Examining the cognitive costs of counterfactual language comprehension: Evidence from ERPs

Heather J Ferguson

James E Cane

University of Kent

Correspondence to:
Heather Ferguson
School of Psychology
Keynes College
University of Kent
Canterbury
Kent CT2 7NP
England, UK

email: h.ferguson@kent.ac.uk
Tel: +44 (0) 1227 827120 Fax: +44 (0) 1227 827030

Abstract

Recent empirical research suggests that understanding a counterfactual event (e.g. 'If Josie had revised, she would have passed her exams') activates mental representations of both the factual and counterfactual versions of events. However, it remains unclear *when* readers switch between these models during comprehension, and whether representing multiple 'worlds' is cognitively effortful. This paper reports two ERP studies where participants read contexts that set up a factual or counterfactual scenario, followed by a second sentence describing a consequence of this event. Critically, this sentence included a noun that was either consistent or inconsistent with the preceding context, and either included a modal verb to indicate reference to the counterfactual-world or not (thus referring to the factual-world). Experiment 2 used adapted versions of the materials used in Experiment 1 to examine the degree to which representing multiple versions of a counterfactual situation makes heavy demands on cognitive resources by measuring individuals' working memory capacity. Results showed that when reference to the counterfactual-world was maintained by the ongoing discourse, readers correctly interpreted events according to the counterfactual-world (i.e. showed larger N400 for inconsistent than consistent words). In contrast, when cues referred back to the factual-world, readers showed no difference between consistent and inconsistent critical words, suggesting that they simultaneously compared information against both possible worlds. These results support previous dual-representation accounts for counterfactuals, and provide new evidence that linguistic cues can guide the reader in selecting which world model to evaluate incoming information against. Crucially, we reveal evidence that maintaining and updating a hypothetical model over time relies upon the availability of cognitive resources.

Keywords: Counterfactual conditionals; discourse comprehension; event-related potentials; N400; working memory

1. Introduction

Counterfactual thinking refers to an understanding of events that are 'counter to reality' or false, and involves the comparison of reality to a hypothetical alternative (Fauconnier & Turner, 2003). For example, understanding a counterfactual sentence, such as, 'If Josie had revised over Christmas, she would have passed her exams', requires the reader to temporarily accept false information as true (that *Josie revised over Christmas*), and to accommodate subsequent events according to that hypothetical model of the world (that *she passed her exams*). However, readers must also infer the implied facts from the conditional frame (that *Josie did not revise over Christmas*), in order to understand the sentence's implied meaning (that *she failed her exams*).

Counterfactual worlds can be triggered by numerous linguistic structures, including negatives, verbs (e.g. wish or hope), if-then conditionals, and modal terms (e.g. could or might), and their plausibility can be further influenced by manipulating tense (Cowper, 1999; Kratzer, 1991). A long history of theoretical research has postulated on the nature of the mental representations that are activated by counterfactual utterances. Much of this work stems from general models of language comprehension, which suggest that to understand a text readers must construct a mental representation of the information (Garnham, 1981; Johnson-Laird, 1983). Fauconnier (1985, 1997) suggests that in the case of counterfactual conditionals two such mental spaces are produced; one is the reality and the other is the counterfactual hypothetical space. Counterfactuality, therefore, is a case of forced incompatibility between these two spaces. Byrne and colleagues (e.g. Byrne, 2002, 2005; Byrne & Tasso, 1999; Johnson-Laird & Byrne, 2002) advocate a similar mental model approach, whereby the models represent logically equivalent situations that are valid possibilities. Thus, reasoners represent both the true possibilities and the contrary-to-fact possibilities, while keeping track of each model's epistemic status. The aim of the current paper is to examine reader's ability to switch between these factual and counterfactual representations of the world, and to assess whether suppressing access to the alternative

representation is cognitively effortful.

Traditionally, research on counterfactuals has focused on reasoning, and what sort of constraints there are on the kinds of counterfactual thoughts people are likely to generate in a variety of circumstances (e.g. Kahneman & Miller, 1986; Byrne, 1997; Markman & Tetlock, 2000). However, given the largely theoretical nature of this work, very little was known about how counterfactual mental models might be set up and accessed online during understanding. In recent years, this question (and others) has begun to be empirically tested by psychologists, with an emerging body of research testing comprehenders' ability to make inferences about counterfactual events. Some of this work has demonstrated that comprehenders have good access to the counterfactual model of the world. For example, using the visual world paradigm, Ferguson, Scheepers and Sanford (2010) showed that listeners can modify their expectations about forthcoming events according to a fictional counterfactual version of the world (e.g. "If cats were vegetarians…"). Analysis of eye movements around a concurrent visual scene (depicting, among other things, a *fish* and some *carrots*) revealed that listeners anticipate reference to the contextually appropriate counterfactual-world object (*carrots*) at least 200ms before the auditory onset of this critical word. Similarly, using an eye-tracked reading task, Ferguson and Sanford (2008) showed that readers can accommodate a novel counterfactual world, leading to a reversal of typical real-world anomaly detection effects when that information was presented within an appropriate counterfactual context (see also Ferguson, Sanford, & Leuthold, 2008 who showed similar effects for negated counterfactual contexts). This pattern was evidenced by longer total reading times on critical words that were inconsistent with the described counterfactual world (Ferguson & Sanford, 2008 only), increased regressions out from the words that directly followed this critical word (showing that readers regressed back in the text to make sense of the inconsistency), and longer 'go past' reading times on this post-critical region (showing that readers re-read earlier portions of text to make sense of the difficulty).

Using neuroimaging methods, Nieuwland has shown that counterfactual events are just as

easy to integrate as factual events, provided that the propositional truth-value of the statement is high. In this research, Nieuwland and Martin (2012) recorded event-related brain potentials (ERPs) while participants read negated counterfactual sentences (e.g. "If N.A.S.A. had not developed its Apollo project, the first country to land on the moon would have been Russia/America"), and real-world sentences (e.g. "Because N.A.S.A. developed its Apollo project, the first country to land on the moon was America/Russia"). Results showed that false words elicited larger N400 components compared to true words. The fact that these N400 responses were indistinguishable between real-world and counterfactual contexts suggests that readers can integrate the consequences of events in a counterfactual world with no delay, and without interference from the implied factual version of events. Similar results have been found using unrealistic counterfactual sentences (e.g. "If dogs had gills, Dobermans would breathe underwater"), indicating that the ability to make counterfactual inferences does not rely on pre-existing knowledge about reality (Nieuwland, 2013). Nieuwland (2012) has also presented evidence that different neural circuits might underlie these truth-value judgements for real-world and counterfactual-world utterances. Specifically, research using fMRI showed that while real-world and counterfactual false sentences elicited comparable activity increases in the left inferior frontal gyrus, counterfactuals elicited a larger truth-value difference in the right inferior frontal gyrus compared to real-world sentences.

Complementing this research showing rapid access to the explicitly described counterfactual world, is recent empirical evidence showing fast access to the inferred factual world, based on a counterfactual context. For example, Ferguson (2012) reports two experiments where participants were eye-tracked while they read short narratives, in which a context sentence set up a realistic counterfactual world (e.g. "If Joanne had remembered her umbrella, she would have avoided the rain"), and a subsequent critical sentence described an event that was either consistent or inconsistent with the implied factual world ("By the time she arrived at school, Joanne's hair was wet/dry"). Results showed that readers detected the contextually inconsistent

word in the earliest moments of reading, leading to an increased incidence of regressive eye movements in this condition, and increased total reading times on the inconsistent word.

Using self-paced reading and probe word tasks, de Vega and colleagues have also demonstrated rapid access to the factual world within a counterfactual narrative (de Vega & Urrutia, 2012; de Vega, Urrutia, & Riffo, 2007). Here, participants read short stories that included a factual (e.g. "As Juan had enough time, he went to the cafe to drink a beer") or counterfactual context ("If Juan had enough time, he would have gone to the cafe to drink a beer"), then subsequently responded to a probe word that was either related to the initial events of the story (*typing*), or to the recently described event (*drink*). Results showed that the initial situation was more accessible within a counterfactual story than a factual story, as evidenced by shorter probe response times. Similar effects were found using an ERP task (Urrutia, de Vega, & Bastiaansen, 2012), where continuation sentences that referred back to the initial situation elicited more negative ERP waves when they followed a factual discourse than a counterfactual discourse. These findings suggest that readers cancel discourse updating following a counterfactual; readers make the inference that the described events did not occur.

Intriguingly, emerging experimental evidence seems to support the previously described theoretical models, in showing that counterfactuals activate representations of *both* the counterfactual and factual versions of the world, possibly even simultaneously. Early evidence for these dual mental representations comes from Santamaria and colleagues (2005), who observed that counterfactual utterances such as, "If it had rained, the plants would have bloomed" prime faster comprehension of both negative (i.e., factual; not *p* and not *q*) and affirmative (i.e., counterfactual; *p* and *q*) conjunctions. Other researchers (de Vega et al., 2007; de Vega & Urrutia, 2012; Urrutia et al., 2012) advocate a two-stage model of counterfactual understanding, whereby the alternative "as if" meaning of counterfactuals is momentarily activated alongside the presupposed factual meaning, but is later suppressed in favour of the factual version of events (thus explaining the 'cancelled updating' effects described above). Evidence for this early dual

meaning comes from the fact that probe words and sentence continuations referring to the new situation were equally accessible following factual and counterfactual contexts.

Indeed, even in the eye-tracking and ERP studies reported above there was evidence of initial interference from the alternative factual/counterfactual world, despite later measures reflecting the overall sentence meaning. For example, Ferguson and Sanford (2008) found that following a counterfactual context (e.g. "If cats were vegetarians…"), continuations that violated real-world assumptions (i.e., cats eating carrots) led to disruptions during the early stages of critical word integration (increased first-pass reading times). Similarly, Ferguson et al. (2008) found that readers' initial interpretation of events following a negated counterfactual context (e.g. "If cats were not carnivores…") reflected a conflict between real-world expectations and this fictional context. That is, reading times on the critical word (first-pass fixation duration and total reading time) did not differ between counterfactual-consistent and -inconsistent continuations, and the N400 response was larger for critical words that violated real-world knowledge than those that violated the counterfactual premise. Further, Ferguson (2012) found that initial reading times on the critical word (first-pass reading times and first fixation duration) were increased even when the continuation was consistent with the counterfactual context. Taken together, these eye-tracking and ERP studies converge to suggest that readers maintain access to both the counterfactual and factual interpretations of events, and when encountering the critical word they simultaneously evaluate incoming information against both versions of the world, just prior to accommodation into the wider counterfactual proposition. This final interpretation of events appears to be guided by cues from the evolving linguistic input, with modal verbs, such as "could" and "would", signaling a continuation of the counterfactual world, and factual cues (e.g. "That evening…"), indicating reference back to the factual version of events. However, so far, observations of these simultaneous mental representations have been reported as incidental 'interference' effects in tasks that primarily aimed to establish the speed of factual or counterfactual inferences within the wider discourse's propositional meaning. Thus, to our

knowledge, no study has directly tested whether these dual representations might be selectively accessed online given cues from the ongoing discourse. This direct comparison will be conducted here.

Here we describe two experiments that aimed to test the relative availability of each of these mental representations within a counterfactual discourse, by assessing whether readers can switch between factual and counterfactual models of the world during comprehension. To this end, these studies employed an anomaly detection reading paradigm while recording ERPs, and aimed to exploit the brain's clear sensitivity to stimulus predictability and semantic integration processes during language comprehension; the N400 ERP (Kutas & Hillyard, 1980). This component is a centroparietally distributed, negative-going deflection in the ERP, which peaks approximately 400 ms after word-onset. Extensive research in psycholinguistics and cognitive neuroscience has shown that the amplitude of this ERP component is directly influenced by inconsistencies of local and contextual information (e.g. Hagoort, Hald, Bastiaansen, & Petersson, 2004; Van Berkum, Zwitserlood, Hagoort, & Brown, 2003), as well as the degree of plausibility for the overall message (DeLong, Quante, & Kutas, 2014). For example, typical N400 responses to local semantic anomalies (e.g. *the peanut was <u>in love</u>*) have been shown to reverse within an appropriate discourse context that makes the described events plausible (Filik & Leuthold, 2013; Nieuwland & Van Berkum, 2006). Similarly, N400 effects are activated when a narrative describes a character's inappropriate emotional response to a given social situation (Leuthold, Filik, Murphy, & MacKenzie, 2012), when a statement conflicts with a person's moral values (Van Berkum, Holleman, Nieuwland, Otten, & Murre, 2009), or when their actions mismatch their (false) beliefs (Ferguson, Cane, Douchkov, & Wright, in press). Importantly for the current study, plausibility effects have been shown to supersede typical sentence-level semantic anomaly responses within an appropriate counterfactual discourse (Nieuwland, 2013; Nieuwland & Martin, 2012). Thus, by comparing N400 amplitudes for world-consistent and world-inconsistent conditions we aimed to examine to what extent counterfactual and factual

representations were being used to interpret the unfolding discourse.

The current experiments recorded ERPs while participants read short narratives in which a counterfactual context set up a realistic hypothetical situation (as in (1)).

(1) If it had rained this morning, Susan would have rushed to get to work.

In this example, readers need to construct a hypothetical version of the world in which the described events- that it rained this morning and Susan rushed to get to work- are true. However, the conditional framing of this statement means that readers should also make the inference that in fact, it did not rain this morning and therefore she did not rush to get to work. In order to assess the relative *availability* of each of these representations of the world, the linguistic structure of a second critical sentence that described a consequence of the preceding context was manipulated so that it either implied reference to the factual world (as in (2), below), or included a modal inflection that signalled a continuation of events according to the counterfactual world (as in (3)). Crucially, this second sentence contained a sentence-final word that was either consistent or inconsistent with the overall preceding context. Thus, when referring back to the factual interpretation of events, it makes sense for Susan to arrive at work late. In contrast, when the modal verb 'would' implies reference to the counterfactual version of events, it makes sense for Susan to arrive at work early. Note that neither of these critical words is semantically anomalous, and both 'early' and 'late' are plausible and relevant words given the general topic (i.e. someone going to work); it is the fit with the preceding factual/counterfactual premise that defines the consistent/inconsistent event. To confirm this, factors such as plausibility, predictability and semantic relatedness (which are known to modulate the N400) will be controlled for in the current experiments.

(2) In the end, Susan arrived at work late.

(3) In the end, Susan would have arrived at work early.

Stories depicting factual contexts (established with the conjunction, "because" instead of "if") were included as a baseline of normal contextual integration.

The two experiments presented here aim to test the hypothesis that readers maintain access to both the counterfactual and factual world possibilities within a discourse, and use cues from the evolving linguistic input to direct their expectations about subsequent events. Based on previous research, three possible patterns of effect can be predicted. First, if readers have established a full representation of the counterfactual context and are using this to generate expectations about forthcoming events, then they will elicit a larger N400 for critical words that are inconsistent with that hypothetical world than words that are consistent with that hypothetical world. This prediction assumes that readers do not hold a simultaneous representation of the inferred factual world (or at least are not using this to guide interpretation), therefore processing unfolding events according to the counterfactual model should be just as easy as processing events according to a factual model. Alternatively, if readers have cancelled updating of the counterfactual scenario in favour of the inferred factual model, then we might expect to see a reversed inconsistency effect whereby critical words that continue the counterfactual world (but are inconsistent with the factual world) elicit larger N400 responses compared to those that are inconsistent with the counterfactual world. Finally, if the counterfactual context prompts readers to represent both the explicit counterfactual scenario and the implied factual scenario, with equal access to both, then they we should see evidence of a conflict in the earliest moments of integrating the critical word. This conflict could be evidenced by the absence of an N400 effect between consistent and inconsistent conditions, as each is partially activated by the respective contexts.

2. Results

2.1 Experiment 1

Grand average ERP waveforms are presented for the factual conditions in Figure 1, for the counterfactual-counterfactual conditions in Figure 2, and for the counterfactual-factual conditions in Figure 3. It can be seen that following factual and counterfactual-counterfactual contexts, inconsistent target words triggered a more negative-going deflection (N400) than consistent target words, starting at about 250ms after target word onset. In contrast, this pattern appears to be reversed for the counterfactual-factual condition, with consistent target words eliciting a slightly more negative-going N400 component from around 400ms after target word onset. Scalp topographies of these consistency effects in each condition are shown in Figure 4.

---------- Figures 1, 2 and 3 here ----------

---------- Figure 4 here ----------

2.1.1 Early N400 Analyses (300-400ms)

Analysis of the N400 amplitude in the 300-400ms time interval revealed a significant main effect of consistency over lateral electrodes [$F(1, 29) = 5.95$, $p < .05$, ${}_p\eta^2 = .17$], showing a more negative-going N400 wave overall for inconsistent compared to consistent critical words (.56 *vs.* .82 μV). A significant interaction between ant-pos and context [$F(2, 58) = 3.11$, $p < .05$, ${}_p\eta^2 = .1$] revealed similar amplitude ERP waveforms across context conditions over anterior sites, but a more negative-going deflection for critical words following a counterfactual-counterfactual context (.83 μV), compared to either a factual (1.07 μV) or counterfactual-factual context (1.11 μV), over posterior sites. In addition, there was a significant three-way interaction between ant-pos, context and consistency [$F(2, 58) = 3.56$, $p < .05$, ${}_p\eta^2 = .11$]. Examining effects over anterior sites revealed no significant effects [all $F < 1$], however, analysis over posterior sites revealed a main effect of consistency [$F(1, 29) = 6.91$, $p < .01$, ${}_p\eta^2 = .19$], such that overall, the N400 was more negative-going for inconsistent compared to consistent conditions (.78 *vs*. 1.23 μV).

Moreover, there was a significant context * consistency interaction [$F(2, 58) = 3.96$, $p < .05$, $_p\eta^2 = .12$], which was further examined by comparing effects of consistency at each context level. A comparable interaction was revealed over the midline electrodes, with a significant three-way interaction between electrode, context and consistency [$F(4, 116) = 2.53$, $p < .05$, $_p\eta^2 = .08$]. Here, the context * consistency interaction was only significant at electrode Pz [$F(2, 58) = 3.36$, $p < .05$, $_p\eta^2 = .11$].

Factual contexts revealed a clear effect of consistency (lateral: $t(29) = 2.61$, $p < .01$; midline: $t(29) = 2.24$, $p < .05$), reflecting the expected increased N400 amplitude following a contextually inconsistent target word compared to a consistent target word (.74 *vs*. 1.41 μV). Similarly, following a counterfactual-counterfactual context the main effect of consistency was significant (lateral: $t(29) = 3.01$, $p < .01$; midline: $t(29) = 2.05$, $p < .05$), revealing a more negative-going wave for inconsistent target words compared to consistent target words (.44 *vs*. 1.22 μV). However, no significant difference was found between inconsistent and consistent conditions following a counterfactual-factual context (*t*s < .78; 1.17 *vs*. 1.06 μV).

2.1.2 Late N400 Analyses (400-500ms)

Analysis of the N400 amplitude in the 400-500ms time interval revealed an interaction between ant-pos and consistency over lateral electrodes [$F(1, 29) = 3.11$, $p < .05$, $_p\eta^2 = .1$] and an interaction between electrode and consistency over midline electrodes [$F(2, 58) = 3.82$, $p < .05$, $_p\eta^2 = .12$]. These effects showed that while the N400 was more negative-going for inconsistent than consistent critical words over posterior sites (1.07 *vs*. 1.29 μV), the pattern was reversed over anterior sites (1.15 *vs*. .91 μV). Further, the three-way interactions between ant-pos, context and consistency over lateral sites [$F(2, 58) = 3.21$, $p < .05$, $_p\eta^2 = .1$] and between electrode, context and consistency over midline sites [$F(4, 116) = 3.66$, $p < .01$, $_p\eta^2 = .11$] were significant. In contrast to the earlier N400 time window, in this later N400 window none of the effects reached significance over posterior sites [all Fs < 1.8], however the context * consistency

interaction was significant over anterior sites [$F(2, 58) = 3.15$, $p < .05$, $_p\eta^2 = .1$] and at electrode Fz [$F(2, 58) = 3.11$, $p < .05$, $_p\eta^2 = .1$]. Examining the effects of consistency in each context condition revealed a more positive waveform in the inconsistent condition compared to the consistent condition within a factual context (1.33 *vs*. .82 μV; lateral: $t(29) = 2.02$, $p = .053$; midline: $t(29) = 2.02$, $p = .053$), and within a counterfactual-counterfactual context (1.34 *vs*. .81 μV; lateral: $t(29) = 1.71$, $p = .1$; midline: $t(29) = 2.02$, $p = .053$). No significant difference was found between inconsistent and consistent conditions following a counterfactual-factual context (ts < 1.1).

In summary, results from Experiment 1 showed that readers can detect context-inconsistent information within 400ms of encountering the critical word in factual and counterfactual-counterfactual contexts. This suggests that despite the multiple world representations involved in counterfactuals (as shown in de Vega et al., 2007; de Vega & Urrutia, 2012; Ferguson, 2012; Ferguson & Sanford, 2008; Ferguson et al., 2008; Santamaria et al., 2005; Urrutia et al., 2012), readers had rapid access to the explicitly described counterfactual version of events, which elicited comparable brain responses to those for narratives that involved a single factual version of events. In contrast, no significant difference was found between inconsistent and consistent conditions following a counterfactual context that required readers to refer to the implied factual version of events. This suggests that both versions of the world were equally available to readers at the point of integration, and that readers have not yet switched to favour the factual interpretation of events (as they do in offline tasks, such as sentence completion); the counterfactual version of events continues to interfere with understanding.

2.2 Experiment 2

Experiment 2 aimed to replicate and extend this work using the same reading task as used in Experiment 1, modified in two important ways. Firstly, experimental items were edited so that

critical words were no longer sentence-final, since sentence-final words are likely to reflect a collection of 'wrap-up' processes in language comprehension, which may dilute or mask readers' immediate responses to the critical word (Hagoort, 2003). Thus, we added three 'neutral' words to the final target sentence of each item (identical across conditions within an item). Secondly, previous research has begun to demonstrate the importance of working memory in counterfactual thinking, with studies showing that counterfactual judgements are impaired under high working memory load (Goldinger, Kleider, Azuma, & Beike, 2003), and that reduced working memory capacity correlates with poorer performance on tasks that require counterfactual thinking (Drayton, Turley-Ames, & Guajardo, 2011). However, we are aware of no study that has examined how working memory might influence the incremental processing of counterfactual utterances compared to factual utterances (that only activate representation of a single version of the world). Moreover, recent electrophysiological research has demonstrated that individual differences in executive function skills can impact on the brain's responses to words in a discourse. For example, Nakano, Saron and Swaab (2010) showed that when presented with verb-argument animacy violations in otherwise grammatical sentences (e.g. 'The box is biting the mailman') individuals with low working memory spans elicit an N400 (reflecting lexical/semantic processing) while individuals with high working memory spans elicit a P600 (reflecting combinatorial processes). Similarly, Van Petten, Weckerly, McIsaac and Kutas (1997) showed that individuals with low working memory capacity were less able than individuals with medium or high working memory capacity to use sentence context to facilitate word integration (as indicated by N400 modulations), though they were equally good at using lexical relationships between words. Taken together, these findings suggest that individuals with low working memory capacity may be limited in their ability to interpret events in a discourse according to a counterfactual model of the world.

Therefore, Experiment 2 further examined readers' access to representations of counterfactual and factual worlds, and examined how understanding is influenced by individual

differences in working memory capacity. In this way, we aim to assess whether representing multiple versions of a counterfactual situation makes heavier demands on cognitive resources than factual events that only elicit a single representation. Moreover, by comparing effects between counterfactuals that refer back to the counterfactual *versus* factual version of events we aim to establish the relative availability of each 'world', and the cognitive impact of suppressing the alternative event. This will also allow us to further our understanding of how individual differences in working memory capacity influence language comprehension processes across an extended discourse. Thus, to directly examine the need for cognitive effort involved in counterfactual discourse processing, and whether switching between worlds or inhibiting access to one world is impaired when cognitive resources are low, we divided participants into high and low working memory capacity groups using a median split[1] and compared the direction and timing of anomaly detection effects in each context condition.

Grand average ERP waveforms in each working memory group are presented for the factual conditions in Figure 5, for the counterfactual-counterfactual conditions in Figure 6, and for the counterfactual-factual conditions in Figure 7. Difference waveforms (inconsistent minus consistent), showing the time course and amplitude differences across conditions for high and low working memory groups, are shown in Figure 8. N400 effects were analysed in two N400 time-windows following critical word onset (i.e. 300-400ms and 400-500ms). In addition, the P2 component was analysed between 200 and 250 ms after critical word onset, since visual inspection of the data suggested that differences between conditions/groups might be present prior to the N400 window. All statistical analyses were conducted over lateral electrodes using an ANOVA with context (factual *vs.* counterfactual-counterfactual *vs.* counterfactual-factual), consistency (consistent *vs.* inconsistent), hemisphere (left *vs.* right) and ant-pos (anterior *vs.*

[1] The median split was based on scores from the OSPAN task. The low working memory group consisted of 15 individuals with a mean score of 22.8, and the high working memory group consisted of 15 individuals with a mean score of 35.4 ($t(28) = 6.38, p < .001$).

posterior) as the within-subjects variables, and working memory capacity (high *vs.* low) as the between-subjects variable. ERP amplitudes over midline electrodes were analysed using a context (factual *vs.* counterfactual-counterfactual *vs.* counterfactual-factual) x consistency (consistent *vs.* inconsistent) x electrode (Fz *vs.* Cz *vs.* Pz) x working memory capacity (high *vs.* low) ANOVA.

---------- Figures 5, 6 and 7 here ----------

---------- Figure 8 here ----------

2.2.1 P2 Analyses (200-250ms)

The P2 was maximal over central and anterior scalp regions (lateral: [$F(1, 28) = 43.2$, $p < .001$, $_p\eta^2 = .61$]; midline: [$F(2, 56) = 13.4$, $p < .001$, $_p\eta^2 = .32$]), and was marginally larger in the high working memory group than the low working memory group [lateral: $F(1, 28) = 3.85$, $p = .06$, $_p\eta^2 = .12$; midline: $F(1, 28) = 3.83$, $p = .06$, $_p\eta^2 = .12$]. Interestingly, the interaction between group, context and consistency was significant over lateral electrodes [$F(2, 56) = 3.94$, $p < .05$, $_p\eta^2 = .11$]. Follow-up analyses revealed that the context x consistency interaction was only significant in the high working memory group [$F(2, 28) = 4.81$, $p < .05$, $_p\eta^2 = .26$], reflecting a larger P2 peak for inconsistent than consistent critical words within a counterfactual-factual context ($t(29) = 2.73$, $p < .05$). Comparisons between consistency conditions within factual and counterfactual-counterfactual contexts were not significant (ts < 1). None of the remaining effects reached significance.

2.2.2 Early N400 Analyses (300-400ms)

Analysis of the N400 amplitude in the 300-400ms time interval revealed a main effect of group [lateral: $F(1, 28) = 10.23$, $p < .01$, $_p\eta^2 = .27$; midline: $F(1, 28) = 8.15$, $p < .01$, $_p\eta^2 = .23$], with the low working memory capacity group showing a more negative-going N400 than the high

working memory capacity group (.31 *vs*. 1.47 μV). Over lateral electrodes, the interaction between ant-pos and consistency [$F(1, 28) = 4.37$, $p < .05$, $_p\eta^2 = .14$], and between hemisphere and consistency [$F(1, 28) = 4.17$, $p < .05$, $_p\eta^2 = .13$] was significant. In addition, the interaction between context and consistency was significant [$F(2, 56) = 3.34$, $p < .05$, $_p\eta^2 = .13$], as well as the three-way interactions between ant-pos, context and consistency [$F(2, 56) = 3.92$, $p < .05$, $_p\eta^2 = .12$], and between working memory group, context and consistency [$F(2, 56) = 4.24$, $p < .05$, $_p\eta^2 = .13$]. Over midline electrodes, the three-way interaction between electrode, context and consistency was significant [$F(4, 112) = 2.83$, $p < .05$, $_p\eta^2 = .1$], and the working memory group x context x consistency interaction was marginal [$F(2, 56) = 2.93$, $p = .06$, $_p\eta^2 = .09$].

To follow up the ant-pos * context * consistency interaction, further analyses examined effects over anterior and posterior sites separately. Analysis over anterior sites showed no significant effects [all $Fs < 1$], however over posterior sites the context * consistency interaction was significant [$F(2, 58) = 5.36$, $p < .01$, $_p\eta^2 = .16$]. Thus, the influence of consistency at each context level over the posterior scalp regions was examined. As in Experiment 1, following a factual context, the expected consistency effect was significant ($t(29) = 2.09$, $p < .05$), revealing a more negative-going N400 wave following an inconsistent word than a consistent word (.54 *vs*. 1.43 μV). Similarly, the counterfactual-counterfactual condition showed a comparable main effect of consistency ($t(29) = 2.37$, $p < .05$), with contextually inconsistent target words eliciting a more negative-going wave than a consistent word (.06 *vs*. 1.13 μV). However, analysis of the counterfactual-factual context condition revealed no significant difference between inconsistent and consistent target words ($t = 1.6$; 1.06 *vs*. .52 μV). Over midline electrodes the context * consistency interaction was only significant at electrode Pz [$F(2, 56) = 4.72$, $p < .01$, $_p\eta^2 = .14$]. Here, the consistency effect was only significant within a counterfactual-counterfactual context ($t(29) = 2.13$, $p < .05$), revealing a more negative-going N400 wave following an inconsistent word than a consistent word. The difference was not significant within either a factual or counterfactual-factual context ($ts < 1.6$).

To follow up the group * context * consistency interaction over lateral and midline electrodes, analyses examined effects in the high and low working memory capacity groups separately. Results revealed that individuals with low working memory capacity did not elicit significantly different electrophysiological responses to consistent/inconsistent critical words in any of the context conditions, as shown by the non-significant main effect of consistency and context * consistency interaction [Fs < 1.4]. In contrast, individuals with high working memory capacity showed a significant context * consistency interaction [lateral: $F(2, 28) = 8.85$, $p < .01$, $_p\eta^2 = .39$; midline: $F(2, 28) = 4.36$, $p < .05$, $_p\eta^2 = .24$], reflecting a more negative-going N400 wave over lateral scalp sites for factual-inconsistent compared to factual-consistent critical words (.92 *vs*. 2.06 μV; lateral: $t(14) = 2.16$, $p < .05$; midline: t = 1.18, *ns*). Similarly, the counterfactual-counterfactual condition elicited a more negative-going wave over lateral scalp sites when the critical word was inconsistent (versus consistent) with the context (.22 *vs*. 1.73 μV; lateral: $t(14) = 2.14$, $p < .05$; midline: t = .81, *ns*). However, the counterfactual-factual condition showed the opposite pattern of effect, where a consistent critical word elicited a more negative-going N400 than an inconsistent critical word (.92 *vs*. 2.32 μV; lateral: $t(14) = 2.85$, $p < .01$; midline: $t(14) = 2.71$, $p < .05$). It should be noted however that this pattern is the same as that found on the P200, thus might not reflect a genuine N400 effect at all.

2.2.3 Late N400 Analyses (400-500ms)

In the 400-500ms time interval the main effect of group was only significant over midline electrodes [lateral: $F(1, 28) = 2.54$, $p = .12$; midline: $F(1, 28) = 15.28$, $p < .001$, $_p\eta^2 = .35$], showing that the low working memory capacity group elicited a more negative-going N400 than the high working memory capacity group (.95 *vs*. 2.85 μV). There was also a significant interaction between ant-pos and consistency over lateral electrodes [$F(1, 28) = 13.49$, $p < .001$, $_p\eta^2 = .33$] and an interaction between electrode and consistency over midline electrodes [$F(2, 56) = 7.92$, $p < .001$, $_p\eta^2 = .22$], and three-way interactions between ant-pos, context and consistency

[$F(2, 56) = 6.97$, $p < .005$, $_p\eta^2 = .2$] and electrode, context and consistency [$F(4, 112) = 3.49$, $p < .01$, $_p\eta^2 = .11$]. Analysis over anterior sites found no significant effects [all $Fs < 1$], however analysis over posterior sites showed a significant interaction between context and consistency [$F(2, 58) = 5.57$, $p < .01$, $_p\eta^2 = .16$]. Here, only the factual context condition showed a significant effect of critical word consistency (all other ts < 1), with inconsistent words eliciting a more negative-going N400 wave compared to consistent words (.56 *vs*. 1.95 μV; $t(14) = 3.28$, $p < .005$). Over midline electrodes, the context x consistency interaction was significant at Cz [$F(2, 56) = 3.72$, $p < .05$, $_p\eta^2 = .12$] and Pz [$F(2, 58) = 5.05$, $p < .01$, $_p\eta^2 = .15$], again showing a more negative-going N400 wave for inconsistent words than consistent words within a factual context (Cz: $t(14) = 1.83$, $p = .08$; Pz: $t(14) = 2.81$, $p < .01$). None of the interactions involving group were significant, which is supported by planned comparisons of the consistency effects in each group, which showed that both low (.33 *vs*. 1.37 μV; lateral: $t(14) = 2.22$, $p < .05$; midline: $t(14) = 1.94$, $p = .07$) and high (.8 *vs*. 2.53 μV; lateral: $t(14) = 2.44$, $p < .05$; midline: $t(14) = 2.24$, $p < .05$) working memory capacity groups elicited a more negative-going N400 for factual inconsistent versus consistent critical words. Neither group showed evidence of detecting the inconsistent word in the counterfactual-counterfactual or counterfactual-factual context conditions (ts < 1).

3. Discussion

This paper aimed to examine the relative availability of factual and counterfactual possibilities during the comprehension of a counterfactual discourse. Specifically, it examined whether readers can use cues from the evolving linguistic input to direct their expectations about subsequent events, such that they inhibit access to one of these versions of the world and show preference for the other. Thus, two ERP experiments were conducted where participants read short narratives in which a counterfactual context set up a realistic hypothetical situation (e.g. "If David had been wearing his glasses, he would have been able to read the poster easily"),

followed by a second target sentence that described a consequence of this event. The structure of this target sentence was manipulated so that it either implied reference to the factual world ("From this distance, David found that the words were blurry/clear"), or included a modal inflection that signaled a continuation of events according to the counterfactual world ("From this distance, David would have found that the words were clear/blurry"). N400 amplitude was examined as evidence of the ease with which the consistent/inconsistent final word was integrated into the preceding context.

The ERP results reported in both studies fit with previous psycholinguistic research in showing that context has an early influence on discourse comprehension (Van Berkum et al., 2003; Nieuwland & Van Berkum, 2006). Following a factual context (e.g. "Because David had been wearing his glasses, he was able to read the poster easily"), a more negative-going N400 component was found for critical words that were inconsistent with the pragmatic constraints of the narrative discourse (*clear*), compared to critical words that were consistent with this discourse (*blurry*); this pattern was apparent in both early and late parts of the N400. Thus, the factual context had the effect of easing integration of the consistent word. Importantly, similar effects were found following a counterfactual-counterfactual context, which described the equivalent factual scenario in a hypothetical form. Analysis of the N400 revealed a clear inconsistency effect at the critical word, with a more negative-going N400 over posterior sites in the early time-window following a word that was consistent with the factual version of events (but counterfactual-inconsistent), compared to a word that was consistent with the counterfactual version of events. This early effect was followed by a more positive waveform in the inconsistent condition compared to the consistent condition over anterior sites in the late time-window. These results provide further evidence that readers have set up a mental representation of the described counterfactual world, and can interpret events according to this plausible counterfactual world in the earliest moments of processing (e.g. Ferguson et al., 2010; Ferguson & Sanford, 2008; Nieuwland, 2013; Nieuwland & Martin, 2012). Moreover, the fact that this pattern of

inconsistency detection did not differ between the counterfactual-counterfactual and factual context conditions suggests that similar pragmatic constraints were activated by this counterfactual world as within a factual context.

Interestingly, comparison of effects in individuals with high and low working memory capacity showed that the speed with which readers detected contextually inconsistent input differed according to their available cognitive resources[2]. Specifically, in Experiment 2 N400 responses showed that readers with a high working memory capacity experienced difficulty integrating the factual inconsistency from 300ms after the word onset, suggesting that context rapidly facilitated retrieval of the world-consistent target word. In contrast, readers with a low working memory capacity did not elicit comparable responses to these factual inconsistencies until 400ms after word onset. This finding fits with previous psycholinguistic research, which has shown that the detection of discourse-based anomalies relies upon effortful cognitive resources to update and maintain information within the narrative over time (Sanford & Garrod, 2005; Yang, Perfetti, & Schmalhofer, 2005, 2007). Importantly, we demonstrate that despite this initial delay in integrating incoming information with the wider discourse, individuals with low working memory capacity are able to make context-based inferences online.

Similarly, the results indicated that representing the multiple events in a counterfactual world requires cognitive effort. This effect is partly evidenced in Experiment 1 by the larger N400 within a counterfactual-counterfactual discourse compared to either a factual or a counterfactual-factual discourse. More striking, however, is the finding in Experiment 2 that while individuals with high working memory capacity detected the inconsistent critical word

[2] Note: Whilst the present study employed the OSPAN task (measuring Operation Span) there is growing evidence in support of a domain–general account of Working Memory Capacity (e.g. Kane, Hambrick, Tuholski, Payne et al., 2004; Pardo-Vasquez & Fernandez-Rey, 2012; Chein, Moore, & Conway, 2011). Furthermore, a number of studies have identified positive correlations between Operation Span performance and reading span performance, and relationships between Operation Span and verbal comprehension and ability (Unsworth, Fukuda, Awh, & Vogel, 2014; Turley-Ames & Whitfield, 2003). Therefore, these results may have implications beyond the Operation Span WMC procedure employed.

within a counterfactual-counterfactual context from 300ms after word onset, individuals with low working memory capacity did not show any signs of using the counterfactual representation to distinguish between world-consistent and -inconsistent continuations, either in the early or late N400 windows. Therefore, in contrast to factual inferences (which were simply delayed when cognitive resources were low), counterfactual inferences may not be made at all when working memory capacity is low. Increased cognitive effort during the comprehension of counterfactuals has been observed recently in a study that recorded electrophysiological measures over longer factual/counterfactual narratives (Urrutia et al., 2012). This research reported a 'global' context effect in the final 'wrap-up' words following a counterfactual discourse, with a more negative-going N400 wave and decreased gamma band activity (indicative of semantic unification) compared to a factual context. The authors suggest that these differences might simply be due to the complex semantics of counterfactuals, since the end of a sentence is thought to hold a special status for 'wrap-up' processing (e.g. Just & Carpenter, 1980; Rayner, Kambe, & Duffy, 2000). Importantly here, the fact that individuals with low working memory capacity never showed signs of detecting the inconsistent word within a counterfactual context, but did show delayed detection responses in the factual context, suggests that the difficulties with counterfactuals reflects more than complex semantics. Specifically, we note three key processes that are likely to underpin successful counterfactual language understanding: representing multiple versions of the world, suppressing access to an alternative world, and maintaining a counterfactual world over time (readers must keep track of the epistemic nature of this counterfactual model, given incoming language (Johnson-Laird & Byrne, 1991)).

In contrast to the clear evidence for early access to the explicitly-stated counterfactual world, the ERP data failed to reveal an inconsistency effect following a counterfactual-factual context (there was no difference between consistent/inconsistent critical words in early or late time-windows), and in fact showed an early reversed effect (i.e. larger N400 for contextually-consistent than inconsistent words) in the high working memory capacity group in Experiment 2.

As such, the ERP data could be taken as evidence that the counterfactual model of the world is more available than the factual model. These findings appear to go against previous research which has shown rapid inferences based on the implied factual premise of a counterfactual utterance (Ferguson, 2012; Santamaria et al., 2005; de Vega et al., 2007; de Vega & Urrutia, 2012), as well as the offline sentence completion pre-test data reported here which showed that participants were significantly more likely to complete counterfactual-factual passages using consistent compared to inconsistent critical words (33.6% *vs.* 11.3%). Moreover, analysis of the P2 component showed that this reversed consistency effect in the high working memory group was present from 200ms after critical word onset, meaning that it is unlikely to be a genuine N400 expectancy effect at all. Thus, we propose that the apparent 'lack of effect' in the ERP data here simply reflects delayed suppression of the counterfactual version of events, as readers maintain parallel representations of both possibilities at the point of disambiguation. Thus, incoming information must be simultaneously 'checked' against both possible worlds, which is likely to delay anomaly detection brain responses at the critical word. In Experiment 1 the critical word was sentence-final, which meant that it was not possible to observe any 'switches' of world preference downstream of the critical word. However, in Experiment 2 we extended the final target sentence to include three 'neutral' words (identical across conditions within an item), which allowed us to examine such downstream effects. Statistical analysis of the N400 response to the three wrap-up words did not reveal any significant effects of consistency in the counterfactual-factual condition (Fs < 1.5; see Figure 9), suggesting that both versions of the world remained equally available. Further, the fact that high working memory readers initially showed the reversed pattern of contextual integration, then showed equal responses in the later time-windows, suggests that the explicit counterfactual version of events may momentarily be more accessible than the implied factual version of events.

---------- Figure 9 here ----------

The suggestion of simultaneous checking fits well with results from Ferguson's (2012) eye-tracking study, which showed that readers hold dual-representations of the counterfactual and factual events within a counterfactual narrative. Simultaneous access was evidenced by distinct anomaly detection responses for contextually consistent and inconsistent information upon first encountering the critical word. Specifically, counterfactual consistent continuations prompted readers to spend longer during initial reading of a critical word, but counterfactual inconsistent continuations elicited a higher incidence of first-pass regressions back from that critical word. It is important to note, however, that later measures of reading in Ferguson's study demonstrate that overall, readers favoured the implied factual meaning within a counterfactual-factual discourse. Recall also that this factual interpretation was preferred in the sentence completion and plausibility pre-tests that were reported here. These off-line tests, which allowed participants ample time to evaluate each story's overall meaning, showed the expected consistency effects in both counterfactual context conditions (i.e. higher cloze probabilities and plausibility ratings for contextually consistent compared to inconsistent words). Importantly, these consistency effects did not differ between the two context conditions.

It is likely therefore that differences in paradigm sensitivity might account for the previously reported differences in early/late processing biases of counterfactuals. For example, paradigms that have examined counterfactual understanding within a single sentence (e.g. Nieuwland & Martin, 2012; Nieuwland, 2013) do not report any interference from the alternative world. These studies show that within a conditional *if-then* structure, readers immediately integrate the consequence of a counterfactual premise according to the sentence's overall propositional truth-value. However, paradigms that allow analysis of natural reading over longer narratives, including increased processing time and opportunities to revisit important parts of the text, are able to capture both the initial dual representation as well as the later presupposed factual/counterfactual preference (e.g. de Vega et al., 2007; Ferguson, 2012; Ferguson & Sanford,

2008; Ferguson et al., 2008; Urrutia et al., 2012). The time course of such an effect has been examined by de Vega and Urrutia (2012), who found that immediately following a counterfactual context, both the initial and new information were equally accessible. However, when readers were tested 1500 ms after sentence conclusion, the counterfactual version of events had become less accessible than the implied factual version. The current study demonstrates that access to the counterfactual world can in fact be maintained beyond the counterfactual sentence conclusion, and can continue to influence processing of subsequent events in discourse. Indeed, even when the continuing discourse implied reference back to the factual world, readers here continued to experience interference from the hypothetical counterfactual world.

Finally, in Experiment 2 we observed some differences in processing between the high and low working memory capacity groups. Specifically, the N400 amplitude was larger overall among individuals with low working memory capacity compared to those with high working memory capacity. However, this between-group difference was already present in the P2 peak, with the high working memory capacity group eliciting a more positive P2 than the low working memory capacity group. Since data loss was comparable between groups (high WM = 3%, low WM = 8%), it is unlikely that these differences were due to unequal trial counts. Thus, the group difference could be due to 'irrelevant' individual differences between groups, but is more likely to reflect some general differences in attention (Luck & Hillyard, 1994), or early semantic access in the high working memory capacity group (Martin, Garcia, Breton, Thierry, & Costa, 2014).

In conclusion, the current study has demonstrated the powerful effects that counterfactuals can have on discourse comprehension. The results fit with previous suggestions that counterfactual contexts activate dual representations of the world (e.g. de Vega et al., 2007; de Vega & Urrutia, 2012; Ferguson, 2012; Ferguson & Sanford, 2008; Ferguson et al., 2008; Santamaria et al., 2005; Urrutia et al., 2012), comprising both the explicit hypothetical model (*p* and *q*) and the implied factual model (not *p* and not *q*). Moreover, the results show that linguistic cues provided within a narrative can guide the reader in selecting which of these world models to

evaluate incoming information against. Specifically, when reference to the counterfactual world is maintained by the ongoing discourse (i.e. through the use of modal inflections, 'could' or 'would'), the counterfactual premise is preferentially used as the point of comparison for semantic integration. However, maintaining and updating this hypothetical model over time requires increased cognitive effort, compared to simply representing a single model of the world. In contrast, when cues refer back to the factual world, or are ambiguous, readers perform a simultaneous comparison against both possible worlds (evidenced by the absence of an inconsistency detection response in the N400), which has the effect of delaying evaluation of the overall discourse propositional meaning (offline measures showed a preference for the factual world). Analysis of online brain responses in individuals with high and low working memory capacity, as well as offline plausibility and cloze probability judgements, has provided new insights on the reader's ability to switch between factual and counterfactual versions of the world, and the cognitive effort involved in these processes.

4. Experimental Procedures

4.1 Experiment 1

4.1.1 Participants

A total of thirty native English speakers from the University of Kent took part in Experiment 1 (M_{age} = 20.8, SD_{age} = 2.8), and were either paid for participating or received course credits. Of these, 21 were female, and 26 were right-handed. Handedness was measured using the Oldfield Edinburgh Handedness Inventory (Oldfield, 1971). Participants did not have dyslexia and had vision that they reported to be normal or corrected-to-normal. All participants were naïve to the purpose of the study and had not taken part in any of the experimental item pre-tests.

4.1.2 Materials and Design

One hundred and eighty experimental items were created as in Table 1. Each item consisted of two sentences: Sentence one introduced a factual ("Because…") or a realistic counterfactual ("If…") scenario. The second sentence described a consequence of this event, which either referenced events to the factual world (e.g. "From this distance, Dave found that the words were…"), or included a modal inflection that signalled a continuation of events according to the counterfactual world (e.g. "From this distance, Dave would have found that the words were…"). Crucially, this second sentence contained a sentence-final critical word that was either consistent or inconsistent with the preceding context (e.g. "clear" *vs*. "blurry"). This resulted in a within-subjects design that crossed three levels of context (factual *vs*. counterfactual-counterfactual *vs*. counterfactual-factual) with two levels of consistency (consistent *vs*. inconsistent).

---------- Table 1 about here please ----------

Experimental items were tested and modified using two rounds of pre-tests that tested cloze probability and critical word plausibility. Cloze probability was tested by a total of ninety students from the University of Kent using an online questionnaire platform (Qualtrics). Given the large number of items, the full set of 180 experimental items was split into three sets of 60, and within each set, ten participants completed one of three lists with one version of each item appearing in each list. Items were presented one at a time, truncated before the final critical word, and participants were instructed to complete the sentence with the first sensible word coming to mind. Cloze probability was computed as the percentage of trials that elicited the intended consistent or inconsistent critical words, and modifications were made to the items where a different but related critical word dominated responses (e.g. happy/pleased). Mean cloze probability scores per condition for the final set of items are shown in Table 2. Statistical analyses crossed context with consistency, and revealed a main effect of consistency [$F(1, 89) = 575.55$, $p < .001$, ${}_p\eta^2 = .87$], reflecting a higher probability of participants completing the

sentence using the consistent compared to inconsistent critical word (36.4% *vs.* 8.5%). The main effect of context was not significant [F < .57], however, the context * consistency interaction was significant [$F(2, 178) = 28.51$, $p < .001$, $_p\eta^2 = .24$]. Analysis of the simple main effects revealed that although the previously described pattern of consistency held across all context conditions, this difference was larger following a factual context ($t(89) = 27.97$, $p < .001$) compared to either a counterfactual-counterfactual ($t(89) = 11.91$, $p < .001$) or counterfactual-factual context ($t(89) = 10.63$, $p < .001$). Follow-up tests directly compared effects in each of the counterfactual context conditions using a 2 (context) x 2 (consistency) ANOVA, and revealed a main effect of consistency [$F(1, 89) = 221.03$, $p < .001$, $_p\eta^2 = .71$], but no main effect of context [F < .35] or a context * consistency interaction [F < .06].

Next, critical word plausibility was assessed by a different set of ninety students from the University of Kent using the same online questionnaire platform. Once again, the full set of 180 experimental items was split into three sets of 60, and within each set, five participants completed one of six lists with one version of each item appearing in each list. Items were presented in full, with the final word written in capital letters. Participants were asked to rate each story according to how plausible this final word was, given the preceding context, and used a sliding scale from -2 (highly implausible) to +2 (highly plausible) to indicate their response. Mean critical word plausibility scores per condition are shown in Table 2. Statistical analyses, crossing context with consistency, revealed a main effect of consistency [$F(1, 89) = 574.22$, $p < .001$, $_p\eta^2 = .87$], reflecting higher plausibility ratings for items that contained a consistent compared to an inconsistent critical word (1.05 *vs.* -.69). The main effect of context was also significant [$F(2, 178) = 6.03$, $p < .005$, $_p\eta^2 = .06$], showing that participants rated sentences as overall more plausible in the counterfactual-factual context (.26) compared to either the factual (.14) or counterfactual-counterfactual (.15) contexts. Moreover, the context * consistency interaction was significant [$F(2, 178) = 48.81$, $p < .001$, $_p\eta^2 = .35$]. Similar to the cloze probability results, analysis of the simple main effects revealed that although the previously described pattern of

consistency held across all context conditions, this difference was larger following a factual context ($t(89) = 40.41$, $p < .001$) compared to either a counterfactual-counterfactual ($t(89) = 8.92$, $p < .001$) or counterfactual-factual context ($t(89) = 6.68$, $p < .001$). Follow-up tests comparing effects in each of the counterfactual context conditions revealed a main effect of consistency [$F(1, 89) = 159.3$, $p < .001$, $_p\eta^2 = .64$], and a main effect of context [$F(1, 89) = 7.1$, $p < .01$, $_p\eta^2 = .07$] but no interaction between context * consistency [$F < 1.97$], suggesting that both counterfactual-factual and counterfactual-counterfactual contexts elicited similar effects of consistency on plausibility ratings.

---------- Table 2 about here please ----------

Further pre-tests were performed on the final set of items in order to verify that the low-level properties of the language were matched across conditions. As shown in Table 2, this included comparisons of the semantic relatedness between the critical word and preceding context in each condition (using Latent Semantic Analysis, LSA, Landauer & Dumais, 1997), as well as the critical word length, log-frequency, familiarity, concreteness and imaginability (using the MRC Psycholinguistics Database; Wilson, 1988). Analysis of the LSA indicates the strength of the lexical-semantic relationship between the critical word and the words in the preceding words for each item/condition and was computed separately for the full passage and the target sentence; strong relationships elicit higher LSA values, which are associated with reduced N400 amplitudes. Statistical analysis was performed using an ANOVA crossing context with consistency. Results showed a main effect of context for both the full passage analysis [$F(2, 358) = 97.78$, $p < .001$, $_p\eta^2 = .35$] and the target sentence analysis [$F(2, 358) = 38.01$, $p < .001$, $_p\eta^2 = .18$], showing that overall LSA values were highest in the counterfactual-counterfactual context condition followed by the counterfactual-factual and factual context conditions. Neither the main effect of consistency or the context * consistency interactions reached significance [$Fs < 2.87$, ps

> .09]. Planned comparisons of the consistency effect in each context condition also revealed no significant effects (full passage: *t*s < 1.8, *p*s > .07; target sentence: *t*s < 1, *p*s > .37), however when the full passage was considered there was a trend for higher LSA values for the consistent (versus inconsistent) condition in factual and counterfactual-counterfactual context conditions, but the reverse pattern in the counterfactual-factual context condition. Direct comparison of the effects in the two counterfactual context conditions revealed a main effect of context [full passage: $F(1, 179) = 84.97$, $p < .001$, $_p\eta^2 = .32$; target sentence: $F(1, 179) = 55.48$, $p < .001$, $_p\eta^2 = .24$] (pattern as described above), but no main effect of consistency or interaction between context and consistency [Fs < 2.32, *p*s > .13]. Statistical comparisons between conditions on the remaining measures found no significant differences (All *t*s < 1).

Six presentation lists were created, with each list containing one hundred and eighty experimental items, thirty in each of the six conditions. The one hundred and eighty experimental items in each list were interspersed randomly among ninety unrelated filler sentences to create a single random order and each subject only saw each target sentence once, in one of the six conditions. Five participants were randomly assigned to read each list.

4.1.3 Procedure

Participants were informed about the EEG procedure and experimental task. After electrode application they were seated in a booth where they read the materials from a computer screen. There were four practice trials to familiarize them with the procedure, after which the experimenter answered any questions. Each trial began with the presentation of a single centrally-located red fixation cross for 500ms to signal the start of a new trial. After this time, a white fixation cross appeared for 500ms. Next, the context sentence was presented on the screen, and participants were instructed to read this sentence and press spacebar on a keyboard to continue when ready. A blank screen appeared for 500ms, followed by a fixation cross (500ms). The target sentence was then presented word-by-word, with each word appearing at the centre of the screen

for 300ms, with a 200ms blank-screen interval between words. Target words were always sentence final, and thus appeared with a full stop. A 2500ms blank-screen interval followed each item. There was no secondary task. Trials appeared in ten blocks of twenty-seven trials. Each block was separated by a break, the duration of which was determined by the participant. Thus, participants were tested in a single session that lasted approximately one hour, during which they were seated in a comfortable chair located in an isolated room.

4.1.4 Electrophysiological Measures

A Brain Vision Quickamp amplifier system was used with an ActiCap cap for continuous recording of electroencephalographic (EEG) activity from 62 active electrodes over midline electrodes Fz, Cz, CPz, Pz, POz, and Oz, over the left hemisphere from electrodes Fp1, AF3, AF7, F1, F3, F5, F7, FC1, FC3, FC5, FC7, C1 C3, C5, T7, CP1, CP3, CP5, TP7, A1, P1, P3, P5, P7, PO3, PO7, PO9, O1, and from the homologue electrodes over the right hemisphere. EEG and EOG recordings were sampled at 1000 Hz, and electrode impedance was kept below 10kΩ. Off-line, all EEG channels were recalculated to an average mastoid reference.

Prior to segmentation, EEG and EOG activity was band-pass filtered (0.01-30 Hz, 12 dB/oct), and EEG activity containing blinks was corrected using a semi-automatic ocular ICA correction approach (Brain Vision Analyzer 2). The continuous EEG record was then segmented into epochs of 1200ms, starting 200ms prior to the onset of the target word. Thus, the post-stimulus epoch lasted for a total duration of 1000ms. Semi-automatic artifact detection software (Brain Vision Analyzer 2) was run, to identify and discard trials with non-ocular artifacts (drifts, channel blockings, EEG activity exceeding ± 75μV). This procedure resulted in an average trial-loss of 3.2% per condition.

4.1.5 ERP Data Analysis

For analysis of the EEG data, the signal at each electrode site was averaged separately for each experimental condition time-locked to the onset of the target word. Before the measurement of ERP parameters, the waveforms were aligned to a 200ms baseline prior to the onset of the target word. To analyze experimental effects on the N400, mean ERP amplitude was determined in two time intervals relative to target word onset: an early N400 window between 300-400ms and a late N400 window between 400-500ms. These windows were chosen based on visual inspection of the ERP waveforms and following previous studies that have employed similar experimental designs (De Grauwe et al., 2010; Kreher et al., 2008; Chwilla et al., 2000; Chwilla & Kolk, 2003; Van Petten & Kutas, 1987).

ERP amplitudes were analysed using four regions of interest (ROIs). Lateral electrodes were divided along a left-right dimension, and an anterior-posterior dimension. The two ROIs over the left hemisphere were: left-anterior (Fp1, AF3, AF7, F1, F3, F5, F7, FC1, FC3, FC5, FT7), and left-posterior (CP1, CP3, CP5, TP7, P1, P3, P5, P7, PO3, PO7, O1); two homologue ROIs were defined for the right hemisphere. ERP amplitudes over midline electrodes (Fz, Cz, Pz), where the N400 is maximal, were analysed separately from data recorded over lateral electrode sites.

For the statistical analysis of the N400 in each condition, an ANOVA was performed over lateral electrodes with variables context (factual *vs.* counterfactual-counterfactual *vs.* counterfactual-factual), consistency (consistent *vs.* inconsistent), hemisphere (left *vs.* right) and ant-pos (anterior *vs.* posterior). ERP amplitudes over midline electrodes were analysed using a context (factual *vs.* counterfactual-counterfactual *vs.* counterfactual-factual) x consistency (consistent *vs.* inconsistent) x electrode (Fz, Cz, Pz) ANOVA.

4.2 Experiment 2

4.2.1 Participants

A total of thirty native English speakers from the University of Kent took part in Experiment 2 (M_{age} = 21.7, SD_{age} = 5.1), and were either paid for participating or received course credits. Of these, 21 were female, and 29 were right-handed. Participants did not have dyslexia and had vision that they reported to be normal or corrected-to-normal. All participants were naïve to the purpose of the study and had not taken part in any of the experimental item pre-tests, or Experiment 1.

4.2.2 Materials and Procedure

Materials consisted of one hundred and eighty experimental items, crossing context (factual *vs*. counterfactual-counterfactual *vs*. counterfactual-factual) and consistency (consistent *vs*. inconsistent). Items were identical to those used in Experiment 1 except that three words were appended to the final sentence following the critical word (e.g. "From this distance, Dave (would have) found that the words were [clear/blurry] on the poster"). These three words were the same across the six conditions for each item.

In addition, prior to the main task, participants completed the OSPAN task (La Pointe & Engle, 1990; Turner & Engle, 1989), which measures individuals' working memory capacity. In this task participants responded to a mathematical equation (e.g. (4/2)+3=5), stating whether the answer shown was true or false, then read out loud a word that was presented on the screen. After a series of equation-word pairs participants were required to type in the words they had seen in the order they appeared. This version of the OSPAN task consisted of 12 trials in total, which included 2, 3, 4 or 5 equation-word pairs. The task was run using 'E-prime 2' software, and responses were recorded using the keyboard. Participants pressed 'y' to indicate a correct equation and 'n' for incorrect equations, and pressed the 'space' bar to proceed after reading aloud the word that followed each equation. Working memory capacity scores were calculated by summing the number of words in correctly recalled word sequences; sequences only contributed to the working memory capacity score total where *all* words were recalled correctly in the right

order. Therefore working memory capacity scores could range from 0 - 42, with higher scores indicating higher working memory capacity.

The main task procedure, electrophysiological measures and ERP data analysis was identical to that described in Experiment 1.

Author Notes

This work was carried out with the support of a grant from the Experimental Psychology Society. Thanks are due to Julien Leblond, Caroline Mitra, Meredith Puddefoot, Rowan Quas-Morris, Hannah Skinner, and Maria Gallagher for help with data collection.

References

Byrne, R.M.J. (2005). *The Rational Imagination: How People Create Alternatives to Reality*. Cambridge, M.A.: MIT Press.

Byrne, R.M.J. (2002). Mental models and counterfactual thoughts about what might have been. *Trends in Cognitive Sciences, 6(10),* 426-431.

Byrne, R.M.J. (1997). Cognitive processes in counterfactual thinking about what might have been. *The Psychology of Learning and Motivation, Advances in Research and Theory.* Vol 37. San Diego, CA: Academic Press. pp.105-154.

Byrne, R.M.J. & Tasso, A. (1999). Deductive reasoning with factual, possible, and counterfactual conditionals. *Memory & Cognition, 27(4)*, 726-740.

Chein, J.M., Moore, A.B., & Conway, A.R.A. (2011). Domain-general mechanisms of complex working memory span. *NeuroImage, 54*, 550-559.

Chwilla, D.J., Kolk, H.H.J., & Mulder, G. (2000). Mediated priming in the lexical decision task: Evidence from event-related potentials and reaction time. *Journal of Memory and Language, 42*, 314-341.

Chwilla, D.J. & Kolk, H.H.J. (2003). Event-related potential and reaction time evidence for inhibition between alternative meanings of ambiguous words. *Brain and Language, 86*, 167-192.

Cowper, E. (1999). Grammatical aspect in English. *The Linguistic Review, 16*, 205- 226.

De Grauwe, S., Swain, A., Holcomb, P.J., Ditman, T., & Kuperberg, G.R. (2010) Electrophysiological insights into the processing of nominal metaphors. *Neuropsychologia, 48*, 1965-1984.

DeLong, K., Quante, L., & Kutas, M. (2014). Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia, 61*, 150-162.

de Vega, M., Urrutia, M., & Riffo, B. (2007). Cancelling Updating in the Comprehension of Counterfactuals embedded in narratives. *Memory and Cognition, 35*, 1410-1421.

de Vega, M. & Urrutia, M. (2012). Discourse updating in texts with a counterfactual event. *Psicológica, 33*, 157-173.

Drayton S., Turley-Ames K.J., & Guajardo N.R. (2011). Counterfactual thinking and false belief: The role of executive function. *Journal of Experimental Child Psychology, 108*, 532-548.

Fauconnier, G. (1985). *Mental Spaces: Aspects of Meaning Construction in Natural Language*, Cambridge, MA: MIT Press.

Fauconnier, G. (1997). *Mappings in thought and language*, Cambridge University Press.

Fauconnier, G. & Turner, M. (2003). *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. New York: Basic Books.

Ferguson, H.J. (2012). Eye movements reveal rapid concurrent access to factual and counterfactual interpretations of the world. *Quarterly Journal of Experimental Psychology, 65(5)*, 939-961.

Ferguson, H.J., Douchkov, M., Wright, D., & Cane, J.E. (In Press). Empathy predicts false belief reasoning ability: Evidence from the N400. *Social Cognitive and Affective Neuroscience*.

Ferguson, H.J. & Sanford, A.J. (2008). Anomalies in real and counterfactual worlds: An Eye-Movement Investigation. *Journal of Memory & Language, 58*, 609-626.

Ferguson, H.J., Sanford, A.J., & Leuthold, H. (2008). Eye-movements and ERPs reveal the time-course of processing negation and remitting counterfactual worlds. *Brain Research, 1236*, 113-125.

Ferguson, H.J., Scheepers, C., & Sanford, A.J. (2010). Expectations in counterfactual and theory of mind reasoning. *Language and Cognitive Processes, 25*, 297-346.

Filik, R. & Leuthold, H. (2013). The role of character-based knowledge in online narrative comprehension: Evidence from eye movements and ERPs. *Brain Research, 1506*, 94-104.

Garnham, A. (1981). Mental models as representations of text. *Memory and Cognition*, *9*, 560-565.

Goldinger, S.D., Kleider, H.M., Azuma, T., & Beike, D. (2003). "Blaming the victim" under memory load. *Psychological Science, 14*, 81-85.

Hagoort, P. (2003). Interplay between syntax and semantics during sentence comprehension: ERP effects of combining syntactic and semantic violations. *Journal of Cognitive Neuroscience, 15*, 883-899.

Hagoort, P., Hald, L., Bastiaansen, M.C.M., & Petersson, K.M. (2004). Integration of Word Meaning and World Knowledge in Language Comprehension. *Science, 304 (5669),* 438-440.

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

Johnson-Laird, P.N. & Byrne, R.M.J. (2002). Conditionals: a theory of meaning, pragmatics, and inference. *Psychological Review, 109,* 646-678.

Johnson-Laird, P.N. & Byrne, R.M.J. (1991). *Deduction*. Hove, UK: Erlbaum.

Just, M. & Carpenter, P. (1980). A theory of reading: from eye fixations to comprehension. *Psychological Review, 4*, 329-354.

Kahneman, D. & Miller, D.T. (1986). Norm Theory - Comparing reality to its alternatives. *Psychological Review, 93,* 136-153.

Kane, M.J., Hambrick, D.Z., Tuholski, S.W., Wilhelm, O., Payne, T.W., & Engle, R.W. (2004). The generality of working memory capacity: a latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, 133, 189-217.

Kratzer, A. (1991). Modality. In A von Stechow & D. Wunderlich (Eds.). *Semantik: Ein internationals Handbuch der zeitgnössischen Forschung* (pp. 639-650). Berlin: Walter de Gruyter.

Kreher, D.A., Holcomb, P.J., Goff, D., & Kuperberg, G.R. (2008). Neural evidence for faster and further automatic spreading activation in schizophrenic thought disorder. *Schizophrenia Bulletin, 34*, 473-482.

Kutas, M. & Hillyard, S.A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science, 207*, 203-205.

Landauer, T.K. & Dumais, S.T. (1997). A solution to Plato's problem: The Latent Semanctic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-140.

La Pointe, L.B., & Engle, R.W. (1990). Simple and complex word spans as measures of working memory capacity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 1118-1133.

Leuthold, H., Filik, R., Murphy, K., & Mackenzie, I.G. (2012). The on-line processing of socio-emotional information: Inferences from brain potentials. *Social Cognitive and Affective Neuroscience, 7,* 457-466.

Luck, S.J. & Hillyard, S.A. (1994). Electrophysiological correlates of feature analysis during visual search. *Psychophysiology, 31*, 291-308.

Markman, K.D. & Tetlock, P.E. (2000). 'I couldn't have known': Accountability, foreseeability and counterfactual denials of responsibility. *British Journal of Social Psychology*, *39(3),* 313-325.

Martin, C.D., Garcia, X., Breton, A., Thierry, G., & Costa, A. (2014). From literal meaning to veracity in two hundred milliseconds. *Frontiers in Human Neuroscience, 8* (40).

Nakano, H., Saron, C., & Swaab, T.Y. (2010). Speech and span: Working memory capacity impacts the use of animacy but not of world knowledge during spoken sentence comprehension. *Journal of Cognitive Neuroscience*, *22*, 2886-2898.

Nieuwland, M.S. (2013). "If a lion could speak …": Online sensitivity to propositional truth-value of unrealistic counterfactual sentences. *Journal of Memory & Language, 68*, 54-67.

Nieuwland, M.S. (2012). Establishing propositional truth-value in counterfactual and real-world contexts during sentence comprehension: Differential sensitivity of the left and right inferior frontal gyri. *NeuroImage, 59*, 3433-3440.

Nieuwland, M.S. & Martin, A.E. (2012). If the real world were irrelevant, so to speak: The role of propositional truth-value in counterfactual sentence comprehension. *Cognition, 122*, 102-109.

Nieuwland, M.S. & Van Berkum, J.J.A. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, *18,* 1098-1111.

Pardo-Vazquez, J.L. & Fernandez-Rey, J. (2012). Working Memory Capacity and Mental Rotation: Evidence for a Domain-General View. *The Spanish Journal of Psychology, 15*, 881-890.

Rayner, K., Kambe, G., & Duffy, S.A. (2000). The effect of clause wrap-up on eye movements during reading. *Quarterly Journal of Experimental Psychology, 53*, 1061-1080.

Sanford, A.J. & Garrod, S.C. (2005). Memory-based approaches and beyond. *Discourse Processes, 39*, 205-224.

Turley-Ames, K.J. & Whitfield, M.M. (2003). Strategy training and working memory task performance. *Journal of Memory and Language, 49*, 446-468.

Turner, M.L., & Engle, R.W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, *28*, 127-154.

Unsworth, N., Fukuda, K., Awh, E., & Vogel, E.K. (2014). Working memory and fluid intelligence: Capacity, attention control, and secondary memory retrieval. *Cognitive Psychology, 71*, 1-26.

Urrutia, M., De Vega, M., & Bastiaansen, M. (2012). Understanding counterfactuals in discourse modulates ERP and oscillatory gamma Rhythms in the EEG. *Brain Research, 1455*, 40-55.

Van Berkum, J. J. A., Holleman, B., Nieuwland, M. S., Otten, M., & Murre, J. (2009). Right or wrong? The brain's fast response to morally objectionable statements. *Psychological Science, 20*, 1092-1099.

Van Berkum, J.J.A., Zwitserlood, P., Hagoort, P., & Brown, C.M. (2003). When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect. *Cognitive Brain Research, 17,* 701-718.

Van Petten, C. & Kutas, M. (1987). Ambiguous words in context: An event-related potential analysis of the time course of meaning activation. *Journal of Memory and Language, 26*, 188-208.

Van Petten, C.K., Weckerly, J., McIsaac, H.K., & Kutas, M. (1997). Working Memory Capacity Dissociates Lexical and Sentential Context Effects. *Psychological Science*, *8*, 238-242.

Wilson, M.D. (1988) The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. *Behavioural Research Methods, Instruments and Computers, 20*, 6-11.

Yang, C.L., Perfetti, C.A., & Schmalhofer, F. (2007). Event-related potential indicators of text integration across sentence boundaries. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 55-89.

Yang, C.L., Perfetti, C.A., & Schmalhofer, F. (2005). Less skilled comprehenders' ERPs show sluggish word-to-text integration processes. *Written Language & Literacy, 8*, 233-257.

Figure captions

Figure 1: Grand average ERPs elicited by critical words in the target sentence for factual-consistent and factual-inconsistent conditions, in Experiment 1. Note that negativity is plotted upwards.

Figure 2: Grand average ERPs elicited by critical words in the target sentence for counterfactual-counterfactual-consistent and counterfactual-counterfactual-inconsistent conditions, in Experiment 1. Note that negativity is plotted upwards.

Figure 3: Grand average ERPs elicited by critical words in the target sentence for counterfactual-factual-consistent and counterfactual-factual-inconsistent conditions, in Experiment 1. Note that negativity is plotted upwards.

Figure 4: Topographic maps of the ERP difference waveform for each context condition (inconsistent *minus* consistent) in Experiment 1, for the time intervals 300-400ms and 400-500ms (N400) relative to critical word onset.

Figure 5: Grand average ERPs elicited by critical words in the target sentence for factual-consistent and factual-inconsistent conditions, for the low working memory capacity group (left panels) and the high working memory group (right panels), in Experiment 2. Topographic maps show the ERP difference waveform (inconsistent *minus* consistent).

Figure 6: Grand average ERPs elicited by critical words in the target sentence for counterfactual-counterfactual-consistent and counterfactual-counterfactual-inconsistent conditions, for the low working memory capacity group (left panels) and the high working memory group (right panels), in Experiment 2. Topographic maps show the ERP difference waveform (inconsistent *minus* consistent).

Figure 7: Grand average ERPs elicited by critical words in the target sentence for counterfactual-factual-consistent and counterfactual-factual-inconsistent conditions, for the low working memory capacity group (left panels) and the high working memory group (right panels), in

Experiment 2. Topographic maps show the ERP difference waveform (inconsistent *minus* consistent).

Figure 8: Difference waveforms (inconsistent minus consistent), showing the time course and amplitude differences for factual, counterfactual-counterfactual, and counterfactual-factual context conditions in low (left panels) and high (right panels) working memory groups, in Experiment 2.

Figure 9: Grand average ERPs elicited by the three 'wrap-up' words that followed the critical word in the target sentence for counterfactual-factual-consistent and counterfactual-factual-inconsistent conditions, in Experiment 2.