



Published in final edited form as:

Sci Stud Read. 2021 ; 25(2): 141–158. doi:10.1080/10888438.2020.1831503.

The Process and Product of Coherence Monitoring in Young Readers: Effects of Reader and Text Characteristics

Nicola K. Currie^a, Gillian Francey^a, Robert Davies^a, Shelley Gray^b, Mindy S. Bridges^c, Maria Adelaida Restrepo^b, Marilyn S. Thompson^d, Margeaux F. Ciraolo^b, Jinxiang Hu^e, Kate Cain^a

^aDepartment of Psychology, Lancaster University, Lancaster, UK

^bCollege of Health Solutions, Arizona State University, Tempe, Arizona, USA

^cDepartment of Hearing and Speech, University of Kansas Medical Center, Kansas City, Kansas, USA

^dDenny Sanford School of Social and Family Dynamics, Arizona State University, Tempe, Arizona, USA

^eDepartment of Biostatistics & Data Science, University of Kansas Medical Center, Kansas City, Kansas, USA

Abstract

We examined sixth graders' detection of inconsistencies in narrative and expository passages, contrasting participants who were monolingual speakers ($N = 85$) or Spanish-English DLLs ($N = 94$) when recruited in pre-kindergarten (PK). We recorded self-paced reading times and judgments about whether the text made sense, and took an independent measure of word reading. Main findings were that inconsistency detection was better for narratives, for participants who were monolingual speakers in PK, and for those who were better word readers. When the text processing demands were increased by separating the inconsistent sentence and its premise with filler sentences there was a stronger signal for inconsistency detection during reading for better word readers. Reading patterns differed for texts for which children reported an inconsistency compared to those for which they did not, indicating a failure to adequately monitor for coherence while reading. Our performance measures indicate that narrative and expository texts make different demands on readers.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

CONTACT Kate Cain k.cain@lancaster.ac.uk Department of Psychology, Lancaster University, Lancaster, LA1 4YF, UK.

Ethics approval

The study conforms to the US Federal Policy for the Protection of Human Subjects and was approved by the Institutional Review Boards or Research Ethics Committees at each university.

Informed consent

Informed consent was provided by legal guardians and children gave their assent prior to participation

Conflicts of interest

The authors have no conflicts of interest to declare.

Reading comprehension involves the construction of an integrated and coherent representation of the information presented in the text (Johnson-Laird, 1983; Kintsch, 1998). This representation is updated continuously as the text unfolds and readers integrate successive ideas and concepts into the existing model (Rapp & Kendeou, 2007). Theoretically, monitoring the coherence of a text is critical in the construction of a mental representation of its content. Individuals who evaluate the adequacy of their comprehension will detect when information within the text is hard to integrate into the existing mental model, and may take action such as re-reading or inference making (Gernsbacher, 1990; Kintsch, 1998; Rapp & van den Broek, 2005). We extend previous research on coherence monitoring in young readers by examining the influence of critical reader and text characteristics on both the product and process of this skill.

The importance of coherence monitoring (also referred to as comprehension monitoring) to successful reading comprehension is evident from studies examining individual differences in reading comprehension and its development: Coherence monitoring is weak in children with poor reading comprehension (Ehrlich, Remond, & Tardieu, 1999; Oakhill, Hartt, & Samols, 2005) and predictive of concurrent and subsequent reading comprehension between 7 to 12 years, over and above word reading, vocabulary, and grammar (Kim, 2015; Language and Reading Research Consortium [LARRC] & Yeomans-Maldonado, 2017; Oakhill & Cain, 2012). Thus, coherence monitoring is well established as critical for successful reading comprehension.

Coherence monitoring is typically assessed using an error detection task, in which participants are presented with materials that include deliberate anomalies, such as nonwords, prior knowledge violations, or internal inconsistencies where two details in the text contradict. Many studies require readers (or listeners) to judge whether or not the material makes sense and to identify information that is not coherent with the whole (Baker, 1984; LARRC & Yeomans-Maldonado, 2017; Oakhill et al., 2005). Such measures, taken after the passage has been presented, capture the quality and coherence of the mental representation – the *product* of text processing. Fewer studies have examined both the product and the *process* of coherence monitoring in children – what happens when they encounter a coherence break during reading, particularly for passages (Harris, Kruithof, Terwogt, & Visser, 1981; Helder, Van Leijenhof, & van den Broek, 2016; Zabrocky & Ratner, 1992). This is surprising because one of the earliest studies of coherence monitoring contrasted product and process measures and highlighted the need to examine both to clarify the locus of difficulty (Harris et al., 1981). Processing measures are interpreted in relation to the online detection of a coherence break, with longer times indicating detection and consequent integration difficulty; product measures are interpreted to reflect what information is encoded into the mental model.

Harris et al. (1981) presented 8- and 11-year-olds with short narratives and recorded reading times as participants moved a screen to reveal each new sentence. In two experiments, both age groups took longer to read a sentence that was inconsistent with the passage title relative to a consistent condition. However, on completion of the passage, younger children were less likely to correctly report detection of an inconsistency or identify the anomaly. These findings have been reproduced in computer-based reading time studies. Helder et al. (2016)

presented 8 to 9- and 10 to 11-year-old good and poor reading comprehenders with narratives, sentence-by-sentence. In half the materials, a coherence break was created by making the second sentence of the passage inconsistent with a characteristic of the protagonist or situation presented in the target final sentence. All readers read the target sentence more slowly when inconsistent, but younger children and poorer comprehenders were less likely to report a coherence break on passage completion. Zabrocky and Ratner (1992) found a similar pattern of findings for 11–12-year-olds. Comparing findings for the product and process measures, each study concluded that all readers detected inconsistencies when present but only the better comprehenders routinely encoded this information in their mental model. Evaluating the adequacy of comprehension is an important skill for reading comprehension in general, and may become more critical in the later grades when reading to learn from text. We examine the influence of reader and text characteristics on both product and process measures to gain a better understanding of the source of coherence monitoring difficulties.

The majority of research on coherence monitoring has focused on monolingual speakers. Here, we contrast the performance of monolingual English speakers with children who were identified in PK as Spanish-English dual-language learners (DLLs) whose home language was Spanish, but who were schooled in English. This latter group comprises a sizable population in the U.S. and often shows a poor reading comprehension profile presenting age-appropriate word reading but, by nine years of age, reading comprehension and oral language that lags normative samples even when supported by bilingual education (Lesaux, Crosson, Kieffer, & Pierce, 2010; Nakamoto, Lindsey, & Manis, 2007).

Two studies have investigated whether coherence monitoring is a source of DLLs' weak reading comprehension (Denton et al., 2015; Lesaux & Harris, 2017). Lesaux and Harris (2017) found that 11–13-year-old language minority students reported regular engagement in self-monitoring and in strategic processing to extract meaning from text but demonstrated limited awareness of their comprehension difficulties. As the authors note, these findings contrast with those from a study using a think-aloud procedure with readers aged 12–16, comprising 50% Hispanic students (Denton et al., 2015); reading comprehension difficulties were associated with less frequent use of comprehension processes such as coherence monitoring. Denton et al. (2015) did not report the proportion of Hispanic students in their poor reader sample, limiting comparison between the two studies, and therefore understanding of the reading strengths and weaknesses of DLLs. Our study provides an important extension to this work by contrasting in children identified as monolingual or DLL in PK both the process of coherence monitoring and its product.

Text as well as reader characteristics may influence coherence monitoring (Rapp & van den Broek, 2005). Here, we consider the processing demands of the task and genre. Poor comprehenders are less likely to detect inconsistencies when a text has high processing demands involving integration of information across several sentences (Oakhill et al., 2005; van der Schoot, Reijntjes, & van Lieshout, 2012). van der Schoot et al. (2012) found that 10–12-year-old poor comprehenders took longer to read an inconsistent sentence when separated from its predicate by one sentence (low processing demands), but not when there were 5–7 sentences between the two (high processing demands), whilst good comprehenders

took longer to read the target sentences in both conditions. The authors concluded that, in contrast to good comprehenders, poor comprehenders do not routinely encode all information into their mental model, so that they only detected inconsistencies when both pieces of information were still active in working memory. Building on this, we examine how the distance between critical information in the text affects both the product and process of coherence monitoring.

Few studies of coherence monitoring have contrasted performance between narrative and expository texts (Denton et al., 2015; Zabucky & Ratner, 1992). Expository texts are considered more challenging than narratives because they can take a variety of structures and contain new information, alongside specialized vocabulary, placing a higher demand on the integration of information within the text and with prior knowledge to support learning (Best, Floyd, & McNamara, 2008; Graesser, McNamara, & Louwerse, 2003). In a recent brain imaging study, 8–10-year-olds made greater use of top-down regions, believed to support the strategic processing associated with coherence monitoring and integration, when reading expository compared to narrative text (Aboud, Bailey, Del Tufo, Barquero, & Cutting, 2019). In addition, Denton et al. (2015) found less evidence of integration, monitoring, and mental model building in think-alouds for expository compared to narrative texts (see also Zabucky & Ratner, 1992). They suggest that readers in this age group might not go beyond a basic textbase level (or locally coherent) representation for expository texts, indicating that standards of coherence might differ by genre (van den Broek, Ridsen, & Husebye-Hartman, 1995). If so, readers might be less accurate in detecting inconsistencies in expository texts relative to narratives, but equally likely to accept consistent texts as coherent in both genres. We are the first to compare processing and product measures of coherence monitoring for narrative and expository texts, to determine the locus of difficulty for each genre.

Current study

Our study extends previous research on young readers' coherence monitoring in several important ways. We examined performance on narrative and expository texts in two groups found to differ in reading comprehension: monolingual English speakers and Spanish-English DLLs. Unlike much previous research, our samples were identified in PK and experienced English-language instruction throughout schooling. Despite this, the evidence that monolingual and DLL children's reading comprehension skills diverge by around nine years led us to predict that the DLL group would perform more poorly, on average. Texts were either fully consistent or included an inconsistency between two sentences. We manipulated the distance between these two sentences to create two conditions contrasting lower and higher processing demands. We recorded sentence reading times, as well as responses to a sense judgment question after each text. We predicted that coherence monitoring would be poorer for expository relative to narrative texts, and for inconsistent texts with higher processing demands. Our design enables us to determine how differences in language group, genre or processing demands modulate the locus of difficulty detection or encoding of coherence breaks. We also included a measure of children's word reading, which we predicted would relate to task performance (Perfetti & Stafura, 2014). The close relation of our study design to that used in previous work on monolingual readers' coherence

monitoring, for narrative passages (e.g., Helder et al., 2016; van der Schoot et al., 2012), enables us to test the reproducibility of those study findings, and further to assess the generalizability of their theoretical accounts for different reader groups and genres.

Method

Participants

Eighty-five monolingual English speakers and 94 children who entered pre-kindergarten (PK) as Spanish-English DLLs participated in this study when in grade six (Spring, 2018). Participants were originally enrolled in a 5-year multi-site longitudinal study investigating the language bases of reading comprehension from PK (~4 years) to third grade (~9 years) (for details, see (LARRC, Farquharson, & Murphy, 2016) and reconsented in Grade 6 for the current study. The DLL sample resided in Arizona; the monolingual sample in Arizona, Kansas, and Nebraska. Because the samples were recruited in PK, they attended a number of different schools. Lower income levels were more predominant in the DLL sample (see Table 1). The study conforms to the US Federal Policy for the Protection of Human Subjects and was approved by the Institutional Review Boards or Research Ethics Committees at each university. Informed consent was provided by legal guardians and children gave their assent prior to participation.

Measures and procedure

Children were tested individually in a quiet room in their school or at a university lab.

Sight word reading

Children completed the Sight Word Efficiency subtest of the Test of Word Reading Efficiency – Second Edition (TOWRE-2; Torgesen, Wagner, & Rashotte, 1999), which measures the number of English words, ranging from high to low frequency, pronounced correctly in 45 seconds. The test was administered and scored by trained assessors in line with the manual. The average test-retest reliability reported in the manual is .93.

Coherence monitoring

Each child read 18 eight-sentence narrative texts and 18 eight-sentence expository texts written for this age group. The narratives concerned human characters and events focused on typical activities for children, such as parties, schools, and friendships. The expository texts focused on facts about a specific animal and its habitat. The texts were piloted with 11–12-year-olds using a paper and pen task (where children underlined inconsistencies) to check that target inconsistencies were detected. The administration procedure was piloted with a different sample to check that task instructions and feedback during practice items were clear. Coh-matrix text analysis (Graesser, McNamara, & Kulikowich, 2011)¹ confirmed that the two sets of texts did not differ on word concreteness ($M_{\text{narrative}} = 94.83$ (SD = 6.88); $M_{\text{expository}} = 95.61$ (SD = 8.12); $t < 1.00$) but did differ on narrativity ($M_{\text{narrative}} = 50.00$ (SD = 18.57); $M_{\text{expository}} = 22.00$ (16.50); $t(34) = 4.78$, $p < .001$).

Notes

1. We note that our texts were below the 200 word minimum length suggested for reliable Coh-matrix analysis.

Within genre, each text was either fully consistent ($n = 6$) or contained two sentences with contradictory information ($n = 12$), based on materials from previous studies (e.g., LARRC & Yeomans-Maldonado, 2017; Oakhill et al., 2005). There were two versions for each inconsistent text: in one, the critical sentences were separated by 1–2 sentences (near condition, low processing demands); in the other, they were separated by 3–5 sentences (far condition, high processing demands). The inconsistent items were counterbalanced across two presentation lists to ensure that each participant read only one version of each inconsistent text (to avoid priming) and only completed six passages in each condition. The same six consistent passages were used in both lists. Examples are provided in Table 2. The two sets of texts were equated for length ($M_{\text{narrative}} = 100.66$ ($SD = 10.32$), $M_{\text{expository}} = 103.94$ ($SD = 8.79$), $t(34) = 1.03$, $p = .31$) and did not differ in the number of intervening words in the near vs. far inconsistent conditions ($M_{\text{narrative}} = 27.33$ ($SD = 9.79$), $M_{\text{expository}} = 23.17$ ($SD = 7.28$), $t(22) = 1.18$, $p = .25$).

Theoretical interest lies in the difference between the response to sentences in consistent compared to inconsistent near/far conditions. In seeking to estimate the impact of variation in the conditions under which sentences are presented, at the design phase researchers in our field are faced with three options. We could (a) present the same target sentences under all conditions to every participant; (b) present the same target sentences under different conditions to different (sub-groups of) participants; or (c) present different target sentences under different conditions to every participant. Option (a) would be analytically helpful because it could be assumed that any difference in response could not be attributed to differences between-stimuli or between- participants. However, differences between conditions would be confounded with differences in stimulus repetition or order of presentation. Option (b) precludes the risk of confounding the difference between conditions with differences between stimuli but it permits the risk of confounding the difference between conditions with differences between participants or, under counterbalanced designs, between sub-groups of participants. Option (c) precludes the risk of confounding the difference between conditions with differences between participants but it permits the risk of confounding the difference between conditions with differences between stimuli. We chose option (c) because we expected that the inferential risks associated with estimating the effects of conditions in the context of differences between participants would be more important than the risks associated with estimating the differences between conditions in the context of differences between stimuli. Hence, the critical consistent sentences were different from the inconsistent sentences but were presented to elicit responses from the same individuals. (As will be seen, the analysis approach we undertook to examine responses – mixed-effects models – allows us to directly verify our assumptions about the relative balance of inferential risks.)

The coherence monitoring task was run using E-Prime 3.0 (Psychology Software Tools, 2016). Instructions outlined the procedure and included an example of an inconsistency. Children completed two practice passages with feedback to ensure their interpretation of the sense question was focussed on the detection of an inconsistency in the text. They viewed the texts on a laptop, advancing to each new sentence by pressing a key on E-Prime's SR-BOX button box. The reading time for each sentence was recorded. After each text, participants answered a yes/no sense question "Did this text make sense?" (for inconsistent

passages a correct response was “no”; for consistent passages a correct response was “yes”). They also answered a comprehension question to encourage reading for meaning (87% correct responses for monolinguals and 80% for DLLs). Each task took approximately 20 minutes to complete. One child did not complete the narrative task and three did not complete the expository task, however, their partial data were included.

Overview of data analysis

Reading time and sense question accuracy data were analyzed with (generalized) linear mixed-effects models (GLMMs) using the lme4 package for R (Bates, Maechler, Bolker, & Walker, 2015). Models were fitted to estimate the predicted (fixed) effects of critical variables (language status, word reading ability, genre and condition) and their interactions while taking into account random effects associated with differences between sampled children or texts. Categorical variables were contrast coded. Word reading scores were standardized. Models specified with maximal random effects structure (Barr, Levy, Scheepers, & Tily, 2013) did not converge so we report models which did converge and which include all fixed effects plus those random effects that were supported by the data (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017).

Results

We pre-registered our data preparation and analysis plans (<https://osf.io/r69ae>), reporting any deviations from our plans in the following. We share our data and analysis code through OSF (<https://osf.io/sj28g/>).

Responses to sense questions (product of comprehension)

We fitted a GLMM to estimate the effects influencing the log odds that a child’s response to the sense question would be correct, estimating the effects of language status, word reading ability, genre and condition, along with the effects of the interactions between these factors. The mean proportions of correct responses to the sense question in each condition are reported in Table 3. Our model included random effects corresponding to by-items deviations in intercepts, and by-participant deviations in intercepts, and in the slopes of the genre and condition effects. The model summary is reported in Table 4. The positive intercept coefficient shows that participants were more likely to answer the sense question correctly than incorrectly. Given the contrast coding of conditions, the significant positive language status effect (coefficient $B = 0.29$) shows that monolingual children were 6% more accurate and DLLs were 6% less accurate than the grand mean of the sample.² The significant positive genre effect indicates that sense judgments following narratives were more accurate than those following expository texts. The significant negative coefficients

2. In our model the coefficient of the intercept is .80 which represents odds of $\exp(.80) = 2.23$ which in turn represents an overall probability of being correct of $p = 2.23/1 + 2.23 = .69$ which is our models’ best estimate for the grand mean for the data (and is in line with our sample grand mean of $(.68 + .59)/2 = .635$ (See Table 3)). Using our model coefficients, the log odds of being correct for monolingual participants is $.80 + .29 = 1.09$ and for DLL is $.80 - .29 = .51$. The odds of being correct for each group are Monolingual $\exp(.80 + .29) = 2.97$ and DLL $\exp (.80 - .29) = 1.67$. The probability of being correct for each group is Monolingual $2.97/(1 + 2.97) = 0.75$ and DLL $1.67/(1 + 1.67) = 0.63$. Using our model the probability of being correct for Monolingual participants is .06 above the grand mean $(.75 = .69 + .06)$ and for DLL is .06 below the grand mean $(.63 = .69 - .06)$. Therefore the .29 beta coefficient represents an estimated 6% change in accuracy either side of the grand mean for the two language status groups. The 6% estimate is reflected in differences between the actual sample means Monolingual = .68 and DLL = .59 around the actual sample grand mean .635.

representing the effects of differences between the consistent versus the near or far inconsistent conditions show that sense judgments were less accurate for texts containing inconsistencies. The significant positive word reading effect shows that better word reading was associated with more accurate sense judgments.

These main effects were qualified by three significant interactions: language status x genre, language status x condition (near), and genre x condition (near). The nature of these interactions is clearly revealed in Figure 1.

Traditionally, interaction effects have been explored by sub-setting data to examine the effect of one factor (e.g., language) separately at each level of another factor (e.g., genre). There are important concerns about this approach (see Von der Malsburg & Angele, 2017, for a relevant discussion) that render significance tests problematic but the coefficients estimates from such sub-set analyses are helpful as *descriptions* of the average differences between conditions (or groups) in outcomes. Thus, in the following, we report estimates but not *p*-values.

Examination of the genre x language status interaction suggests that monolingual participants were more accurate than DLLs for both genres but that the difference due to language status was greater for narrative compared to expository texts (Figure 1a). If we estimate the effect of language status separately for each genre, we see that it is larger for narrative texts (coefficient $B = 0.37$ ($SE = 0.07$)) than for expository texts ($B = 0.22$ ($SE = 0.06$)). Analysis of the language status x condition interaction (Figure 1b) showed that monolingual children's sense judgments were correct more often than those of DLL children, for consistent (language status effect, $B = 0.52$ ($SE = 0.10$)) and far inconsistent ($B = 0.21$ ($SE = 0.08$)) but not for near inconsistent texts ($B = 0.16$ ($SE = 0.09$)). Analysis of the genre x condition interaction (Figure 1c) showed that sense questions following inconsistent texts were answered accurately more often for narrative than for expository texts in both the near (genre effect, $B = 0.69$ ($SE = 0.15$)) and far ($B = 0.65$ ($SE = 0.16$)) but not the consistent condition ($B = 0.02$ ($SE = 0.16$)). (See OSF <https://osf.io/sj28g/for> interaction model summaries.)

Sentence reading times (process of comprehension)

Our analysis of process focused on the comparison of critical sentence reading times for different sentences read by the same children under different conditions. As explained, we assumed that controlling for between-participant differences was more important than controlling for between-stimulus differences. In our pre-registered design, we had planned to estimate the effect of consistency by comparing critical sentence reading times in consistent, near inconsistent, and far inconsistent passages, congruent with the accuracy analysis. (This analysis is reported on OSF <https://osf.io/sj28g/>). However, we reasoned that while between-stimulus differences were not *as* important we should still seek to minimize them. Analyzing the effect of condition by comparing sentence reading times in between-passage comparisons confounds condition differences with passage differences. Also, critical sentences varied in length, another potential confound. We resolved both problems as follows.

We compared the reading time of each critical sentence in the inconsistent near or far conditions with the time taken to read the sentence located immediately prior (n-1) to the critical sentence in the same text. The n-1 sentence is assumed to have been processed under consistent text conditions because it occurs prior to the critical inconsistent sentence. The comparison of critical and n-1 sentences is within-passage thus removing the confound between passage differences and condition differences. Reading times were scaled to millisecond per word times to remove the potential confound with differences in sentence length. All times were within ± 3 SDs of an individual's condition means. Condition means are reported in Table 5.

Reading times for within-text consistent, inconsistent near and inconsistent far conditions

The process model included the same fixed effects as the accuracy model plus sentence type (consistent, inconsistent). It included random effects accounting for between-participant and between-sentence differences in intercepts, between-participant differences in the effects of genre, condition and sentence type, and between-item differences in the effect of word reading. The coefficients show that faster reading times were associated with monolingual status, narrative texts, and better word reading (Table 6); consistent sentences were read more quickly than inconsistent sentences.

The genre sentence type interaction was statistically significant (illustrated in Figure 2). Analyses of the effect of sentence type, considered separately for narrative and expository texts, indicate that consistent sentences were read more quickly than inconsistent sentences for narratives ($B = -13.74$ ($SE = 2.09$)) but not for expository texts ($B = -2.87$ ($SE = 2.78$); see OSF <https://osf.io/sj28g/for> interaction model summaries).

We assumed that inferential risks were greater in comparing conditions if manipulated between-participants than if manipulated between-stimuli. We can examine this assumption directly. The random effects variances of our analyses (see Tables 6–8) are estimates but, the variances associated with random differences between participants are considerably larger than the random effects variances associated with random differences between stimuli. This suggests it is preferable, given available options, to compare responses to different sentences under different conditions for the same participants.

Critical sentence reading times in correct and incorrect responses

Similar to Helder et al. (2016), we examined whether the pattern of effects differed when children answered correctly or incorrectly. Sentence reading times were analyzed separately conditional on sense question response accuracy.

The coefficients show that, given correct sense judgments, just as in the analysis of all responses, faster reading times were associated with monolingual status, narrative texts, better word reading and consistent sentences (Table 7). There was a genre x sentence type interaction (See Figure 3) where sub-set analyses indicated that consistent sentences were read more quickly than inconsistent sentences in narrative ($B = -18.21$ ($SE = 2.54$)) but not expository texts ($B = -7.07$ ($SE = 4.12$)). There was a significant condition x sentence type interaction, qualified by a three-way interaction with word reading ability (See Figure 3). Sub-set analyses indicated that in the near condition, word reading ability did not modulate

the difference in reading times between consistent and inconsistent sentences ($B = 2.41$ ($SE = 3.35$)) while, in the far condition, better word reading was associated with a greater difference in reading times between the two sentences types ($B = -9.68$ ($SE = 3.62$); see OSF <https://osf.io/sj28g/for> model summaries).

In comparison, given incorrect sense judgments, faster reading times were associated with monolingual status, narrative texts, and better word reading (Table 8) but the effect of sentence type was not significant. There was a significant language status x genre interaction (Figure 4) which, in sub-set analyses, appeared because genre had a significant effect for monolingual ($B = -30.66$ ($SE = 7.72$)) but not DLL participants ($B = -15.23$ ($SE = 9.97$); see <https://osf.io/sj28g/for> model summaries).

Discussion

Our examination of 11–12-year-olds' reading of consistent and inconsistent texts provides unique insight into how reader and text attributes influence coherence monitoring. We extend previous research on coherence monitoring by demonstrating that readers are more likely to detect and report an inconsistency if they are monolingual speakers compared to Spanish-English DLLs (when recruited in PK), if they are better word readers, and when passage content is narrative. The processing demands of the task differentially influenced the strength of the signal that a coherence break was present for good and poor word readers while, overall, participants took longer to read inconsistent sentences only for passages eliciting accurate sense judgments. Our findings suggest that the primary locus of coherence monitoring failure lay in not detecting an inconsistency while reading, rather than not encoding this information.

Similar to previous reading time studies of passage-level coherence monitoring (Harris et al., 1981; Helder et al., 2016; van der Schoot et al., 2012; Zabrocky & Ratner, 1992), children differentiated between consistent and inconsistent text. They took longer to read inconsistent sentences and reported these coherence breaks on completion of the passage. (Agreement on consistent texts was high across genre and language status groups.) Critically, we found reading time differences between consistent and inconsistent sentences only for passages eliciting accurate sense judgments. Thus, we propose that when a coherence break was not reported (incorrect sense judgment after reading) it was not detected during reading. This contrasts with proposals that coherence monitoring failures arise when children detect, but do not encode, a coherence break in their mental model.

Most previous studies of the process of coherence monitoring have used only narrative texts (Harris et al., 1981; Helder et al., 2016; van der Schoot et al., 2012). We found that genre matters, with lower levels of detection during reading and subsequent reporting of coherence breaks for inconsistent expository texts. This pattern suggests that readers are more likely to process expository text in a piecemeal sentence-by-sentence manner, perhaps adopting different standards of coherence for different genres (van den Broek et al., 1995), and thus failing to integrate information across sentences to construct a coherent mental model of the whole text (Denton et al., 2015). We found longer reading times for expository texts in general, further indicating they were more challenging than the narratives (Best et al., 2008;

Graesser et al., 2003). Given recent imaging research showing that narrative and expository text make different processing demands (Aboud et al., 2019), future research should examine how the genre-related differences in standards of coherence or processing that we observe are related to genre differences in content (structure, vocabulary) and reading goals (learning vs. pleasure) (Graesser et al., 2003). Imaging studies using temporal measures could further elucidate the locus of difficulty for different genres.

DLLs were less likely than monolinguals to monitor text for coherence, but the locus of difficulty for both groups appeared to be the same: a failure to detect inconsistencies when reading. The DLLs in our study were not selected (at PK) to be poor comprehenders but previous research, consistent with our findings, suggests that many may have a poor comprehender profile (Lesaux et al., 2010; Mancilla- Martinez & Lesaux, 2011; Nakamoto et al., 2007). Contrary to predictions, the manipulation of the processing demands of our texts did not influence detection or reporting of inconsistencies (see also, Zabucky & Ratner, 1992). However, poorer word readers showed a smaller processing time difference between consistent and inconsistent sentences when the processing demands were high. This suggests that a coherence break is more readily detected by stronger (compared to weaker) readers under high demand conditions. This pattern mirrors that reported for good and poor comprehenders by van der Schoot et al. (2012).

Our study has several important strengths, including the comparison of product and process measures, of genre, and of language groups, besides our use of mixed-effects models to account for random differences. We discuss the limitations here. First, like other work in this field, our groups differed in relation to both language background and socio-economic status. Thus, despite English language schooling from PK, our DLLs may have shown poorer coherence monitoring because of unobserved effects of low-income and language exposure at home (Hoff, 2013). Isolation of the influence of each factor is needed to inform targeted support. Second, we included both product and process measures to permit identification of the most likely source of coherence monitoring failures. Through this approach, our readers were necessarily alerted to the presence of inconsistencies and may have adopted different standards of coherence and strategies compared to “typical” reading. Examining the effect of task instructions on the process of reading could provide important information about the influence of different reading goals, and insight into ways to foster better coherence monitoring. Relatedly, we varied the position of the inconsistent sentence to minimize strategic processing. We note that the inconsistent sentence in Helder et al.’s (2016) materials was always in the same sentence-final position, which may have encouraged more strategic anticipatory processing with the resultant higher accuracy scores than we report here. Finally, our reading time paradigm did not permit examination of whether readers looked back to check preceding text when an inconsistency was detected. Future studies could use eye tracking methods to do this (Connor et al., 2015) and determine whether this behavior differs by reader or genre (Zabucky & Ratner, 1992).

In summary, we have advanced understanding of young readers’ coherence monitoring, indicating for the first time that this skill is weaker in Spanish-English DLLs. Our findings indicate the most likely source of poor coherence monitoring is a failure to construct a coherent mental model when reading, rather than a failure to encode a break when detected.

This is particularly evident for expository texts. Future work should identify how the teaching and activation of relevant vocabulary, background knowledge and reading strategies could support the development of coherence monitoring, a critical skill for learning, and a fundamental skill for expository text comprehension.

Funding

This work was supported by Grant R01HD093003 from the National Institute of Child Health and Human Development. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

Appendix A. Sense question accuracy

Model Specification

In R notation the model for the sense question accuracy main analysis was:

$$\text{Question Accuracy} \sim (\text{Language Status} + \text{Word Reading}) * \text{Genre} * \text{Condition} + (\text{Genre} + \text{Condition} + 1 | \text{Participant}) + (1 | \text{Item})$$

The model formulae are shown (here and for reading times analyses, following) in the style required to specify mixed-effects models for lme4 model fitting functions, to aid results reproducibility. The A*B*C notation requires a model to be fit including the fixed effects of the three-way interaction (A x B x C) as well as all lower-order two-way interactions (A x B, B x C and A x C) and all lower-order main effects (A, B, and C). The random effects are specified in parentheses, including the random effect of participants (|Participant) or of text passage (|Item) on intercepts (. . . 1 | . . .), and the random effect, here, of participants on the slopes of the genre and condition effects (Genre + Condition . . . |Participant).

In standard notation the model for the sense question accuracy main analysis was:

$$\begin{aligned} \text{Question Accuracy} = & \beta_0 + \beta_1 \text{language status}_C + \beta_2 \text{word reading}_C + \beta_3 \text{genre}_C + \\ & \beta_4 \text{condition}_C + \beta_5 \text{language status}_C * \text{genre}_C + \beta_6 \text{word reading}_C * \text{genre}_C + \beta_7 \text{language} \\ & \text{status}_C * \text{condition}_C + \beta_8 \text{word reading}_C * \text{condition}_C + \beta_9 \text{genre}_C * \text{condition}_C + \beta_{10} \text{language} \\ & \text{status}_C * \text{genre}_C * \text{condition}_C + \beta_{11} \text{word reading}_C * \text{genre}_C * \text{condition}_C + u_0 + u_1 \text{genre}_C + \\ & u_2 \text{condition}_C + v_0 + e. \end{aligned}$$

β_0 = fixed intercept, β_{1-11} = fixed effects, C = centered, u_0 = by-participant random intercept, u_{1-2} by-participant random slopes, v_0 = by-item random intercept, e = random error.

Appendix B. Sentence reading times

Model Specification, Within Texts Main Analysis

In R notation the model for the within texts analysis was:

$$\text{Sentence RT} \sim (\text{Language Status} + \text{Word Reading}) * \text{Genre} * \text{Condition} * \text{Sentence type} + (\text{Genre} + \text{Condition} + \text{Sentence type} + 1 | \text{Participant}) + (\text{Word Reading} + 1 | \text{Item})$$

In standard notation the final model for the within texts analysis was:

$$\begin{aligned} \text{Sentence RT} = & \beta_0 + \beta_1 \text{language statusC} + \beta_2 \text{word readingC} + \beta_3 \text{genreC} + \beta_4 \text{conditionC} + \\ & \beta_5 \text{sentence typeC} + \beta_6 \text{language statusC*genreC} + \beta_7 \text{word readingC*genreC} + \beta_8 \text{language} \\ & \text{statusC*conditionC} + \beta_9 \text{word readingC*conditionC} + \beta_{10} \text{genreC*conditionC} + \beta_{11} \text{language} \\ & \text{statusC*sentence typeC} + \beta_{12} \text{word readingC*sentence typeC} + \beta_{13} \text{genreC*sentence typeC} \\ & + \beta_{14} \text{conditionC*sentence typeC} + \beta_{15} \text{language statusC*genreC*conditionC} + \beta_{16} \text{word} \\ & \text{readingC*genreC*conditionC} + \beta_{17} \text{language statusC*genreC*sentence typeC} + \beta_{18} \text{word} \\ & \text{readingC*genreC*sentence typeC} + \beta_{19} \text{language statusC*conditionC*sentence typeC} + \\ & \beta_{20} \text{word readingC*conditionC*sentence typeC} + \beta_{21} \text{genreC*conditionC*sentence typeC} + \\ & \beta_{22} \text{language statusC*genreC*conditionC*sentence typeC} + \beta_{23} \text{word} \\ & \text{readingC*genreC*conditionC*sentence typeC} + u_0 + u_1 \text{genreC} + u_2 \text{conditionC} + \\ & u_3 \text{sentence typeC} + v_0 + v_1 \text{word readingC} + e. \end{aligned}$$

β_0 = fixed intercept, β_{1-23} = fixed effects, C = centered, u_0 = by-participant random intercept, u_{1-3} by-participant random slopes, v_0 = by-item random intercept, v_1 = by-item random slope, e = random error.

Critical sentence reading times in correct responses (within texts).

Model specification, within texts analysis, correct responses.

In R notation the model for the within texts correct responses analysis was:

$$\text{Sentence RT} \sim (\text{Language status} + \text{Word Reading}) * \text{Genre} * \text{Condition} * \text{Sentence type} + (\text{Genre} + \text{Condition} + 1 | \text{Participant}) + (1 | \text{Item})$$

In standard notation the final model for the within texts correct responses analysis was:

$$\begin{aligned} \text{Sentence RT} = & \beta_0 + \beta_1 \text{language statusC} + \beta_2 \text{word readingC} + \beta_3 \text{genreC} + \beta_4 \text{conditionC} + \\ & \beta_5 \text{sentence typeC} + \beta_6 \text{language statusC*genreC} + \beta_7 \text{word readingC*genreC} + \beta_8 \text{language} \\ & \text{statusC*conditionC} + \beta_9 \text{word readingC*conditionC} + \beta_{10} \text{genreC*conditionC} + \beta_{11} \text{language} \\ & \text{statusC*sentence typeC} + \beta_{12} \text{word readingC*sentence typeC} + \beta_{13} \text{genreC*sentence typeC} \\ & + \beta_{14} \text{conditionC*sentence typeC} + \beta_{15} \text{language statusC*genreC*conditionC} + \beta_{16} \text{word} \\ & \text{readingC*genreC*conditionC} + \beta_{17} \text{language statusC*genreC*sentence typeC} + \beta_{18} \text{word} \\ & \text{readingC*genreC*sentence typeC} + \beta_{19} \text{language statusC*conditionC*sentence typeC} + \\ & \beta_{20} \text{word readingC*conditionC*sentence typeC} + \beta_{21} \text{genreC*conditionC*sentence typeC} + \\ & \beta_{22} \text{language statusC*genreC*conditionC*sentence typeC} + \beta_{23} \text{word} \\ & \text{readingC*genreC*conditionC*sentence typeC} + u_0 + u_1 \text{genreC} + u_2 \text{conditionC} + v_0 + e. \end{aligned}$$

β_0 = fixed intercept, β_{1-23} = fixed effects, C = centered, u_0 = by-participant random intercept, u_{1-2} by-participant random slopes, v_0 = by-item random intercept, e = random error.

Critical sentence reading times in incorrect responses (within texts).

Model specification, within texts analysis, incorrect responses.

In R notation the model for the within texts incorrect responses analysis was:

Sentence RT ~ (Language status + Word Reading)*Genre*Condition*Sentence type + (Genre + 1|Participant) + (1|Item)

In standard notation the model for the within texts incorrect responses analysis was:

$$\text{Sentence RT} = \beta_0 + \beta_1 \text{language statusC} + \beta_2 \text{word readingC} + \beta_3 \text{genreC} + \beta_4 \text{conditionC} + \beta_5 \text{sentence typeC} + \beta_6 \text{language statusC*genreC} + \beta_7 \text{word readingC*genreC} + \beta_8 \text{language statusC*conditionC} + \beta_9 \text{word readingC*conditionC} + \beta_{10} \text{genreC*conditionC} + \beta_{11} \text{language statusC*sentence typeC} + \beta_{12} \text{word readingC*sentence typeC} + \beta_{13} \text{genreC*sentence typeC} + \beta_{14} \text{conditionC*sentence typeC} + \beta_{15} \text{language statusC*genreC*conditionC} + \beta_{16} \text{word readingC*genreC*conditionC} + \beta_{17} \text{language statusC*genreC*sentence typeC} + \beta_{18} \text{word readingC*genreC*sentence typeC} + \beta_{19} \text{language statusC*conditionC*sentence typeC} + \beta_{20} \text{word readingC*conditionC*sentence typeC} + \beta_{21} \text{genreC*conditionC*sentence typeC} + \beta_{22} \text{language statusC*genreC*conditionC*sentence typeC} + \beta_{23} \text{word readingC*genreC*conditionC*sentence typeC} + u_0 + u_1 \text{genreC} + v_0 + e.$$

β_0 = fixed intercept, β_{1-23} = fixed effects, C = centered, u_0 = by-participant random intercept, u_1 by-participant random slope, v_0 = by-item random intercept, e = random error.

References

- About KS, Bailey SK, Del Tufo SN, Barquero LA, & Cutting LE (2019). Fairy tales versus facts: Genre matters to the developing brain. *Cerebral Cortex*, 29(11), 4877–4888. doi:10.1093/cercor/bhz025 [PubMed: 30806463]
- Baker L (1984). Spontaneous versus instructed use of multiple standards for evaluating comprehension; effects of age, reading proficiency, and type of standard. *Journal of Experimental Child Psychology*, 38(2), 289–311. doi:10.1016/0022-0965(84)90127-9
- Barr DJ, Levy R, Scheepers C, & Tily HJ (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. doi:10.1016/j.jml.2012.11.001
- Bates D, Maechler M, Bolker B, & Walker S (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. doi:10.18637/jss.v067.i01
- Best RM, Floyd RG, & McNamara DS (2008). Differential competencies contributing to children's comprehension of narrative and expository texts. *Reading Psychology*, 29(2), 137–164. doi:10.1080/02702710801963951
- Connor CM, Radach R, Vorstius C, Day SL, McLean L, & Morrison FJ (2015). Individual differences in fifth graders' literacy and academic language predict comprehension monitoring development: An eye-movement study. *Scientific Studies of Reading*, 19(2), 114–134. doi:10.1080/10888438.2014.943905 [PubMed: 27065721]
- Denton CA, Enos M, York MJ, Francis DJ, Barnes MA, Kulesz PA, . . . Carter S (2015). Text-processing differences in adolescent adequate and poor comprehenders reading accessible and challenging narrative and informational text. *Reading Research Quarterly*, 50(4), 393–416. doi:10.1002/rrq.105
- Ehrlich MF, Remond M, & Tardieu H (1999). Processing of anaphoric devices in young skilled and less skilled comprehenders: Differences in metacognitive monitoring. *Reading and Writing*, 11(1), 29–63. doi:10.1023/A:1007996502372
- Language and Reading Research Consortium; Farquharson K, & Murphy K (2016). Ten steps to a large, multi-site, longitudinal investigation of language and reading in young children. *Frontiers in Developmental Psychology*, 7, 419. doi:10.3389/fpsyg.2016.00419
- Gernsbacher MA (1990). *Language comprehension as structure building*. Lawrence Erlbaum.

- Graesser AC, McNamara DS, & Louwse MM (2003). What do readers need to learn in order to process coherence relations in narrative and expository text? In Sweet AP & Snow CE (Eds.), *Rethinking reading comprehension* (pp. 82–98). Guilford.
- Graesser AC, McNamara DS, & Kulikowich JM (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5), 223–234. doi:10.3102/0013189X11413260
- Harris PL, Kruithof A, Terwogt MM, & Visser T (1981). Children's detection and awareness of textual anomaly. *Journal of Experimental Child Psychology*, 31(2), 212–230. doi:10.1016/0022-0965(81)90013-8
- Helder A, Van Leijenhof L, & van den Broek P (2016). Coherence monitoring by good and poor comprehenders in elementary school: Comparing offline and online measures. *Learning and Individual Differences*, 48, 17–23. doi:10.1016/j.lindif.2016.02.008
- Hoff E (2013). Interpreting the early language trajectories of children from low-SES and language minority homes: Implications for closing achievement gaps. *Developmental Psychology*, 49(1), 4–14. doi:10.1037/a0027238 [PubMed: 22329382]
- Johnson-Laird PN (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge University Press.
- Kim YS (2015). Language and cognitive predictors of text comprehension: Evidence from multivariate analysis. *Child Development*, 86(1), 128–144. doi:10.1111/cdev.12293 [PubMed: 25174258]
- Kintsch W (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.
- Lesaux NK, Crosson AC, Kieffer MJ, & Pierce M (2010). Uneven profiles: Language minority learners' word reading, vocabulary, and reading comprehension skills. *Journal of Applied Developmental Psychology*, 31(6), 475–483. doi:10.1016/j.appdev.2010.09.004 [PubMed: 21243117]
- Lesaux NK, & Harris JR (2017). An investigation of comprehension processes among adolescent English learners with reading difficulties. *Topics in Language Disorders*, 37(2), 182–203. doi:10.1097/TLD.0000000000000120
- Mancilla-Martinez J, & Lesaux NK (2011). The gap between Spanish speakers' word reading and word knowledge: A longitudinal study. *Child Development*, 82(5), 1544–1560. doi:10.1007/s11145-009-9215-5 [PubMed: 21848955]
- Matuschek H, Kliegl R, Vasishth S, Baayen H, & Bates DM (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, 94, 305–315. doi:10.1016/j.jml.2017.01.001
- Nakamoto J, Lindsey KA, & Manis FR (2007). A longitudinal analysis of English language learners' word decoding and reading comprehension. *Reading and Writing*, 20(7), 691–719. doi:10.1007/s11145-006-9045-7
- Oakhill JV, & Cain K (2012). The precursors of reading comprehension and word reading in young readers: Evidence from a four-year longitudinal study. *Scientific Studies of Reading*, 16(2), 91–121. doi:10.1080/10888438.2010.529219
- Oakhill JV, Hartt J, & Samols D (2005). Levels of comprehension monitoring and working memory in good and poor comprehenders. *Reading and Writing*, 18(7–9), 657–713. doi:10.1007/s11145-005-3355-z
- Perfetti CA, & Stafura J (2014). Reading comprehension: Including word knowledge in a theoretical framework. *Scientific Studies of Reading*, 18(1), 22–37. doi:10.1080/10888438.2013.827687
- Psychology software tools I (2016). E-Prime 3.0. Retrieved from <https://www.pstnet.com>
- Rapp DN, & Kendeou P (2007). Revising what readers know: Updating text representations during narrative comprehension. *Memory & Cognition*, 35(8), 2019–2032. doi:10.3758/BF03192934 [PubMed: 18265617]
- Rapp DN, & van den Broek P (2005). Dynamic text comprehension: An integrative view of reading. *Current Directions in Psychological Science*, 14(5), 276–279. doi:10.1111/j.0963-7214.2005.00380.x
- Torgesen JK, Wagner R, & Rashotte C (1999). *TOWRE–2 Test of Word Reading Efficiency*. Austin, TX: Pro-Ed.

- van den Broek PW, Risden K, & Husebye-Hartman E (1995). The role of readers' standards for coherence in the generation of inferences during reading. In Lorch RF & O'Brien EJ (Eds.), *Sources of coherence in reading* (pp. 353–373). Lawrence Erlbaum Associates, Inc.
- van der Schoot M, Reijntjes A, & van Lieshout EC (2012). How do children deal with inconsistencies in text? An eye fixation and self-paced reading study in good and poor reading comprehenders. *Reading and Writing*, 25(7), 1665–1690. doi:10.1007/s11145-011-9337-4 [PubMed: 23293428]
- Von der Malsburg T, & Angele B (2017). False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of Memory and Language*, 94, 119–133. doi:10.1016/j.jml.2016.10.003 [PubMed: 28603341]
- Language and Reading Research Consortium; Yeomans-Maldonado G (2017). Development of comprehension monitoring in beginner readers. *Reading and Writing*, 30(9), 2039–2067.
- Zabrocky K, & Ratner HH (1992). Effects of passage type on comprehension monitoring and recall in good and poor readers. *Journal of Reading Behavior*, 24(3), 373–391. doi:10.1080/10862969209547782

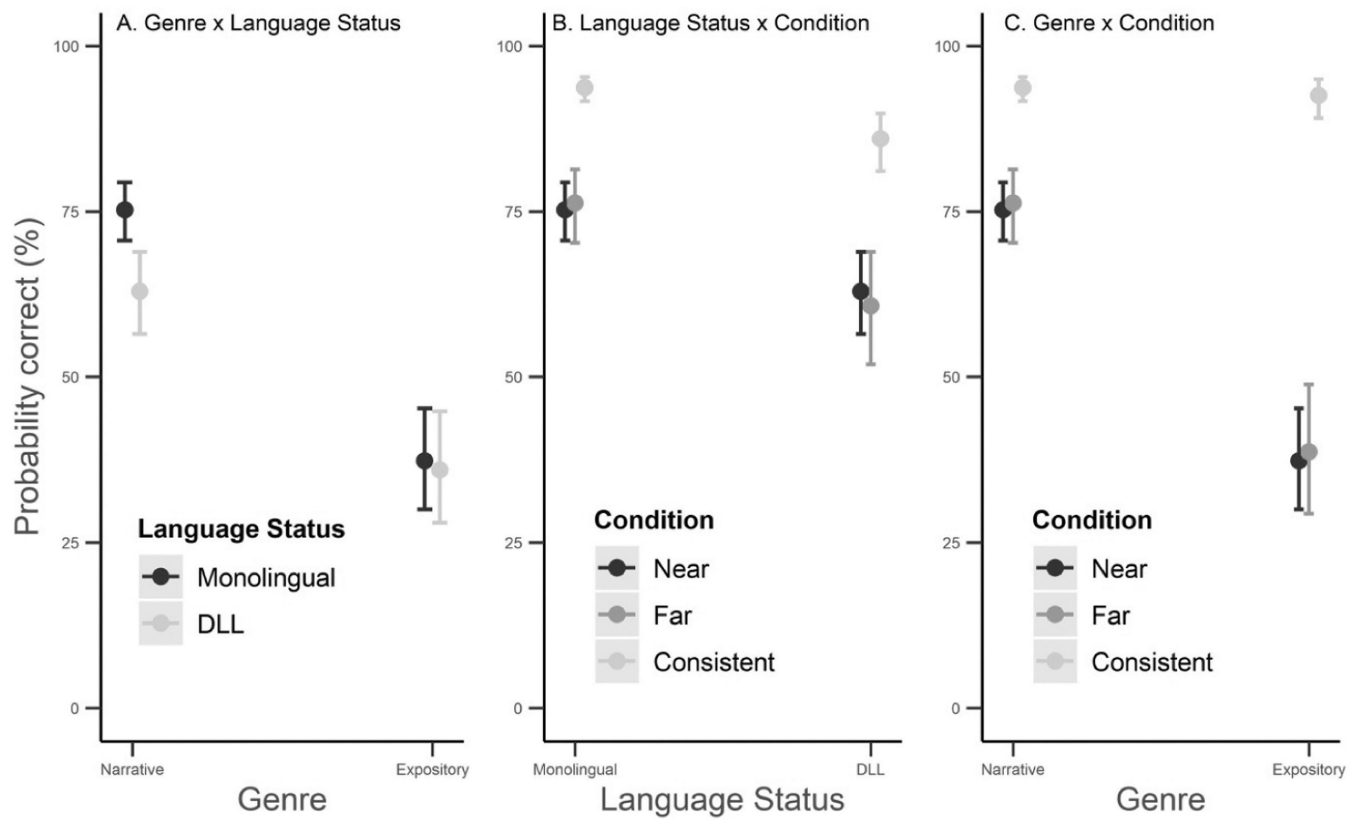


Figure 1. Graphical representations of the significant interactions in the sense question accuracy model (Table 4).

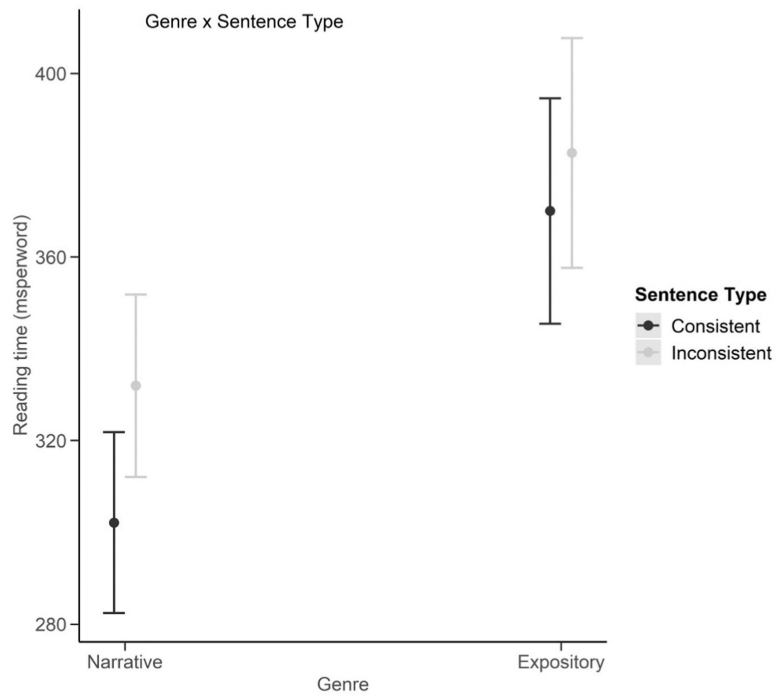


Figure 2. Graphical representation of the significant genre x sentence type interaction in the all data sentence reading time model (Table 6).

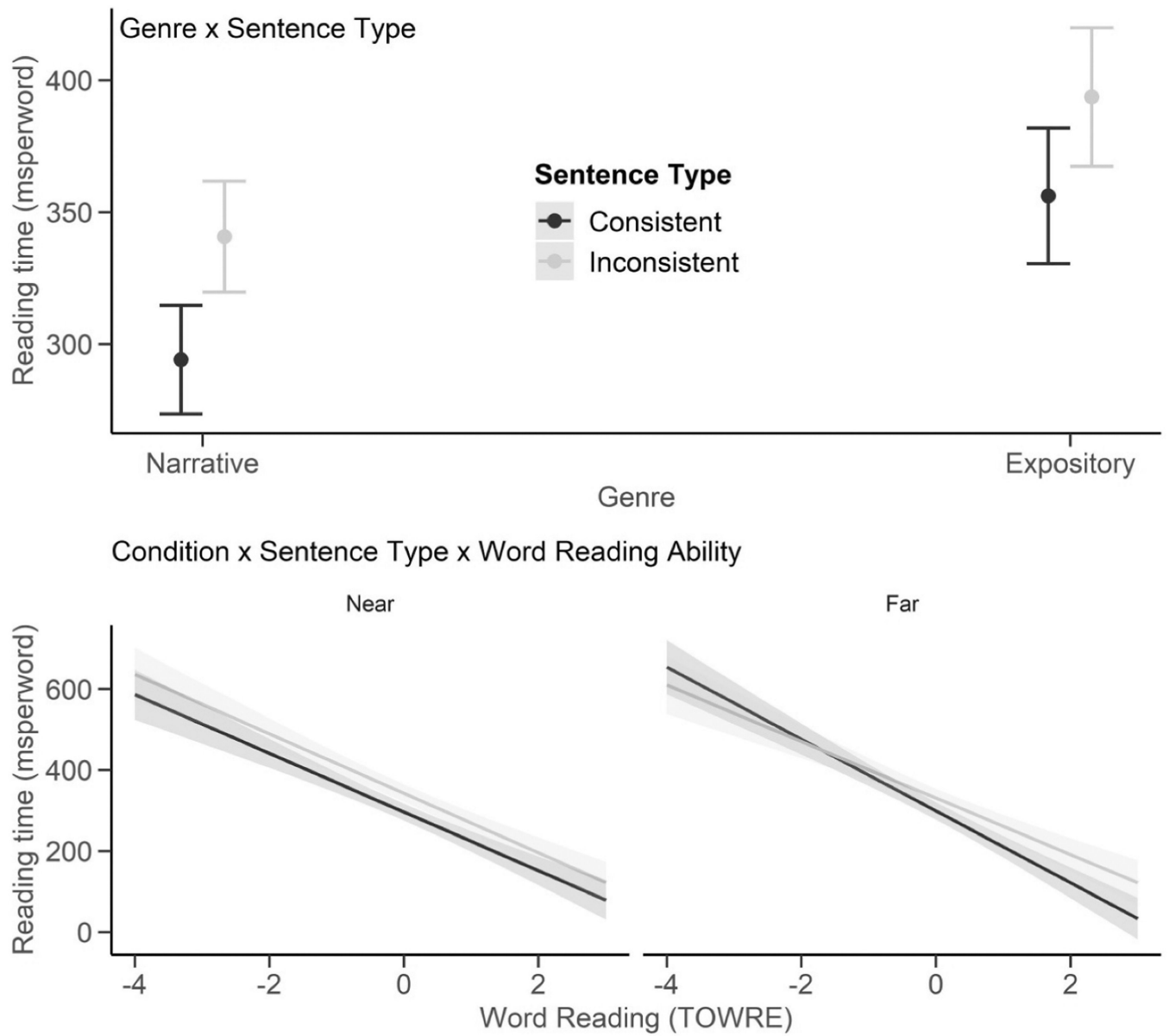


Figure 3. Graphical representation of the significant interactions in the correct only sentence reading time model (Table 7).

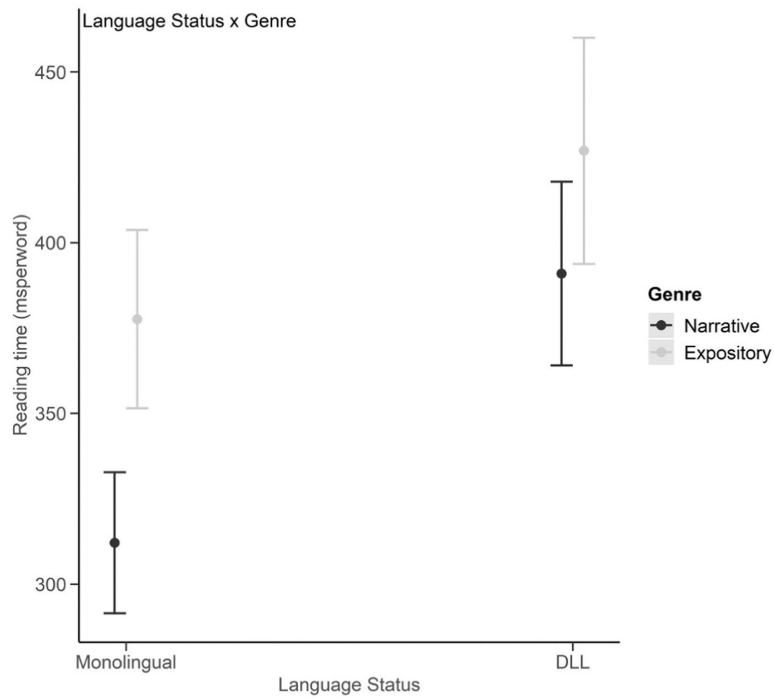


Figure 4. Graphical representation of the significant language status x genre interaction in the incorrect only sentence reading time model (Table 8).

Table 1.

Demographic characteristics of English monolingual and Spanish-English dual language learners.

	English monolingual speakers	Spanish-English Dual Language Learners
<i>N</i> (%female)	85 (44%)	94 (57%)
Age	12 years 1 month	12 years 1 month
SWE (raw scores)	*78.48 (10.50)	75.32 (9.79)
SWE (standard scores)	103.49 (15.60)	99.06 (13.42)
Income *	< 20 k	0
	20, 001– 40 k	10
	40,001– 60 k	9
	60, 0001– 80 k	10
	> 80 k	55
Father/Male Guardian's Education Level	< High school	2
	High school	10
	Some college	12
	Associates/Technical degree	7
	Bachelor's degree	26
	Post graduate degree	26
Mother/Female Guardian's Education Level	< High school	0
	High school	3
	Some college	17
	Associates/Technical degree	8
	Bachelor's degree	25
	Post graduate degree	31
Free/reduced lunch	12	85

* 1 non-responder to SWE and Income. SWE refers to the Sight Word Efficiency subtest of the TOWRE-2 (Torgesen, Wagner, & Rashotte, 1999).

Table 2.

Examples of narrative and expository passages in the inconsistent near and far conditions and the consistent condition.

Narrative Inconsistent-Near	Expository Inconsistent-Near
<p>Sarah got some roller skates for her birthday. She had never skated before and was surprised at how fast she could skate along the sidewalk. Dad had warned her that she must be very careful not to go too fast, until she got the hang of it. <i>All of a sudden Sarah fell over and very badly hurt her arm.</i> Dad took her to the hospital to get checked by a doctor. <i>The doctor took an X-Ray of Sarah's leg.</i> The hospital was very busy and they had a long wait ahead of them. Dad promised he would buy Sarah an ice cream on the way home to cheer her up.</p>	<p>The monarch butterfly is America's most familiar butterfly. Its wings have a recognizable black, orange, and white pattern. <i>Monarch butterflies flap their wings more slowly than any other butterfly.</i> They migrate up to three thousand miles each fall. They then fly back again in the spring, traveling up to 350 miles a day. <i>They can make this incredible journey because they can flap their wings more quickly than any other butterfly.</i> The monarch is the only butterfly known to make a two-way migration as birds do. Adult monarch butterflies feed off the nectar of wildflowers and the blossom on fruit trees.</p>
<p><u>Narrative Inconsistent-Far</u></p> <p>Sarah got some roller skates for her birthday. She had never skated before and was surprised at how fast she could skate along the sidewalk. <i>All of a sudden Sarah fell over and very badly hurt her arm.</i> Dad had warned her that she must be very careful not to go too fast, until she got the hang of it. Dad took her to the hospital to get checked by a doctor. The hospital was very busy and they had a long wait ahead of them. <i>The doctor took an X-Ray of Sarah's leg.</i> Dad promised he would buy Sarah an ice cream on the way home to cheer her up. SENSE QUESTION: Did this story make sense? NO COMPREHENSION QUESTION: Did Sarah and her dad have to wait at the hospital? YES</p>	<p><u>Expository Inconsistent-Far</u></p> <p>The monarch butterfly is America's most familiar butterfly. Its wings have a recognizable black, orange, and white pattern. <i>Monarch butterflies flap their wings more slowly than any other butterfly.</i> They migrate up to three thousand miles each fall. They then fly back again in the spring, traveling up to 350 miles a day. The monarch is the only butterfly known to make a two-way migration as birds do. <i>They can make this incredible journey because they can flap their wings more quickly than any other butterfly.</i> Adult monarch butterflies feed off the nectar of wildflowers and the blossom on fruit trees. SENSE QUESTION: Does this passage make sense? NO COMPREHENSION QUESTION: Can monarch butterflies fly more than 500 miles a day? NO</p>
<p><u>Narrative Consistent</u></p> <p>Olivia always gets up early in the morning to get ready for school. She often helps to get her little brother Liam ready for school too. This morning she tied her brother's shoelaces and combed his messy hair. Their mom was busy filling their water bottles and getting the lunch boxes ready. Eventually they were all ready to leave the house. Olivia skipped down the garden path and jumped over the little wall at the end of the garden. Liam is only five years old and very short. He always tries to copy her and ends up falling over the wall instead. SENSE QUESTION: Did this story make sense? YES COMPREHENSION QUESTION: Did Liam comb his own hair? NO</p>	<p><u>Expository Consistent</u></p> <p>Tortoises are land-dwelling reptiles, with hard protective shells. Because they are reptiles, female tortoises lay eggs. Female tortoises do not sit on their eggs like a bird. Instead, they lay the eggs in a burrow and cover them with sand and soil to stay warm. Tortoises are the longest living land animal in the world. Some species of tortoise live for more than 150 years. Their age can be estimated from the rings on the pattern on their shell. The rings can be counted in the same way that we count rings on a tree to estimate its age. SENSE QUESTION: Does this passage make sense? YES COMPREHENSION QUESTION: Do tortoises sit on their eggs? NO</p>

Bold italicized text indicates inconsistent information in the inconsistent passages. The same consistent passages were used in both lists.

Table 3.

Mean proportion of correct responses (and Standard Deviations) for the sense question.

Genre	Condition	Language Status		Total
		Monolingual	DLL	
Narrative	Consistent	0.91 (0.29)	0.80 (0.40)	
	Near	0.70 (0.46)	0.60 (0.49)	
	Far	0.72 (0.46)	0.58 (0.49)	0.72 (0.45)
Expository	Consistent	0.90 (0.30)	0.75 (0.43)	
	Near	0.41 (0.49)	0.39 (0.49)	
	Far	0.42 (0.49)	0.39 (0.49)	0.54 (0.50)
Total		0.68 (0.47)	0.59 (0.49)	

“Yes” and “no” are the correct responses for consistent and inconsistent texts, respectively.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 4.

Summary GLMM for (log odds) sense question accuracy.

Fixed effects	Estimated coefficient	SE	<i>z</i>	<i>p</i>
(Intercept)	0.80	0.12	6.61	<.001
Language status	0.29	0.06	5.06	<.001
Word reading	0.14	0.06	2.40	.02
Genre	0.05	0.11	4.40	<.001
Condition (Near)	-0.66	0.10	-6.86	<.001
Condition (Far)	-0.65	0.10	-6.79	<.001
Language status x Genre	0.08	0.04	2.02	.04
Word reading x Genre	-0.02	0.04	-0.50	.62
Language status x Condition (Near)	-0.13	0.06	-2.15	.03
Language status x Condition (Far)	-0.08	0.06	-1.33	.19
Word reading x Condition (Near)	-0.07	0.06	-1.10	.27
Word reading x Condition (Far)	-0.07	0.06	-1.13	.26
Genre x Condition (Near)	0.19	0.09	2.17	.03
Genre x Condition (Far)	0.16	0.09	1.87	.06
Language status x Genre x Condition (Near)	0.05	0.04	1.20	.23
Language status x Genre x Condition (Far)	0.08	0.04	1.76	.08
Word reading x Genre x Condition (Near)	-0.03	0.04	-0.71	.48
Word reading x Genre x Condition (Far)	-0.04	0.04	-0.82	.41
Random effects			Variance	<i>SD</i>
Participant	(intercept)		1.07	1.03
	Genre		0.24	0.49
	Condition		2.43	1.56
Item	(intercept)		0.40	0.63
R ² marginal ^c = 0.22, R ² conditional ^d = 0.45				

Observations = 6336^a; Participants = 178^b; Items = 36.^aOne participant had missing narrative data and 3 had missing expository data.^bOne Monolingual child did not have a TOWRE score. R² calculated using the MuMIn package in R.^crepresents the variance explained by the fixed effects,^drepresents the variance explained by the entire model including both fixed and random effects. Effects in bold are statistically significant. All categorical fixed effects were contrast coded in order to be able to interpret the lower order (main) effects. Language status: Monolingual = +1, DLL = -1; Genre: Narrative = +1, Expository = -1; Condition: Near = +1, Far = +1, Consistent = -1. TOWRE scores were centered and scaled. See Appendix A for the model specification in R and in standard notation.

Table 5.

Critical sentence reading times (Milliseconds per Word).

Text Type	Condition	Language Status	
		Monolingual	DLL
Narrative	Inconsistent Near	318.91 (151.35)	410.01 (230.55)
	Inconsistent Far	317.73 (148.62)	397.73 (196.42)
	Consistent Near	289.54 (143.07)	378.84 (190.66)
	Consistent Far	284.10 (150.27)	380.35 (235.76)
Expository	Inconsistent Near	366.85 (208.25)	438.60 (252.32)
	Inconsistent Far	360.94 (215.47)	426.58 (202.66)
	Consistent Near	356.21 (212.10)	436.19 (214.55)
	Consistent Far	346.75 (200.36)	431.64 (276.24)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 6.

Summary LMM for critical sentence reading time (Milliseconds per Word): within texts.

Fixed effects	Estimated coefficient	SE	<i>t</i>	<i>P</i>
(Intercept)	372.89	10.05	37.12	<.001
Language status	-28.75	7.29	-3.94	<.001
Word reading	-75.03	7.51	-9.99	<.001
Genre	-24.96	8.11	-3.08	.004
Condition (Near)	3.17	1.91	1.65	.10
Sentence type (Consistent)	-8.31	1.79	-4.63	<.001
Language status x Genre	-3.46	4.14	-0.84	.40
Word reading x Genre	-2.11	4.50	-0.47	.64
Language status x Condition (Near)	-0.64	1.94	-0.33	.74
Word reading x Condition (Near)	2.52	1.94	1.30	.20
Genre x Condition (Near)	-0.87	1.73	-0.50	.61
Language status x Sentence type (Consistent)	-2.46	1.82	-1.36	.18
Word reading x Sentence type (Consistent)	-0.59	1.81	-0.33	.74
Genre x Sentence type (Consistent)	-5.43	1.73	-3.14	.002
Condition x Sentence type (Consistent)	-1.03	1.73	-0.60	.55
Language status x Genre x Condition (Near)	-0.37	1.75	-0.21	.83
Word reading x Genre x Condition (Near)	2.02	1.75	1.15	.25
Language status x Genre x Sentence type	1.29	1.75	0.74	.46
Word reading x Genre x Sentence type	-2.22	1.75	-1.27	.20
Language status x Condition x Sentence type	1.19	1.75	0.68	.50
Word reading x Condition x Sentence type	2.95	1.75	1.69	.09
Genre x Condition x Sentence type	-0.41	1.73	-0.24	.81
Language status x Genre x Condition x Sentence type	0.25	1.75	0.14	.89
Word reading x Genre x Condition x Sentence type	0.85	1.75	0.49	.63
Random effects			Variance	<i>SD</i>
Participant	(intercept)		8573.03	92.59
	Genre		9562.21	97.79
	Condition		478.87	21.88
	Sentence type		159.31	12.62
Text	(intercept)		1177.55	34.32
R ² marginal ¹¹ = 0.17, R ² conditional = 0.45	Word reading		76.27	8.73

Consistent Near and Far are the comparison sentences (n-1) from the inconsistent passages used in the analysis. DLL = dual language learner.

Observations = 8448^a, Participants = 178^b, Texts = 24.

^aThere were two items (consistent/inconsistent) per text. One participant had missing narrative data and 3 had missing expository data.

^bOne Monolingual child did not have a TOWRE score. R² calculated using the MuMIn package in R,

^crepresents the variance explained by the fixed effects,

d^2 represents the variance explained by the entire model including both fixed and random effects. Effects in bold are statistically significant. All categorical fixed effects were contrast coded in order to be able to interpret the lower order (main) effects. Language status: Monolingual = +1, DLL = -1; Genre: Narrative = +1, Expository = -1; Condition: Near = +1, Far = -1, Sentence type: Inconsistent = +1, Consistent = -1. TOWRE scores were centered and scaled. See Appendix B for the model specification in R and in standard notation.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 7.

Summary LMM for reading time (Milliseconds per Word): Correct responses.

Fixed effects	Estimated coefficient	SE	<i>t</i>	<i>P</i>
(Intercept)	374.86	10.49	35.73	<.001
Language status	-28.57	7.50	-3.81	<.001
Word reading	-73.00	7.41	-9.85	<.001
Genre	-25.87	8.90	-2.91	.006
Condition	1.91	2.48	0.77	.44
Sentence type	-12.64	2.29	-5.52	<.001
Language status x Genre	-3.22	4.94	-0.65	.51
Word reading x Genre	-3.06	4.83	-0.63	.52
Language Status x Condition	0.08	2.52	0.03	.98
Word reading x Condition (Near)	0.67	2.46	0.27	.78
Genre x Condition (Near)	0.60	2.38	0.25	.80
Language status x Sentence type (Consistent)	-1.44	2.34	-0.62	.54
Word reading x Sentence type (Consistent)	-2.44	2.28	-1.07	.29
Genre x Sentence type (Consistent)	-5.57	2.29	-2.43	.01
Condition x Sentence type (Consistent)	-5.79	2.29	-2.53	.01
Language status x Genre x Condition	-0.25	2.42	-0.11	.92
Word reading x Genre x Condition	2.49	2.36	1.06	.29
Language status x Genre x Sentence type	0.07	2.34	0.03	.98
Word reading x Genre x Sentence type	-2.05	2.28	-0.90	.37
Language status x Condition x Sentence type	-1.21	2.34	-0.52	.61
Word reading x Condition x Sentence type	4.85	2.28	2.13	.03
Genre x Condition x Sentence type	0.72	2.29	0.31	.75
Language status x Genre x Condition x Sentence type	2.55	2.34	1.09	.27
Word reading x Genre x Condition x Sentence type	0.13	2.28	0.06	.95
Random effects			Variance	<i>SD</i>
Participant	(intercept)		8676.70	93.15
	Genre		11,036.10	105.05
	Condition		316.30	17.79
Text	(intercept)		1320.70	36.34

R² marginal¹¹ = 0.20, R² conditional = 0.48Observations = 4450^a, Participants = 175^b, Texts = 24.^aThere were two items (consistent/inconsistent) per text.^bOne Monolingual child did not have a TOWRE score. Three participants did not respond correctly to any of the items. R² calculated using the MuMIn package in R.^crepresents the variance explained by the fixed effects,^drepresents the variance explained by the entire model including both fixed and random effects. Effects in bold are statistically significant. All categorical fixed effects were contrast coded in order to be able to interpret the lower order (main) effects. Language status: Monolingual = +1,

DLL = -1; Genre: Narrative = +1, Expository = -1; Condition: Near = +1, Far = -1, Sentence type: Inconsistent = +1, Consistent = -1. TOWRE scores were centered and scaled. See Appendix B for the model specification in R and in standard notation.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 8.

Summary LMM for narrative text reading time (Milliseconds per Word): incorrect responses.

Fixed effects	Estimated coefficient	SE	<i>t</i>	<i>P</i>
(Intercept)	369.83	10.52	35.15	<.001
Language status	-32.49	8.16	-3.98	<.001
Word reading	-79.81	8.21	-9.72	<.001
Genre	-23.42	8.16	-2.87	.007
Condition	2.57	3.06	0.84	.40
Sentence type	-2.37	2.82	-0.84	.40
Language status x Genre	-9.34	4.64	-2.01	.05
Word reading x Genre	-6.62	4.72	-1.40	.16
Language Status x Condition	-1.09	3.09	-0.35	.72
Word reading x Condition (Near)	3.87	3.18	1.22	.22
Genre x Condition (Near)	-0.88	2.95	-0.30	.77
Language status x Sentence type (Consistent)	-1.82	2.83	-0.65	.52
Word reading x Sentence type (Consistent)	2.09	2.93	0.72	.47
Genre x Sentence type (Consistent)	-2.34	2.82	-0.83	.41
Condition x Sentence type (Consistent)	4.39	2.82	1.56	.12
Language status x Genre x Condition	-0.68	2.97	-0.23	.82
Word reading x Genre x Condition	2.88	3.07	0.94	.35
Language status x Genre x Sentence type	3.27	2.83	1.16	.25
Word reading x Genre x Sentence type	-0.95	2.93	-0.33	.74
Language status x Condition x Sentence type	3.36	2.83	1.19	.23
Word reading x Condition x Sentence type	1.27	2.93	0.44	.66
Genre x Condition x Sentence type	1.10	2.82	0.39	.70
Language status x Genre x Condition x Sentence type	-0.61	2.83	-0.22	.83
Word reading x Genre x Condition x Sentence type	0.23	2.93	0.08	.94
Random effects			Variance	<i>SD</i>
Participant	(intercept)		8242.00	90.79
	Genre		7814.70	88.40
	Condition		461.20	21.47
Text	(intercept)		1063.80	32.62

R² marginal¹¹ = 0.15, R² conditional = 0.42Observations = 3998^a, Participants = 178^b, Texts = 24.^aThere were two items (consistent/inconsistent) per text.^bOne Monolingual child did not have a TOWRE score. R² calculated using the MuMIn package in R,^crepresents the variance explained by the fixed effects,^drepresents the variance explained by the entire model including both fixed and random effects. Effects in bold are statistically significant. All categorical fixed effects were contrast coded in order to be able to interpret the lower order (main) effects. Language status: Monolingual = +1, DLL = -1; Genre: Narrative = +1, Expository = -1; Condition: Near = +1, Far = -1, Sentence type: Inconsistent = +1, Consistent = -1. TOWRE scores were centered and scaled. See Appendix B for the model specification in R and in standard notation.