doi.org/10.1016/j.jecp.2013.09.005

Gernsbacher, M. A. (1990). *Language comprehension as structure building.* Hillsdale, NJ: Erlbaum.

Gernsbacher, M. A., & Faust, M. E. (1991). The mechanism of suppression: a component of general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(2), 245-262. https://doi.org/10.1037/0278-7393.17.2.245

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*(3), 371-395. https://doi.org/10.1037/0033-295X.101.3.371

Hall, C., & Barnes, M. A. (2017). Inference instruction to support reading comprehension for elementary students with learning disabilities. *Intervention in School and Clinic*, *52*(5), 279-286. https://doi.org/10.1177/1053451216676799

Hsiao, Y., & Nation, K. (2018). Semantic diversity, frequency and the development of lexical quality in children's word reading. *Journal of Memory and Language*, *103*, 114-126. https://doi.org/10.1016/j.jml.2018.08.005

Johnson-Laird, P. N. (1983). *Mental Models*. Cambridge: Harvard University Press.

Karlsson, J., van den Broek, P., Helder, A., Hickendorff, M., Koornneef, A., & van Leijenhorst, L. (2018). Profiles of young readers: Evidence from thinking aloud while reading narrative and expository texts. *Learning and Individual Differences*, *67*, 105-116. https://doi.org/10.1016/j.lindif.2018.08.001

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

Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. *Psychological Review*, *95*(2), 163-182. https://doi.org/10.1037/0033-295X.95.2.163

Language and Reading Research Consortium, Currie, N. K., & Muijselaar, M. M. L. (2019). Inference making in young children: The concurrent and longitudinal contributions of verbal working memory and vocabulary. *Journal of Educational Psychology, 111*(8), 1416–1431. https://doi.org/10.1037/edu0000342

Lepola, J., Lynch, J., Laakkonen, E., Silven, M., & Neimi, P. (2012). The role of inference making and other language skills in the development of narrative listening comprehension in 4–6-year-old children. *Reading Research Quarterly*, 47(3), 259-282. https://doi.org/10.1002/rrq.020

Long, D. L., & Chong, J. L. (2001) Comprehension skill and global coherence: A paradoxical picture of poor comprehenders' abilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(6), 1424-1429. https://doi.org/10.1037/0278-7393.27.6.1424

Lorsbach, T.C., Katz, G. A., & Cupak, A. J. (1998). Developmental differences in the ability to inhibit the initial misinterpretation of garden path passages. *Journal of Experimental Child Psychology, 71*(3), 275-296. https://doi.org/10.1006/jecp.1998.2462

Masterson, J., Stuart, M., Dixon, M., Lovejoy, S. (2010). Children's printed word database: Continuities and changes over time in children's early reading vocabulary. *British Journal of Psychology, 101*(2), 221-242. https://doi.org/10.1348/000712608X371744

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. M. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94*, 305-315. https://doi.org/10.1016/j.jml.2017.01.001

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., & McDaniel, M. A. (2004). Suppressing irrelevant information: Knowledge activation or inhibition?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 465-482. https://doi.org/10.1037/0278-7393.30.2.465

Miller, A. C., & Keenan, J.M. (2009). How word decoding skill impacts text memory: The centrality deficit and how domain knowledge can compensate. *Annals of Dyslexia, 59*(2), 99-113. https://doi.org/10.1007/s11881-009-0025-x

Nation, K. (2017). Nurturing a lexical legacy: Reading experience is critical for the development of word reading skill. *npj Science of Learning*, *2*(3). https://doi.org/ 10.1038/s41539-017-0004-7

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms.* http://www.usf.edu/FreeAssociation/

Perez, A., Cain, K., Castellanos, M. C., & Bajo, T. (2015). Inferential revision in narrative texts: an ERP study. *Memory & Cognition, 43*(8), 1105-1135. https://doi.org/ 10.3758/s13421-015-0528-0

Pérez, A., Joseph, H. S., Bajo, T., & Nation, K. (2016). Evaluation and revision of inferential comprehension in narrative texts: an eye movement study. *Language, Cognition and Neuroscience*, *31*(4), 549-566. https://doi.org/ 10.1080/23273798.2015.1115883

Perfetti, C. (2007). Reading Ability: Lexical Quality to Comprehension. *Scientific Studies of Reading, 11*(4), 357-383. https://doi.org/10.1080/10888430701530730

Rapp, D. N., Broek, P. V. D., McMaster, K. L., Kendeou, P., & Espin, C. A. (2007). Higher-order comprehension processes in struggling readers: A perspective for research and

intervention. *Scientific studies of reading*, *11*(4), 289-312. https://doi.org/10.1080/10888430701530417

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. Journal of Experimental Psychology: Human perception and performance, 3(1), 1-17. https://doi.org/ 10.1037//0096-1523.3.1.1

Schmidt, C. R., & Paris, S, G. (1983). Children's use of successive clues to generate and monitor inferences. *Child Development, 54*(3), 742-759. https://doi.org/10.2307/1130062

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). E-Prime User's Guide. Pittsburgh: Psychology Software Tools Inc.

Tompkins, V., Guo, Y., Justice, L.M. (2013). Inference generation, story comprehension, and language skills in the preschool years. *Reading and Writing, 26*(3), 403-429. https://doi.org/10.1007/s11145-012-9374-7

van den Broek, P., Tzeng, Y., Risden, K., 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 Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension.* New York: Academic Press.