Published in final edited form as:
J Exp Psychol Learn Mem Cogn. 2011 July ; 37(4): 913–934. doi:10.1037/a0022964.

Event-based Plausibility Immediately Influences On-line Language Comprehension

Kazunaga Matsuki,
University of Western Ontario London, Canada

Tracy Chow,
Teachers College, Columbia University New York, NY

Mary Hare,
Bowling Green State University Bowling Green, OH

Jeffrey L. Elman,
University of California, San Diego San Diego, CA

Christoph Scheepers, and
University of Glasgow Glasgow, United Kingdom

Ken McRae
University of Western Ontario London, Canada

Abstract

In some theories of sentence comprehension, linguistically-relevant lexical knowledge such as selectional restrictions is privileged in terms of the time-course of its access and influence. We examined whether event knowledge computed by combining multiple concepts can rapidly influence language understanding even in the absence of selectional restriction violations. Specifically, we investigated whether instruments can combine with actions to influence comprehension of ensuing patients. Instrument-verb-patient triplets were created in a norming study designed to tap directly into event knowledge. In self-paced reading (Experiment 1), participants were faster to read patient nouns such as *hair* when they were typical of the instrument-action pair (*Donna used the shampoo to wash* vs. *the hose to wash*). Experiment 2 showed that these results were not due to direct instrument-patient relations. Experiment 3 replicated Experiment 1 using eyetracking, with effects of event typicality observed in first fixation and gaze durations on the patient noun. This research demonstrates that conceptual event-based expectations are computed and used rapidly and dynamically during on-line language comprehension. We discuss relationships among plausibility and predictability, as well as their implications. We conclude that selectional restrictions may be best considered as event-based conceptual knowledge, rather than lexical-grammatical knowledge.

Correspondence concerning this article should be sent to either Kaz Matsuki (kmatsuki@uwo.ca) or Ken McRae (mcrae@uwo.ca), Department of Psychology, Social Science Centre, University of Western Ontario, London, Ontario, Canada, N6A 5C2, phone: 519-661-2111 x84688, fax: 519-661-3961..

Keywords

Event knowledge; Selectional restrictions; Plausibility; Thematic roles

Historically, psycholinguists have considered lexical versus broader sentential context effects to reflect qualitatively different representations and processing mechanisms, with lexical effects resulting from fast and automatic processing, but broader sentential effects arising from slower, more controlled processes. A number of theories of sentence comprehension make a clear distinction between knowledge that is encoded in the lexicon, versus world or event knowledge (Bornkessel & Schlesewsky, 2006; Chomsky, 1975; Katz, 1972; Schlesinger, 1995; Sperber & Wilson, 1986). On these accounts, the first is linguistically relevant knowledge, and is used rapidly in the time-course of understanding sentences. The second is not represented within the linguistic system, but instead is part of a comprehender's general knowledge about events (see Jackendoff, 2002, for a summary of this distinction). That is, semantic knowledge, including people's knowledge of generalized events in the world, is activated more slowly than lexical (or syntactic) knowledge, and thus is used only after a short delay (Bornkessel & Schlesewsky, 2006; Clifton & Staub, 2008; Frazier, 1995; van Gompel, Pickering, Pearson, & Liversedge, 2005). In stark contrast, other researchers claim that event-based knowledge is used immediately during on-line sentence comprehension (Delong, Urbach, & Kutas; 2005; Garnsey, Pearlmutter, Myers, & Lotocky, 1997; Kamide, Altmann, & Heywood, 2003; McRae, Spivey-Knowlton, & Tanenhaus, 1998; van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005; Vu, Kellas, Petersen, & Metcalf, 2003).

One key theoretical construct that is central to this debate is that of selectional restrictions. Selectional restrictions refer to knowledge of semantic constraints on verbs' arguments, primarily patients (the entities or objects on which an action is performed). One example of a selectional restriction is ±animacy, which is assumed to be stored in the lexical representations of verbs such as *convince*, reflecting the fact that only animate entities can be convinced (Chomsky, 1965). Furthermore, many researchers argue that selectional restrictions are syntactic in nature, which is an additional reason why they have temporal precedence over more general event-based knowledge (Bornkessel & Schlesewsky, 2006). Although there have been divergent views as to whether selectional restrictions reflect syntactic or semantic knowledge, in many theories, they are considered as lexical information that is temporally privileged in terms of access and use during sentence comprehension. The notion of selectional restrictions has played a central role in interpreting the results of numerous sentence comprehension studies (Altmann & Kamide, 1999; Caplan, Hildebrandt, & Waters, 1994; Trueswell, Tanenhaus, & Garnsey, 1994; Warren & McConnell, 2007).

The purpose of this article is to investigate further the proposed architecturally-determined delay in the use of the general cognitive knowledge of aspects of events. In doing so, we focus on empirical results that recently have been used to argue that event-based knowledge is delayed relative to linguistically-relevant knowledge of selectional restrictions (Warren & McConnell, 2007).

Precedence of Selectional Restrictions

Recent research by Rayner et al. (2004) and Warren and McConnell (2007) obtained results that could be taken to indicate that the use of event-based knowledge during language comprehension is delayed to a certain extent (for related studies, see Filik, 2008; Joseph, Liversedge, Blythe, White, Gathercole, Rayner, 2008; Patson & Warren, 2010; Staub, Rayner, Pollatsek, Hyönä, & Majewski, 2007; Warren, McConnell, & Rayner, 2008). Their studies examined the effects of plausibility on eye movements in reading. They contrasted plausibility effects to either the effects of anomaly (or “theta-assigning relation”, p. 1297, Rayner et al.) or of impossibility (or “selectional restriction violations”, p. 770, Warren & McConnell). The authors used sentences that contained an instrument, a verb, and a patient, and varied the plausibility of the patient by manipulating the instrument-verb combinations. They then examined how this manipulation influenced reading times at the patient noun.

Rayner et al. (2004) used three conditions. In anomalous sentences, the patient (or theme) did not make sense with respect to the verb in that it violated the verb's theta-assigning relation, as in John used a pump to inflate the large carrots for dinner last night. That is, the direct object *carrots* “could not plausibly be assigned the role of theme by the verb” *inflate* because carrots cannot be inflated, regardless of what instrument is used (p. 1292, Rayner et al.). In their control (henceforth, “plausible”) condition, the patient was a good fit for the instrument-verb combination, as in John used a knife to chop the large carrots for dinner last night. That is, it is common to chop carrots, to use a knife for chopping, and more specifically, to use a knife for chopping carrots. Finally, they included an intermediate implausible condition, as in John used an axe to chop the large carrots for dinner last night. In this condition, the patient fit the verb (it is common to chop carrots), the instrument did as well (an axe is used to chop things), but the patient did not fit the instrument-verb combination (people do not chop carrots with an axe).

Similarly, Warren and McConnell (2007) used three conditions: impossible-implausible, possible-implausible (henceforth, “implausible”), and possible-plausible (henceforth, “plausible”) conditions. The impossible-implausible condition differed from Rayner et al.'s (2004) anomalous condition because in all of the items, “the target noun violated the verb's selectional restrictions” (Warren & McConnell, p. 772). In contrast, in Rayner et al.'s anomalous condition, “approximately half of the stimuli … included a selectional restriction violation” (Warren & McConnell, p. 771). The implausible and plausible conditions were similar to Rayner et al's.

In both studies, the key comparison for the present purposes involved the plausible and implausible conditions, which we interpret as corresponding to the degree to which the instrument-verb-patient triplet fits people's knowledge of common events. It is important to note that although Rayner et al. (2004) did not describe their conditions in terms of event knowledge, we do so here. In our view (described in further detail below), the instrument can alter the class of events to which the verb refers. The events in which knives are used for chopping (typically chopping food in a kitchen) differ from those in which axes are used for chopping (typically chopping wood or logs in the great outdoors). One way to think about this difference is that instruments can subtly alter the sense of the verb. Chopping with an axe is a somewhat different sense than is chopping with a knife (e.g., the required movements for these types of chopping differ greatly). Because the instrument-verb combinations cue different types of events, or senses of *chop*, certain patients should be more relevant and common given those combinations. In our view, we would expect immediate differences in reading times at the patient.

However, neither Rayner et al. (2004) nor Warren and McConnell (2007) found immediate differences between their plausible and implausible conditions. When Rayner et al. compared the anomalous to the plausible sentences, an early difference emerged at the patient (*carrots*), with longer gaze durations for the anomalous sentences (in eyetracking reading experiments, effects in first fixation or gaze durations on a target word are considered as “early” or “immediate” effects of a manipulation). The same result obtained when comparing the anomalous and implausible sentences. Likewise, Warren and McConnell found an immediate effect of impossibility at the patient, with longer first fixation durations for the impossible-implausible than the implausible condition. Thus when the verbs' theta-assigning relations and/or selectional restrictions were violated, immediate effects were obtained. Critically for our view, however, for the plausible-implausible comparison, the effects were delayed somewhat. In neither study was there a plausible-implausible difference in first fixation or gaze durations at the patient, or even at the post-target region (*for dinner*). In Rayner et al., there was a marginally significant difference at the patient and a fully significant difference in the post-target region in go-past reading times (also called regression-path durations, which include all fixations starting from the first one in a region and ending when the reader moves past the region; i.e., regressions from that region back to earlier parts of the sentence are included). In Warren and McConnell, a significant plausible-implausible difference emerged in go-past and total reading times at the patient.

In both articles, the authors argued that that the plausibility difference between the implausible and plausible conditions is not of the type that can be detected quickly by readers, and thus does not immediately influence reading times. They further argued that lexically-based information imposed by a verb such as theta-assigning relations or selectional restrictions are privileged in terms of availability, and thus influence reading times immediately. Rayner et al. suggested that readers may have rapidly detected violations because, in most of the anomalous sentences, they “could be detected on the basis of purely lexical information, assuming the information associated with a verb's lexical entry may serve to license certain nouns as verb arguments but not others” (p. 1297). They further stated that, “In the implausible conditions, violations may have been detected more slowly because they may have arisen at a stage of processing after theta assignment, when the target word was integrated into a semantic representation of the sentence fragment up to that point. Our results are therefore consistent with the suggestion that qualitatively different types of processing take place at different stages during sentence comprehension” (p. 1297).

Similarly, Warren and McConnell (2007) concluded that “The finding of earlier processing disruption in conditions with a selectional restriction violation than in conditions without such violations is consistent with the hypothesis that information about a verb's selectional restrictions is privileged over other kinds of knowledge in comprehension. This privilege may be because selectional restriction knowledge is represented in the lexicon and is available earlier than world/contextual knowledge” (p. 774). Warren and McConnell then suggested the possibility that “the current data cannot rule out the possibility that unlikelihood/implausibility had a small effect at the earliest stages of interpretation” (p. 774). However, they went on to conclude that, “A more likely alternative, based on the idea that semantic interpretation begins with a coarse-grained analysis that is subsequently refined (Sanford & Garrod, 2005), is that the knowledge involved in the semantic/thematic fit between a noun and verb may be exactly the kind of coarse grained knowledge recruited during initial interpretation. World or contextual knowledge about the likelihood of a multiparticipant event may generally be recruited later, when the initial coarse interpretation is refined” (p. 774).

These results and interpretations suggest that there is linguistically-relevant information that is accessed immediately from the lexicon. In addition, there is knowledge about common events that is not part of one's lexical knowledge, is computationally more complex, and therefore is computed and used more slowly. Reading time differences at the patient should show up immediately when comparing sentences such as *She ate the carrot* versus *She ate the computer*, because *computer* violates the selectional restrictions of *eat*. In contrast, the knowledge that differentiates *She used the shampoo to wash her filthy hair* from *She used the hose to wash her filthy hair* is available only after a delay because it requires a comprehender to combine their knowledge of the instrument (*shampoo* or *hose*) with types of washing events to establish the difference in plausibility of using one or the other to wash one's hair.

Event Knowledge, Verbs, and Thematic Roles

In direct contrast to the above results and interpretations, a number of theories of language comprehension hinge on the rapid influence of event-based plausibility (Garnsey et al., 1997; Kamide, Altmann, & Heywood, 2003; van Berkum et al., 2005; Vu et al., 2003). In our view, thematic role assignment involves detailed experiential event knowledge that can be activated by individual verbs (McRae, Ferretti, & Amyote, 1997; see also Kamide, Altmann, & Haywood, 2003; Knoeferle, Crocker, Scheepers, & Pickering, 2005). A number of studies suggest that relevant event-based knowledge is used rapidly during sentence comprehension (Hare, Elman, Tabaczynski, & McRae, 2009; Kamide et al., 2003; McRae, Spivey-Knowlton, & Tanenhaus, 1998). Activation of such knowledge has been shown to be interactive and dynamic, such that the kinds of representations being activated are rapidly constrained by the surrounding linguistic cues such as tense, aspect, and syntactic argumenthood.

For example, in an eyetracking study, Morris (1994) found that comprehension of patient nouns such as *mustache* may be facilitated by an event-cuing agent, as in sentence fragments such as *The barber trimmed the* compared to *The person trimmed the*. Although Morris did not statistically test the difference between these conditions, there was a substantial 43 ms difference in gaze duration, and 23 ms difference in first fixation duration. Moreover, several studies that have used the visual world paradigm suggest a role for non-verb elements in generating expectancies. The visual world paradigm features the use of visual scenes depicting objects and/or events to investigate how visual attention (measured by eye-movements) shifts as participants hear sentences. For example, Kamide et al. (2003) presented participants with a picture that contained a girl, a man wearing a helmet, a carousel, and a motorcycle. Participants tended to look more at the carousel upon hearing *ride* in *The girl will ride (the carousel)*, but more at the motorcycle at the same point in *The man will ride (the motorcycle)*. Thus, the preferred patient for *ride* depends on the agent noun phrase. Bicknell, Elman, Hare, McRae, and Kutas (2010) present similar results using both self-paced reading and ERP experiments with analogous materials. These results, like those of Morris (1994), suggest that conceptually-based thematic role knowledge is an important component of sentence comprehension.

The Present Study

Given that a number of theories of language comprehension hinge on the rapid influence of event-based plausibility, the fact that Rayner et al. (2004) and Warren and McConnell (2007) found delayed plausibility effects is significant. The present study further investigates whether plausibility as determined by instrument-verb combinations can influence reading times for ensuing patients in the absence of selectional restriction violations.

Rayner et al. (2004) and Warren and McConnell (2007) collected plausibility ratings for their sentences. In both cases, there was a substantial difference between their plausible and implausible items. However, influences of specific variables on early eye tracking (and other reading time) measures often hinge on detailed, subtle aspects of materials. When studying plausiblity, the stength and homogeneity of both the implausible and plausible conditions determine the probability of observing early differences between them. Upon inspection of their items, there is no reason to doubt that their implausible items were sufficiently implausible. That is, their implausible items sufficiently violate people's knowledge of how real-world events often unfold. Furthermore, most of their plausible items, like the *chop* example presented above, nicely capture people's event knowledge (it is quite common to use a knife to chop carrots). However, some of their plausible items do not seem to match undergraduates' real-world event knowledge to a great extent, and this may have resulted in the lack of effects in first fixation and gaze durations. For example, a plausible sentence from Rayner et al. was The woman used a duster to clean the dirty miniatures sitting on the shelf. It is not highly likely that undergraduate students possess a substantial knowledge regarding dusting minatures. As another example, Warren and McConnell's plausible item, Robert used a trap to catch the large goose that weighed ten pounds, may not correspond to an event that is overly common to the average undergraduate. In other words, the fact that some of the plausible items in their experiments described events that are interpretable by, but not familiar or common to, undergraduate subjects may have weakened the plausible condition to the extent that early eye tracking effects were not found. This issue is revisited in the General Discussion when we discuss plausibility and predictability.

To investigate the strength of plausibility manipulated in our experiments, as well as those of Rayner et al. (2004) and Warren and McConnell (2007), we began by conducting two norming studies. Norming Study 1 used a production task to tap into people's event knowledge of instrument-action-patient triplets (e.g., “what do you cut with a knife?”). We included potential items for our experiments, plus Rayner et al.'s and Warren and McConnell's plausible items for comparison. Norming Study 2 used a sentence plausibility rating task using our items derived from Norming Study 1. Because the major goal was to ensure that our implausible items were implausible but not anomalous, we included Rayner et al.'s plausible, implausible, and anomalous items for comparison.

Experiments 1 (self-paced reading) and 3 (eyetracking) used the same materials to test whether participants rapidly combine instruments and actions to generate expectations for different classes of patients. There were a few differences between our experiments and Rayner et al. (2004) and Warren and McConnell (2007). First, we crossed instruments and patients to completely counterbalance the materials. That is, *Donna used the shampoo to wash her filthy* preceded both *hair* (which we call the typical condition because it reflects our notion of typical events) and *car* (atypical). In addition, *Donna used the hose to wash her filthy* preceded both *car* (typical) and *hair* (atypical). Thus, reading times at the patient were compared in conditions featuring materials identical up to the critical word. Second, we omitted the anomalous condition. In an experiment investigating whether people use their knowledge of real world events when comprehending language, the inclusion of sentences that do not make sense could possibly create a situation in which participants are ostensibly informed that real-world knowledge is somewhat irrelevant in the experimental context. In other words, as stated by Warren and McConnell, the inclusion of anomalous sentences might disrupt natural reading (although they commented on studies with higher proportions of anomalous sentences than in their own, p. 771).

In Experiment 2, we examined whether the results of Experiment 1, in which reading time differences were obtained at the patient noun, might have been due to strong semantic relations or lexical associations between instruments and patients. This is important because the aim was to ensure that the instrument combined with the verb to produce expectancies for the upcoming patient, rather than the instrument acting on its own.

Norming Study 1

The first purpose was to construct items for Experiments 1 and 3 using a methodology designed to tap directly into people's knowledge of the types of things that are acted upon using specific instrument-action combinations. This provides an empirical basis for the distinction between the typical and atypical instrument-verb-patient triplets. The second purpose was to use the same norming method to measure the strength of Rayner et al.'s (2004) and Warren and McConnell's (2007) plausibility manipulation. To do so, we included their plausible items and compared them to the items used in the present experiments.

Method

Participants—One-hundred and five undergraduates at the University of Western Ontario participated for course credit. In all studies reported herein, all participants were native English speakers, had normal or corrected-to-normal visual acuity, and participated in only one study.

Materials—Sixty-eight action verbs were selected. For each verb, two to eight instruments (*M* = 4) that can be used to perform each action (e.g., eat with a fork or spoon) were generated based on the experimenters' intuitions. In addition, the 29 instrument-verb pairs from Rayner et al.'s (2004) 30 plausible items were included. There were only 29 because Rayner et al. used *knife-cut* for two of their plausible items (*knife-cut-carrots* and *knife-cut-bread*). From the resulting 220 instrument-verb pairs, four lists were created so that no instrument-verb pair appeared twice on any list (two lists contained 56 items and 2 contained 54). If the same verb or semantically similar verbs (e.g., *cut* and *chop*) appeared in a list, they did not occur on the same page of the norming form. No list contained more than two occurrences of each verb. The task was administered in a pen-and-paper fashion, with nine instrument-verb combinations per page (with two lists containing the two additional items on the final page). An additional list containing 10 instrument-verb pairs from Warren and McConnell's (2007) 30 plausible items were included. The remaining 20 plausible items were part of the first set of 220 items. For this list, the task was administered in a web-based form, with each item presented in random order.

Procedure—Participants were given instrument-verb pairs, and were asked to "List the things or people that have the following actions done to them with the specified instruments." Each item was worded such as "wash using shampoo". Space for five responses was provided for each item, and participants were instructed to write down or type in as many types of things or people they could think of in one minute for each item. The task took less than one hour to complete.

Results and Discussion

Responses were scored on the basis of their production order within a participant and on their production frequency across participants. A weighted score was calculated for each response by multiplying the frequency with which it was produced as the first response by 5, second by 4, and so on, and then summing those products.

Based on the weighted scores, we selected 48 typical instrument-verb-patient triplets (24 unique verbs) using the following criteria. First, whenever possible, the chosen typical patient was the response with the highest weighted score for an instrument-verb pair. Second, we chose typical items that allowed re-pairing of instruments between the two triplets that used the same verb (e.g., *shampoo-wash-car* and *hose-wash-hair* from *hose-wash-car* and *shampoo-wash-hair*) so that they were atypical but not anomalous (see Norming Study 2). The atypical patient nouns had a low or zero weighted score for its instrument-verb pair. These criteria ensured that each triplet contained a typical and an atypical patient for the instrument-verb pair.

The typical patient nouns had weighted scores ranging from 13 to 101, with a mean of 59 (*SD* = 24) and a median of 53. The mean production probability was .64 (i.e., 64% of participants listed the patient in one of the five slots; *SD* = 23%). Among those items, two triplets (i.e., *net-catch-fish* and *hammer-hit-nail*) seemed to possess an overly strong relation between the instrument and patient. Indeed, these two instrument-patient pairs are highly associated (.39 and .28 respectively) according to Nelson, McEvoy, and Schreiber's (1998) word association norms. To control for normative word association as a confounding factor, we replaced the patient nouns for those items with an exemplar (*trout* for *fish*) and a near synonym (*spike* for *nail*) to obscure the normative association. After these items were altered, the mean weighted score was 55 (SD = 28), the median was 51, and the mean production probability was .61 (SD = .26). The atypical patient nouns created by re-pairing the typical instrument-verb combinations had weighted scores ranging from 0 to 11, with a mean of 1 (SD = 3), a median of 0, and a mean production probability of .03.

In contrast, the weighted scores for Rayner et al.'s (2004) 30 plausible items were much lower than for our typical patient nouns, ranging from 0 to 80, with a mean of 21 (SD = 23) and a median of 14. The mean production probability was .24. Eight of the 30 items had a weighted score of 0, and 6 items had a score between 1 and 5. Warren and McConnell's (2007) 30 plausible items had weighted scores ranging from 0 to 96, with a mean of 22 (SD = 8) and a median of 9. The mean production probability was .26. Eight of the 30 items had a weighted score of 0, and 5 items had a score between 1 and 5. That is, in both studies, approximately half of their items were not what participants considered to be typical patients in the type of event denoted by the instrument-action combination. This suggests that their lack of significant early plausibility effects may have resulted because their plausible items did not sufficiently match people's event knowledge.

An analysis of variance was conducted on the plausibility ratings using study as a between-items independent variable. The studies differed in terms of weighted scores for the plausible items, $F_2(2, 105) = 24.09$, $p < .001$. Bonferroni adjusted t-tests showed that the mean weighted score of our typical items was significantly greater than Rayner et al.'s, $t_2(108) = 6.06$, $p < .001$, and Warren and McConnell's plausible items, $t_2(108) = 5.46$, p < . 001. The weighted scores for Rayner et al.'s and Warren and McConnell's plausible items did not differ significantly, $t_2(108) = 0.20$, $p > .8$. Given that our typical items better match the concepts that people generate given an instrument-action pair, early effects at the patient might be obtained with sentences using our materials.

Norming Study 2

The primary purpose was to establish that our atypical items were not anomalous. For this purpose, it sufficed to include only Rayner et al.'s (2004) items as a comparison set. If participants rate our atypical items as anomalous, then our experiments would be replicating Rayner et al.'s and Warren and McConnell's (2007) comparisons between plausible and anomalous items, rather than comparing more and less plausible sentences (although there would not be selectional restriction violations in our items). Norming Study 2 differed slightly from Rayner et al.'s plausibility ratings in that the sentences ended at the patient so that post-patient continuations could not influence plausibility ratings, and that a 7-point scale, instead of 5-point scale, was used. Note that Warren and McConnell used sentences truncated at the patient, and used a 7-point scale.

Method

Participants—Twenty-six undergraduates at the University of Western Ontario participated for course credit.

Materials—Using the 48 typical and atypical instrument-verb-patient sets selected from Experiment 1, we constructed 96 sentences using Rayner et al.'s (2004) template. Two examples per condition are shown below.

Typical:

a. Donna used the hose to wash her filthy car.

b. Donna used the shampoo to wash her filthy hair.

Atypical:

a. Donna used the shampoo to wash her filthy car.

b. Donna used the hose to wash her filthy hair.

The proper name, the verb *used*, the determiners (or possessive pronouns) for the instrument noun and for the patient noun phrase, and the adjective preceding the patient noun were identical among the sentences that featured the same verb. Adjectives modifying the patient nouns were included because Rayner et al.'s and Warren and McConnell's (2007) sentences included them, and were selected so that they did not bias the plausibility of the patient nouns toward either instrument-verb combination. The items are presented in the Appendix, although they are presented there in their full form, that is, continuing past the patient, and including a second sentence.

Rayner et al.'s (2004) 90 items (30 plausible, implausible, and anomalous sentences) were also included as a direct comparison. Two lists were created to minimize the number of items with lexical overlap. For example, for each of our four items with the same verb, the two atypical items were put in one list and the two typical items in the other list so that there was no overlap with respect to instruments and patients (while balancing the lists overall for number of typical and atypical sentences). Rayner et al.'s items were divided so that sentences with the same verb were in separate lists (the plausible and implausible conditions used the same verb, whereas the anomalous items used a different verb). Each list contained half of the items from each condition.

Procedure—Participants were randomly assigned to one of the lists and were asked to rate how likely it is that the event described in each sentence occurs in the real world, on a scale of 1 (very unlikely) to 7 (very likely). Sentences were presented one at a time in random order on the computer screen, along with the rating scale at the bottom of the screen. Each session lasted approximately ten minutes.

Results and Discussion

As presented in Figure 1, the sentence types, in order from least to most plausible were Rayner et al. (2004) anomalous, Rayner et al. implausible, our atypical, Rayner et al. plausible, and our typical sentences. An analysis of variance was conducted on the plausibility ratings using sentence type as a between-items variable. There were differences among sentence types, $F_2(4, 181) = 196.60$, $p < .001$. Bonferroni adjusted *t*-tests showed that our atypical items were rated as more plausible than Rayner et al.'s anomalous items, $t_2(76) = 6.43$, $p < .001$, and also their implausible items, $t_2(76) = 3.00$, $p < .004$. Our typical and atypical items differed on plausibility, $t_2(94) = 14.96$, $p < .001$, as did Rayner et al.'s plausible and implausible items, $t_2(58) = 15.87$, $p < .001$. Finally, our typical items were rated as more plausible than Rayner et al.'s plausible items, $t_2(76) = 2.10$, $p < .04$.

The primary purpose of Norming Study 2 was to establish that our atypical items are not anomalous. With a mean rating of 3.2, they were rated not only as more plausible than Rayner et al.'s (2004) anomalous items, but also as significantly more plausible than their implausible items. Thus our atypical items were simply that, atypical. Note also that none of our atypical items violated selectional restrictions.

Two other results are worth noting, although we simply point them out here. First, the difference between our typical and Rayner et al.'s plausible items was only 0.3, the smallest difference between sentence types found in the plausibility ratings. This is somewhat odd, given that there was a sizeable difference between these two conditions in the production norms of Norming Study 1. The second notable result is that the difference between Rayner et al.'s plausible and implausible items (6.0 - 2.4 = 3.6) is actually larger than that between our typical and atypical items (6.3 - 3.2 = 3.1). Note that in Warren and McConnell's (2007) plausibility ratings, the difference between their plausible and implausible items was 4.2 on a 7-point scale. We return to both points in the General Discussion.

Finally, we note that it apparently did not make a substantial difference to use sentences that ended at the patient, as opposed to the full sentences (as in Rayner et al., 2004). The correlation between the ratings from Norming Study 2 and Rayner et al.'s ratings using all 90 of their items was $r = .87$. $p < .0001$.

Experiment 1

The production norms of Norming Study 1 suggest that our typical items better match people's event knowledge than do the plausible items of Rayner et al. (2004) and Warren and McConnell (2007). Therefore, the purpose of Experiment 1 was to use self-paced reading to investigate whether our items would produce reading-time differences directly at the patient.

Method

Participants—Thirty-six University of Western Ontario students participated for compensation ($10).

Materials—The 96 sentences were the same as those used in Norming Study 2, except that they continued after the patient. An example of the sentences using the verb *wash* is shown below.

Typical:

a. Donna used the hose to wash her filthy car after she came back from the beach.

b. Donna used the shampoo to wash her filthy hair after she came back from the beach.

Atypical:

a. Donna used the shampoo to wash her filthy car after she came back from the beach.

b. Donna used the hose to wash her filthy hair after she came back from the beach.

The target sentences followed Rayner et al.'s template and were identical for at least two words following the critical patient noun. All target sentences were followed by a second sentence to increase the meaningfulness of the narrative. These second sentences were written to make all patient nouns in both conditions make sense with respect to the discourse. The items are presented in the Appendix.

The sentences were divided into four lists so that no participant saw any instrument, verb, or patient more than once. Each list contained 24 target sentences. Eighty-four filler sentence pairs with various constructions were added to each list to ensure that the targets never occurred adjacently nor appeared first in a list. An additional 16 sentence pairs were used for practice. Yes/no comprehension questions regarding the content of the sentences were created for each sentence pair. The proportion of “yes” and “no” responses were matched across lists and items.

Procedure—The sentence pairs were presented on a 17-inch color CRT monitor using a one-word-at-a-time moving-window self-paced reading paradigm (Just, Carpenter, & Wooley, 1982), implemented in E-Prime (Psychology Software Tools, Inc.) on an AMD Athlon 64 Processor 3200+ computer. Texts were presented in 18 point Courier New font. Each trial began with lines of dashes appearing on the screen, with each non-space character of the sentence pair replaced by a dash. Participants pressed the space bar on the computer keyboard to read each word, with each key press revealing the next word and reverting the previous word to dashes. After the second sentence was completed, participants answered a comprehension question by pressing the “F” button for yes and “J” for no. Reading latencies for each word and responses to the comprehension questions were recorded. Each session began with 16 practice items. Participants then read 108 experimental items, taking a break every 20 items. Each session lasted approximately half an hour.

Results & Discussion

All participants scored better than 80% correct on the comprehension questions (M = 92%, SD = 3%). In addition to the patient noun (*hair*), we analyzed each of seven word positions starting at the four words prior to the patient noun (*to wash her filthy*) and continuing to two words following it (*after she*). The effect of typicality in each region was examined in by-participants (F_1) and by-items (F_2) analyses of variance. Table 1 presents mean reading latencies and the associated F-statistics.

There were no significant differences between typical and atypical conditions in any of the four word positions prior to the patient noun. This was expected because, across conditions, the sentences were identical to this point. No difference was greater than 10 ms, and there was only a 1 ms difference at the adjective preceding the patient. A main effect of event typicality was found at the patient. Patient nouns were read 29 ms faster when the concepts they denoted were typical of the event described by the instrument-verb combinations than when they were atypical. In the post-patient regions, there was a significant 24 ms spill-over effect on the second word after the patient noun (*she*), although only a 6 ms difference at the word directly following the patient.

Experiment 1 used sentential stimuli created from the event generation norms, which reflect people's productions based on their event knowledge. This manipulation influenced reading latency at the critical point in time, that is, directly at the patient noun. Patients that were typical of the event denoted by instruments and verbs matched the context, and thus were processed faster than those that are atypical. These results cannot be explained by a strictly verb-based account (e.g., selectional restrictions) because the same verbs were paired with the same two patients in both conditions. That is, *wash-hair* and *wash-car* appeared in both conditions. What differed were the instruments that were paired with those verb-patient combinations. The results strongly suggest that experiential event knowledge is the source of the observed effects.

We note, however, that there is a possible alternative explanation for these results. The difference between the typical and atypical items may reflect some form of lexical priming directly from instruments to patients. In particular, there clearly do exist semantic relations between instruments and patients. People do know that *hose* is related to *car*, and they can tell you why (i.e., because you use a hose to wash your car). The goal of our studies, however, is to investigate whether an instrument can combine with a verb to influence processing of a patient, not whether an instrument influences processing of a patient independently of the specific intervening verb. Thus, although the influence of this sort of priming effect in sentence comprehension is often rejected on the argument that lexical association effects diminish across intervening words (Murray, 2006; Rayner et al., 2004), it seems prudent to rule it out empirically for our specific instrument-patient pairs.

Experiment 2

To verify that the self-paced reading results are not due to semantic relations or lexical associations between our instruments and patients, we conducted a 250 ms SOA priming task in which instruments (*hose* vs. *shampoo*) were used to prime patients (*car*). Given that the mean reading latency between instruments and patients in the Experiment 1 sentences corresponds to an inter-stimulus interval of 1165 ms, it may on the surface seem appropriate to use a longer SOA. However, word-word priming differs from a situation in which there are multiple intervening words, as there were in Experiment 1. Rather than being engaged in understanding the sentence as the intervening words unfold, in a long SOA word-word priming study in which nothing intervenes between prime and target, participants are free to use strategies such as explicitly trying to guess the identity of the upcoming target given the prime. Therefore, it is not appropriate to use an SOA that mimics the temporal parameters of the reading time experiment. If the instruments do indeed directly prime the patients used in the sentences of Experiment 1, a short SOA such as 250 ms should be sufficient to show it.

Method

Participants—Forty students from the University of Western Ontario participated. They were native English speakers and had normal or corrected-to-normal visual acuity.

Materials—We used the instruments and patients from Experiment 1 as primes and targets. Priming effects are determined by comparing decision latencies for related prime-target pairs to those for unrelated pairs. In this experiment, the related items were the instrument-patient pairs from the typical sentences (*shampoo-hair* and *hose-car*), whereas the corresponding unrelated items were taken from the atypical sentences (*hose-hair* and *shampoo-car*). Therefore, there were 48 related and 48 unrelated pairs.

Items were divided into two lists so that no participant saw any word twice. Each list contained 24 related and 24 unrelated instrument-patient pairs. As a manipulation check, we also included 32 prime-target pairs taken from McRae and Boisvert (1998), in which priming was obtained. Thus, each list also included 16 semantically similar concrete noun pairs (*goose–turkey*) and 16 semantically dissimilar pairs (*eagle–catapult*) from McRae and Boisvert. Both lists also included 80 unrelated word-word pairs (*scarf–elevator*) and 160 word-nonword pairs (*skate-dird*). Thus, there were 50% word targets and 50% nonword targets, and the relatedness proportion was .25. Ten unrelated word-word pairs and ten word-nonword pairs were presented first and thus served as the practice trials.

Procedure—Each trial started with a fixation point (+) displayed for 250 ms, followed by the prime for 200 ms, a blank screen for 50 ms, then the target word until the participant responded. The inter-trial interval was 1500 ms. Participants were instructed to read the first letter string and then to make a lexical decision to the second letter string by pressing a button on an E-Prime button box. The button for the “word” decision always corresponded to the participant's dominant hand.

Results and Discussion

The McRae and Boisvert (1998) items and the instrument-patient pairs were analyzed separately. Trials on which an error occurred were excluded from decision latency analyses. Lexical decision latencies greater than three standard deviations above the mean across all experimental trials were replaced by that cutoff value (< 2% of the scores). Relatedness (related vs. unrelated) was within participants (F_1) and items (F_2).

McRae and Boisvert (1998) Pairs—Decision latencies for semantically similar pairs (*M* = 607 ms; *SE* = 12 ms) were 17 ms shorter than for dissimilar pairs (*M* = 624 ms; *SE* = 11 ms), $F_1(1, 39) = 6.73$, $p < .02$, $F_2(1, 31) = 3.81$, $p = .06$. Thus, the McRae and Boisvert items elicited a priming effect, showing that there was nothing odd in the procedure or participant sample in terms of the ability to observe priming.

Instrument-Patient Pairs—The decision latencies for typical instrument-patient pairs (*M* = 560 ms; *SE* = 12 ms) were a marginal 15 ms shorter than for atypical pairs (*M* = 575 ms; *SE* = 12 ms), $F_1(1, 39) = 3.13$, $p < .09$, $F_2(1, 47) = 2.77$, $p > .1$. The 15 ms difference was driven by three items that had notably large priming effects (*cauldron– potion* vs. *kettle – potion*, *band-aid – scrape* vs. *white-out – scrape*, and *payment – interest* vs. *shortcut – interest*). We removed those items along with the ones that were symmetrically paired with the same instruments to retain proper counterbalancing. Without those items, there was a 1 ms priming effect (typical: *M* = 563 ms; *SE* = 12 ms; atypical: *M* = 564 ms; *SE* = 12 ms), $F_1 < 1$, $F_2 < 1$.

Reanalysis of self-paced reading results—We reanalyzed the Experiment 1 self-paced reading data after removing the items that were responsible for the marginal instrument-patient priming effect. The original Experiment 1 data and the reanalyzed data are presented for comparison in Figure 2. The re-analysis revealed the same pattern of results. There was a 34 ms main effect of event typicality at the critical patient noun, $F_1(1, 35) = 7.84$, $p < .008$, $F_2(1, 41) = 5.66$, $p < .03$, and a slightly diminished and now marginal 22 ms spill-over effect two words after the patient noun, $F_1(1, 35) = 3.60$, $p < .07$, $F_2(1, 41) = 3.43$, $p < .08$. No significant difference was found at any other word position. Thus, the self-paced reading results were not driven by direct priming from our instruments to patients. Crucially, it is the combination of the instrument and verb that alters the class of events referred to by the verb, and thus alters the match between the instrument-verb pair and the ensuing patient.

Finally, although we found only marginal priming from instruments to patients that was due to three items, it should be noted that Hare, Jones, Thomson, Kelly, and McRae (2009b) did in fact find robust priming from instruments to patients (things that those instruments are used to act upon) using the same task parameters. It is important to note that only one item overlapped between the two studies, reflecting the researchers' contrasting goals. Hare et al.'s goal was to demonstrate event-based priming directly from instruments to patients, and therefore they purposely chose items sharing relations that were as strong as possible. Our goal when constructing items was to attempt to avoid direct priming of this sort, and instead to demonstrate the influence of the instrument combined with a specific action.

Experiment 3

The studies reported so far support our hypothesis regarding the rapid use of event-based knowledge, and thus contrast with Rayner et al. (2004) and Warren and McConnell (2007). However, the difference in experimental paradigms, namely self-paced reading versus eyetracking, complicates drawing direct comparisons. Furthermore, several researchers have raised issues with self-paced reading in particular. For example, Rayner et al. (2004) commented on Thornton and MacDonald's (2003) demonstration of immediate plausibility effects in self-paced reading study by stating that “self-paced reading is known to slow down the normal reading process (Rayner, 1998), so their results are not definitive with respect to the issue at hand” (p. 1290). Therefore, in Experiment 3, we replicated

Experiment 1 using eyetracking to enable a more direct comparison between our results and those of Rayner et al. and Warren and McConnell.

Method

Participants—Thirty-two native English-speaking undergraduates from the University of Glasgow participated in exchange for £5.

Materials—Stimuli were identical to those of Experiment 1, except that 24 of the filler items were replaced with sentences from a second, unrelated experiment on relative clause processing. In addition, 4 of the 48 experimental items were ‘de-Canadianized’ in non-critical portions of text (i.e., in the second sentences). One-third of the experimental items and about half of the filler items were followed by comprehension questions. Experimental and filler items spanned 2-3 lines on the screen. For all experimental items, the first line contained the first sentence up to (and including) at least three words after the critical patient noun.

Procedure—The experiment was conducted using a SR-Research EyeLink 2000 desk-mounted eye-tracker that has a spatial resolution of 0.01 degrees. The tracker ran at a 1000 Hz sampling rate. Stimulus presentation and data collection were controlled by software developed at University of Massachusetts, Amherst (EyeTrack Version 0.7.9) on the basis of the EyeLink API. Participants were seated about 70 cm from a 21 inch CRT display running at 100 Hz refresh rate in 1280 × 1024 pixel resolution; 2.6 characters equaled one degree of visual angle. Sentences were presented in 20 pt bold Courier New font printed in black on a light grey background (RGB 232,232,232). Line spacing was set to twice the font height such that fixation locations could unambiguously be mapped onto a corresponding line of text. Viewing was binocular, but only the participant's dominant eye was tracked (the right eye for approximately 70% of the participants, as determined by a simple parallax test prior to the experiment). A chin rest was used to keep the viewing distance constant and to prevent strong head movements during reading. Button responses were collected using a hand-held Microsoft USB game pad.

Each participant was assigned to one of four lists containing different versions of experimental items randomly interspersed with the fillers. Randomization was constrained such that there were always six filler trials at the beginning, and each critical trial was preceded by at least one filler. At the start of the experiment, the experimenter performed the standard EyeLink calibration procedure, which involved participants looking at a grid of nine fixation targets in random succession. Then a validation phase followed to test the accuracy of the calibration against the same targets. Calibration and validation were repeated at least once every 25 trials, or if the experimenter noticed a decline in measurement accuracy (e.g. after a change in the participant's posture). Each block of trials following a calibration began with at least one filler.

Each trial started with the presentation of a central fixation dot for drift-correction, followed by a small rectangle in the same position (five characters from the left in the middle row of the screen) as the first character of the upcoming text display; a fixation for at least 200 ms on this target triggered the presentation of the text display, thereby ensuring that participants always started reading in the leftmost character position. Participants were instructed to read at a normal pace (“as if reading a newspaper”) and to press the right-hand button on the game pad when they had finished reading a text display. In 55% of the trials, this triggered the presentation of the following text display; in the remaining trials, a simple yes/no comprehension question was presented which the participant had to answer using either the left (‘no’) or the right (‘yes’) response button. Answering the question trigged the presentation of the next trial in those cases. An experimental session lasted about 40 minutes.

Data Analyses—Fixation coordinates were mapped onto character positions using software developed at University of Massachusetts, Amherst (EyeDoctor Version 0.6.3). Fixations shorter than 80 ms (less than 4% of all fixations) were pooled with preceding or following fixations if these fixations were within one character space of those short fixations. Fixations shorter than 40 ms (less than 1%) were excluded if they were within three character spaces of adjacent fixations, and fixations of more than 900 ms were excluded altogether (cf. Rayner et al., 2004). Also excluded were trials with poor vertical accuracy or with blinks occurring while reading critical portions of text for the first time.

Data are reported for four scoring regions: [R1] the infinitival verb region (*to wash*), [R2] the determiner-plus-adjective region preceding the critical patient noun (*her filthy*), [R3] the critical patient noun (*hair*), [R4] the two (usually short) words immediately following the patient noun (*after she*), and finally [R5] the third word following the patient noun (*came*), with R5 being located at the end of the first line of text in most cases. If a fixation landed on the space between two adjacent regions, it was counted as being to the right of the space.

For each region, fixation data were summarized in terms of five commonly reported eye-tracking measures: (1) *first fixation duration* (the duration of the first fixation on a region); (2) *first pass reading time* (the time from fixating a region for the first time until a different region, either to the left or to the right, is fixated – also called *gaze duration* when applied to single-word regions); (3) *regression path duration* (the time from fixating a region for the first time until a word to the right of the region is fixated – this measure also includes fixations following a regression from the region of interest after encountering it for the first time; cf. *go-past* measure in Rayner et al., 2004); (4) *total reading time* (the sum of all fixations on a region, including those for re-reading the region); and finally (5) percentage of trials on which the patient noun was skipped. Following common procedures, the first three measures (first fixation duration, first pass reading time, and regression path duration) were treated as having missing data if the region of interest was skipped during first pass reading. In total, less than 10% of the available data were excluded due to poor calibration, blinks, or skipping of regions.

Results and Discussion

All participants scored at least 75% accuracy on comprehension questions following the critical trials (mean = 89%; SD = 4%). The results by measure and region are summarized in Tables 2 to 6. Immediate effects of event typicality showed up in first fixation duration and gaze duration at the critical patient noun, in line with the findings from self-paced reading (Experiment 1), but contrasting with Rayner et al. (2004) and Warren and McConnell (2007). Moreover, there was an effect in regression path duration at R5 (second region following the patient noun) which resembles the spillover typicality effect observed in Experiment 1. Given that this regression path duration effect is not reflected in other first-pass measures at R5, it appears that readers tend to launch regressive eye-movements from this region before moving on to read subsequent portions of text, particularly after having encountered an atypical patient noun. There were also reliable typicality effects in total reading time for R1 (infinitival verb), R2 (determiner plus adjective) and R3 (patient noun). Given that these were much larger in magnitude than corresponding first-pass effects in those regions, it can be concluded that relevant regions were frequently re-inspected in the atypical patient condition. Finally, due to the fact that a number of the target patient nouns were short, 16.9% of the typical patients and 12.3% of the atypical patients were skipped in first pass reading. This difference was marginal by participants and significant by items.[1]

As with the self-paced reading data from Experiment 1, we also reanalyzed the eye-tracking data after removing the three items that were responsible for a marginal instrument-patient priming effect in Experiment 2 along with their counterbalanced items. As in Experiment 1, this supplementary analysis revealed the same pattern of results as with the full set of items. Most notably, there now was a slightly stronger (14 ms) first fixation duration effect of event typicality at the patient, $F_1(1, 31) = 6.74$, $p < .02$, $F_2(1, 41) = 8.72$, $p < .01$, a 23 ms gaze duration effect at the patient, $F_1(1, 31) = 8.47$; $p < .01$, $F_2(1, 41) = 10.81$, $p < .01$, and a 46 ms regression path duration effect at R5 (the second region following the critical patient noun), $F_1(1, 31) = 4.31$, $p < .05$, $F_2(1, 41) = 5.98$, $p < .02$. These analyses again confirm that direct instrument-patient priming is not responsible for the current results.

In conclusion, event typicality influences the earliest stages of processing the critical patient noun, which contrasts with the finding of a delayed plausibility effect in Rayner et al. (2004) and Warren and McConnell (2007).

General Discussion

The purpose of this research was to test whether event knowledge that is computed by combining multiple concepts is an integral component of on-line language comprehension, and whether it immediately influences language comprehension. This position contrasts with the claims of researchers such as Bornkessel and Schlesewsky (2006), Clifton and Staub (2008), Frazier (1995), and van Gompel et al. (2005) who have argued that only purely lexical semantic information such as selectional restrictions immediately influences on-line language comprehension.

We focused particularly on investigating whether conceptually combining instruments and actions can have an immediate influence on comprehending patients. The studies of Rayner et al. (2004) and Warren and McConnell (2007) suggest that it does not. In Norming Study 1, production norms were used to construct materials that reflect people's offline event-based knowledge. We also used those data to assess concerns about Rayner et al.'s and Warren and McConnell's items, showing that many of their plausible items may not have optimally matched people's event knowledge. On the other hand, plausibility ratings on sentences up to the patient showed a definite difference between their plausible and implausible items. This inconsistency in event-based productions versus sentence plausibility ratings is discussed below.

In Experiment 1, event-based typicality influenced self-paced reading times at the critical patient noun, as predicted by an account in which people's knowledge of common events and situations influences either rapid conceptual integration or expectancy generation. Participants read a patient noun (*hair*) more quickly when it was typical of the event described by the instrument and action (an event in which shampoo is used for washing hair) compared to when it was atypical (when a hose is used for washing hair). In Experiment 3, an eyetracking study, typicality influenced both first fixation and gaze durations at the patient. These results cannot be due to selectional restrictions because the restrictions imposed by the verb were equivalent in both conditions (the items were completely crossed), and no selectional restrictions were violated. In addition, re-analyses based on the instrument-patient priming results of Experiment 2 show that the results of Experiments 1 and 3 were not due to direct relations between the instruments and patients. Instead, they

[1] The skipping percentage analyses were based on Generalized Estimating Equations (GEE; Hardin & Hilbe, 2003). Unlike standard ANOVA, this procedure allows for the specification of distribution and link functions that are appropriate for categorical data analysis. Here, a binomial distribution and logit link function were assumed (cf. Jaeger, 2008). Typicality was entered as a within-subjects and within-items predictor, assuming a compound symmetry covariance structure for repeated measurements.

support the hypothesis that comprehenders rapidly use event knowledge computed from intra-sentential context, in this case, the combination of instruments and actions.

Plausibility, Predictability, and Reconciling the Results

Although the production norm results of Norming Study 1 are consistent with the difference between our experiments and those of Rayner et al. (2004) and Warren and McConnell (2007), the plausibility ratings of Norming Study 2 are not. When participants were asked to produce the types of patients that are, for example, “cut with a knife”, there were substantial differences between our typical items and Rayner et al.'s and Warren and McConnell's plausible items. However, when participants rated the plausibility of sentences that ended at the patient, the difference between Rayner et al.'s plausible and implausible items (6.0 - 2.4 = 3.6) was actually numerically larger than between our typical and atypical items (6.3 - 3.2 = 3.1). Hence, although participants produced our typical items much more often than Rayner et al.'s, they judged the plausibility manipulation to be stronger in Rayner et al.'s items than in our own (due to differences in the implausible/atypical items). Furthermore, there was only a 0.3 difference in rated plausibility for our typical and Rayner et al.'s plausible items. This raises the issue of why Rayner et al., and Warren and McConnell, who found a plausibility rating difference of 4.2 in their norms, did not find clear immediate differences in reading times between their plausible and implausible sentences, whereas we did.

Before considering explanations of the differences among the studies, it is important to note that the eyetracking results are not tremendously different. In Experiment 3, the plausibility effects at the patient were a significant 11 ms for first fixation duration, a significant 21 ms for gaze duration, and a nonsignificant 12 ms for go-past reading times. In Rayner et al. (2004), the plausibility effects at the patient were −1 ms for first fixation duration, 3 ms for gaze duration, and a marginally significant 13 ms for go-past reading times. In Warren and McConnell (2007), the plausibility effects at the patient were 1 ms for first fixation duration, a nonsignificant 16 ms for gaze duration, and a significant 33 ms for go-past reading times. In other words, Experiment 3 showed differences in the time taken to read the patient noun, whereas the differences in Warren and McConnell, and Rayner et al. as well (marginal effects), were due to regressive saccades that were triggered differentially by the plausible versus implausible patients. Thus, numerically and statistically, the differences among the three studies are not huge, but their theoretical implications are. In numerous previous eye-tracking studies, including Rayner et al. and Warren and McConnell, the absence of statistically significant effects in early measures, which are widely regarded as first fixation and gaze durations, has been interpreted as supporting the conclusion that a psychological construct of interest (in this case, plausibility) has a delayed influence. This delayed influence has been widely interpreted as potentially (though not necessarily) reflecting the architecture of the underlying sentence processing mechanisms (i.e., that some process is delayed because it is further downstream in a sequence of computations). It is therefore critical that we reconcile our results with those of Rayner et al. and Warren and McConnell.

There were a few methodological differences among the experiments. First, we fully crossed conditions in that all instruments, verbs, and patients were used in both the typical (*shampoo-wash-hair, hose-wash-car*) and atypical conditions (*hose-wash-hair, shampoo-wash-car*). In contrast, Rayner et al. (2004) and Warren and McConnell (2007) used the same verbs and patients in their plausible and implausible conditions, which is an important aspect of the design of this type of experiment, but not the same instruments (*knife-chop-carrots* vs. *axe-chop-carrots*). One consequence of this difference is that our items better resembled minimal pairs. Perhaps just as importantly, there were 48 items in each condition in our experiments, as compared to 30 in Rayner et al. and Warren and McConnell. These two differences may have provided us with additional power and sensitivity that was sufficient to observe earlier significant effects.

The second potential explanation concerns the differential sensitivity of norming methods. The discrepancy among reading time measures, production norms, and plausibility ratings perhaps demonstrates that the relationship between comprehension effects and event-based plausibility is nonlinear, as indeed Warren and McConnell (2007) suggested. One potential way to think about plausibility is to consider it as a dimension that ranges from highly anomalous to extremely predictable. On this continuum, plausibility ratings nicely capture differences at the lower end: the ratings show substantial differences between anomalous and implausible items, and, consistent with this, both Rayner et al. (2004) in gaze durations, and Warren and McConnell in first fixation durations, found immediate differences between these conditions.

Event-based production norms, on the other hand, are not sensitive to the lower end of the plausibility continuum. Participants in such studies rarely, if ever, produce anomalous, implausible, or atypical items, and thus production norms do not differentiate among those conditions because all items have a value of essentially zero. Such norms do, however, sensitively capture differences in the upper part of the continuum where plausibility ratings appear to do so only weakly. In plausibility ratings, when the situations being described are quite reasonable, as in our typical and Rayner et al.'s plausible items, there is little variation in the ratings. For example, these conditions were both rated as highly plausible, and differed only by 0.3 on a seven-point scale in Norming Study 2. Such decreased sensitivity in the upper end of the plausibility continuum might even be enhanced when a number of sentences to be rated are anomalous, thus decreasing perceived differences among plausible sentences. In contrast, when people are asked to produce responses, differences at the upper end of the continuum become quite apparent, as they did in Norming Study 1.

Thus, although the plausibility ratings may predict similar reading time effects for the three studies, or even larger effects for Rayner et al. (2004) and Warren and McConnell (2007) than for the present experiments, the event-based production norms mirror the observed patterns of results. Two examples illustrate this difference between norming methods. Both The man used a shovel to spread the steaming asphalt (Rayner et al. plausible item) and *Jamie used a lantern to light the cheap room* (our typical item) had a mean plausibility rating of 5.77. However, in Norming Study 1, no participant produced either *asphalt* or *pavement* (weighted score = 0), whereas *room* had a weighted score of 34. Higher up the plausibility continuum, both Stuart used a ruler to measure the various dimensions (Rayner et al. plausible) and *Leslie used the white-out to cover the minor error* (our typical) had a mean plausibility rating of 6.46. In contrast, the weighted score was 4 for *dimensions* (one participant produced it second), but 81 for *error* (with *error* and *mistake* being combined).

A related issue is that there may be two similar, but perhaps psychologically separate dimensions of plausibility: what one might refer to as *plausibility* and *implausibility*. The focus of the present study somewhat contrasts in this regard with that of Rayner et al. (2004) and Warren and McConnell (2007). Theoretically and empirically, we focused more on a match between people's knowledge of common events and the linguistic input. Therefore, we used highly plausible items in our typical condition, and considered the atypical items as the control condition. That is, our goal was to facilitate reading times by having the typical items match people's event knowledge as well as possible. In contrast, both Rayner et al. and Warren and McConnell discussed their studies in terms of “violation detection”, or detecting a semantic violation, as Warren (in press) put it. Due to this focus, they manipulated the degree of implausibility of their implausible and anomalous items, and put less emphasis on the degree of plausibility of their plausible items. This approach is exemplified by the fact that Rayner et al. called their plausible items the “control” condition. Note that we are not saying that one approach is right whereas the other is wrong; we are simply stating that the approaches differ with respect to focusing on plausibility versus implausibility.

In summary, the typical items in Experiments 1 and 3 were created to match people's event knowledge, and the production norms show that they did. On the other hand, Rayner et al.'s (2004) and Warren and McConnell's (2007) plausible items were created to be significantly more plausible than their implausible (or anomolous) items, and both their and our plausibility norms show that they were. These facts give rise to another potential issue, namely that it is possible that our experiments actually addressed predictability rather than plausibility.

Cloze norms are commonly used to measure the predictability of continuation given a sentence fragment, and the production norms might be viewed as an alternative form of cloze norms (albeit outside of specific sentences). Hence, the differences among studies may actually reflect differences in predictability rather than plausibility, implausibility, or event typicality. That is, the early effects that we obtained may be due to our typical items being more predictable than the plausible items of Rayer et al. (2004) and Warren and McConnell (2007). To test this, we collected cloze norms from 60 University of Western Ontario undergraduate students who did not participate in any of the other studies. We included our items, as well as those of Rayner et al. and Warren and McConnell, who did not report cloze values for their stimuli. We truncated each sentence after the pre-target adjective, divided the unique sets of fragments into three presentation lists to minimize lexical overlap, and asked participants to produce a single-word continuation for each fragment. We obtained average cloze values of 20.8% (range = 0 to 75%) for our typical items, and 1.0% (range = 0 to 10%) for our atypical items. Rayner et al.'s plausible items had lower cloze values (*M* = 10.3%; range = 0 to 55%), with their implausible items being similar (*M* = 1.5%; range = 0 to 15%). The same is true of Warren and McConnell's plausible (*M* = 14.3%; range = 0 % to 90%) and implausible items (*M* = 0.5%; range = 0 to 15%). That is, our items were higher in terms of their mean predictability as measured by cloze norms.

Given these cloze statistics, predictability may be responsible for the discrepancy among studies, and thus it could be argued that the present experiments investigated predictability whereas Rayner et al. and Warren and McConnell investigated plausibility. However, a major and somewhat perplexing issue concerns how and where to draw a line that differentiates predictability from plausibility. The common definition of predictability is the conditional probability that a specific word will occur given the preceding word, sentential fragment, or discourse. This is usually gauged using cloze norms. Experiments investigating predictability effects examine the correspondence between the cloze probability of a given word and reading measures on that word (or the magnitude of N400, as in DeLong et al., 2005, and Van Berkum et al., 2005). In contrast, plausibility, in its most general form, can be defined as the acceptability or likehood of a situation or a sentence describing it, as a whole. Plausibility usually is measured by asking participants to rate, on a Likert scale, “How likely it is that the described event occurs in the real world?” Unlike predictability, plausibility is not inherently conditional or directional, and plausibility ratings do not necessarily reflect the influence of a specific target word or region.

Given these definitions, plausibility and predictability can be contrasted relatively easily. For example, changing a sentence's post-target continuation could alter plausibility ratings, but not predictability of the target because post-target material does not influence it. Critically however, in many experiments that investigate plausibility effects in reading, including the ones under consideration here, the effect is measured at a specific word or region. In such cases, researchers are investigating what may be referred to as conditional plausibility (or implausibility), which could be defined as the acceptability or likelhood of a target word or concept given a preceding word, sentential fragment, or discourse. In this case, predictability and conditional plausibility are extremely difficult to disentangle. This does not suggest that predictability and (conditional) plausibility are the same, but it is does bring attention to the fact that drawing a line between the two is certainly not straightforward.

One way to differentiate the two would be to contrast implausible items with plausible ones for which cloze values of all targets is zero. However, this would appear to be virtually impossible because of the difficulty in constructing items with reasonably plausible but zero cloze targets. Note that even Rayner et al.'s (2004) implausible items had a mean cloze of 1.5% and our atypical items were at 1.0%. Of course, one could hypothetically construct items that were matched for predictability (cloze) but differ in terms of plausibility rating, but this would be extremely difficult, if not impossible, for minimal pairs of stimuli (i.e., those differing only in terms of the target word).

How does this relate to the present studies? The mean cloze probability of Rayner et al.'s and Warren and McConnell's (2007) plausible items is not zero, and there is at least a 9% difference in cloze between their plausible and implausible items. Although the 9% difference in cloze is small compared to many studies investigating predictability effects (Rayner and Well, 1996), and predictability at around 10% to 20% is often considered relatively unpredictable, researchers have argued that such small differences can produce reading time effects. For example, McDonald and Shillcock (2003) demonstrated reading time effects that they interpreted as being due to transitional probability from one word to the next (e.g., high transitional probability: *accept defeat* vs. low: *accept losses*). Frisson, Rayner, and Pickering (2005) argued that there was a cloze difference of 8% versus 0.8% in McDonald and Shillcock's items, and that this difference may be large enough to show early reading-time effects of predictability. Therefore, Frisson et al. argued that McDonald and Shillcock's demonstration of transitional probability effects on early eye-movement measures “might actually be due to predictability” (p. 868). Under this criterion, our study, as well as those of Rayner et al. and Warren and McConnell, investigated predictability effects.

To further test whether the cloze probabilities explain the differential results, we reanalyzed both the self-paced reading and eye-tracking data after removing the items with the highest cloze values (and the matched-verb items to keep proper counterbalancing). After removing 14 items (those that have following 7 verbs: *avoid, cover, brew, color, wash, kill,* and *cut*, which include the items that were removed on the basis of the Experiment 2 priming data), mean cloze was reduced to 13% (ranging from 0% to 45%) for typical and 1% (0 to 10%) for atypical items. The 12% difference falls between Rayner et al.'s (2004) 9% difference and Warren and McConnell's (2007) 14% difference. Tables 7 to 12 present the results by measure and region. The self-paced reading data showed a significant 37 ms effect of event typicality at the patient noun. For the eye-tracking data, we obtained early effects at the patient. There was an 11 ms effect in first fixation duration that was significant by items but not by participants, a 25 ms effect in gaze duration that was significant in both analyses, a significant 42 ms effect in total reading time, and a marginal 5.0% difference in skipping frequencies. Hence, cloze probability does not seem to explain the the contrasting results in early reading times in our study as compared to Rayner et al. and Warren and McConnell.

It should be noted that there were 34 items in the previous analyses, which is approximately equal to that of Rayner et al. and Warren and McConnell. Therefore, the likely explanation for differences across studies is that fully crossing items provided increased sensitivity by controlling for extraneous variability that might occur due to the instruments. Again, we emphasize that the differences among the studies are not numerically large, although they are theoretically significant.

In summary, relatively mild implausibility does not appear to result in early effects on language comprehension unless the condition against which it is compared is highly plausible in terms of matching people's world knowledge. This lack of a difference may occur because people are used to reading or hearing language about somewhat implausible events, with the mild implausibility often being cleared up later in the sentence or discourse. However, early effects of plausibility were obtained with our items, even when they matched Rayner et al.'s (2004) and Warren and McConnell's (2007) in terms of predictability as measured by cloze. Furthermore, severe implausibility, or anomaly, at least when exemplified by a local verb-patient combination that cannot possibly go together, has clear and immediate effects. Note that even the difference between the implausible and anomalous conditions were reliable in both Rayner et al. (gaze durations) and Warren and McConnell (first fixation durations).

Multiple Representations and Selectional Restrictions

Rayner et al. (2004), and to a much greater extent, Warren and McConnell (2007), appealed to the notion of distinct multiple representations that are separable in terms of time course of use to account for the immediate differences obtained between anomalous and implausible items. They suggested that these differences could be due to a “theta assigning relation” that can “be detected on the basis of purely lexical information” (Rayner et al., p. 1297; but see Patson & Warren, 2010) or lexical selectional restriction information associated with a specific verb, which “is privileged over other kinds of knowledge in comprehension” because it “is represented in the lexicon and is available earlier than world/contextual knowledge” (Warren & McConnell, p.774).

When Chomsky (1965) discussed selectional restrictions, he suggested that these were general lexical features that were syntactic in nature. Although the examples on which Chomsky focused were easily accounted for with a relatively small set of abstract verb-general features (e.g., ± animate), this is not always the case in studies that appeal to selectional restrictions (Altmann & Kamide, 1999; Cottrell, 1988). For instance, Myers and Blumstein (2005) define selectional restrictions as “those semantic restrictions that any verb places on the arguments of that verb.” They provide an example in which “the verb ‘to mail’ requires that … the noun in object position must be something ‘mail-able’—it must be an object that does not exceed the size and weight restrictions of the US Postal Service.” (p. 278) Clearly, selectional restrictions such as ± mailable go well beyond abstract and verb-general information such as ± animate. On this definition, there could potentially be as many selctional features as there are verbs. Along this line, when one looks at Warren and McConnell's (2007) items in terms of selectional restriction violations, 12 items include animacy violations, but others require positing selectional restrictions such as ± inflatable, ± catchable, ± cookable, and ± mixable.

A major issue concerns whether anything is gained theoretically by calling this sort of information lexically-based selectional restrictions. Does positing this type of selectional restriction add anything beyond the meaning of the verb itself; that is, stating that the patient of *inflate* must be inflatable, the patient of *catch* must be catchable, and so on? The answer seems to be “no” (see Jackendoff, 2002, for a similar view). Consider a commonly used example of a selectional restriction, based on the verb *eat* (Altmann & Kamide, 1999; Nirenburg & Raskin, 2004). *Eat* requires a patient that is edible. Unlike animacy or humanness, determining whether an object is edible depends at least on who is eating it and knowledge about what the agent can ingest. Likewise, if the verb is *inflate*, what does it mean to say that a patient must denote something that is inflatable? One could assume that people learn, for example, that inflatable things are likely to be made of material that is not porous, are constructed so that they do not leak, are made of material that is expandable, are hollow, and possibly other features as well. But this would correspond to prototype or schema-style representations of thematic role event-based concepts (McRae, Ferretti, & Amyote, 1997), rather than (syntactically-relevant) selectional restrictions. Similarly, one could assume exemplar-based knowledge. That is, people might learn that inflatable things include objects such as balloons, balls, tires, and so on. One exemplar-based view of how children learn thematic role concepts is Tomasello's (1992) verb island hypothesis. In either case, people's knowledge would be based on their experience with inflatable objects, either first-hand experience with inflating these types of objects, or second-hand experience through observing someone else doing so in person, in movies or on television, or hearing or reading about inflating events. In all of these cases, the knowledge of the likelihood of something being inflatable is precisely event-based knowledge.

Conclusion

The present studies show that although there may be a distinction between lexical constraints on the one hand, and conceptual event-based knowledge on the other, this distinction has no relevance for the time course of the activation and use of these types of knowledge. Thus, there is no architecturally-determined delay of event knowledge during sentence comprehension. Furthermore, it seems that selectional restrictions, which are often considered to be lexical-grammatical constraints, and event-based knowledge, which is conceptual, may be, in fact, the same thing. There is consequently no reason for theories of sentence comprehension to attribute them to separate processing stages.

Acknowledgments

We thank Keith Rayner and Tessa Warren for their very helpful comments on an earlier draft. This work was supported by NIH grant HD053136 to KM, JLE, and MH, and NSERC grant OGP0155704 to KM.

Appendix

Sentences used in the self-paced reading (Experiment 2) and their weighted scores in the event generation norms (Experiment 1)

Sentences	Condition	Weighted Score
Jessie used a payment to avoid the annoying interest because he heard that rates were going up. He had the money anyway, and he figured that he could save more by paying now rather than later.	typical	13
Jessie used a shortcut to avoid the annoying interest because he heard that rates were going up. He had tried to cook up a number of complicated schemes, but in the end, he simply paid it from his chequing account.	atypical	0
Jessie used a shortcut to avoid the annoying traffic because he heard that there had been an accident on his usual route. He was in a hurry to get to work because he had an important meeting that morning.	typical	29
Jessie used a payment to avoid the annoying traffic because he heard that the freeway was jammed. He hated paying tolls, but he was in a hurry to get to work that morning.	atypical	0
Linda used a cauldron to brew the medicinal potion for her twelve year old daughter. Her daughter had been home sick for two days now, and Linda was hoping this would help.	typical	51
Linda used a kettle to brew the medicinal potion for her twelve	atypical	0

Sentences	Condition	Weighted Score
year old daughter. Her daughter had been home sick for two days now, and Linda was hoping this would help.		
Linda used a kettle to brew the medicinal tea for her twelve year old daughter. They often had tea as soon as Linda got home from work.	typical	86
Linda used a cauldron to brew the medicinal tea for her twelve year old daughter. They had just moved to a new apartment, and couldn't find the kettle anywhere.	atypical	3
James used a glove to catch the elusive baseball before it fell into someone else's hands. He couldn't believe that he had caught Vernon Wells' home run.	typical	72
James used a net to catch the elusive baseball before it fell into someone else's hands. His buddies had laughed at him for bringing it, but now he was really happy that he did.	atypical	4
James used a net to catch the elusive trout before it fell back into the water. It must have weighed at least 5 pounds.	typical	93
James used a glove to catch the elusive trout before it fell back into the water. He had been told by his guide that when handling a trout, big thick gloves are required.	atypical	0
Nancy used the dye to color her beautiful hair a bright shade of red. This was the fifth time in the last 6 months that she had changed hair color.	typical	64
Nancy used the crayons to color her beautiful hair a bright shade of red. She was trying to imitate her big sister who had just dyed her hair.	atypical	0
Nancy used the crayons to color her beautiful picture a bright shade of red. For some reason, she just felt red that day.	typical	51
Nancy used the dye to color her beautiful picture a bright shade of red. Her art project was supposed to be experimental, and she was trying silk screening for the first time.	atypical	0
John used a joystick to control the brand-new game that he bought yesterday. He had to stop using the regular controller because of the blisters on his both thumbs.	typical	62
John used a remote to control the brand-new game that he bought yesterday. He is now getting used to the non-traditional controller, and enjoying every moment of it.	atypical	0
John used a remote to control the brand-new television that he bought yesterday. He was very happy to see that his old universal remote works perfectly with his new TV.	typical	87
John used a joystick to control the brand-new television that he bought yesterday. He is a heavy gamer, and likes to control everything with his joystick.	atypical	0
Leslie used the white-out to cover the minor error after she had discovered the misspelling. She was too lazy to correct it and print it out again.	typical	89
Leslie used the band-aid to cover the minor error after she had cut her leg while shaving. She vowed to get a better razor as soon as possible because she was tired of cutting herself.	atypical	0
Leslie used the band-aid to cover the minor scrape after she had stopped the bleeding. She couldn't believe that a relatively small scrape would bleed that much.	typical	32
Leslie used the white-out to cover the minor scrape after she had picked up her resume in the parking lot. Luckily, she had some whiteout in her purse, because she didn't have time to print her resume again before the interview.	atypical	0
Susan used the scissors to cut the expensive paper that she needed for her project. She was making a poster for her Grade 11 geography class.	typical	98

Sentences	Condition	Weighted Score
Susan used the saw to cut the expensive paper that she needed for her project. She purposely wanted to create ragged edges on her background.	atypical	0
Susan used the saw to cut the expensive wood that she needed for her project. She was known as one of the best cabinet makers in the city.	typical	83
Susan used the scissors to cut the expensive wood that she needed for her project. The wood that she used for her art was thin and had to be handled with care.	atypical	0
Betty used a fork to eat the homemade pasta that was stuffed with large pieces of crab. She absolutely loved it.	typical	37
Betty used a spoon to eat the homemade pasta that was stuffed with large pieces of crab. She absolutely loved it.	atypical	3
Betty used a spoon to eat the homemade soup that was stuffed with large pieces of crab. She absolutely loved it.	typical	72
Betty used a fork to eat the homemade soup that was stuffed with large pieces of crab. She absolutely loved it.	atypical	5
Helen used a bottle to feed the adorable infant who was born just two weeks ago. She didn't want to do it, but she had been having trouble breastfeeding.	typical	101
Helen used a bucket to feed the adorable infant who was born just two weeks ago on her father's farm. She was worried that the pig wasn't eating well, so she fed him separately from the other pigs.	atypical	0
Helen used a bucket to feed the adorable pig who was born just two weeks ago. She was worried that he wasn't eating well, so she fed him separately from the others.	typical	52
Helen used a bottle to feed the adorable pig who was born just two weeks ago. She was worried that he wasn't gaining weight, so she was giving him some special treatment.	atypical	0
Sandra used a fireplace to heat the frozen cabin that her grandma left to her. She went there every Christmas holiday and skied at the local hill.	typical	23
Sandra used an oven to heat the frozen cabin that her grandma left to her. Her ski chalet had a beautiful old wood burning oven in the middle of it.	atypical	0
Sandra used an oven to heat the frozen pie that her grandma made for her. Her grandma always makes more than Sandra's family can eat.	typical	39
Sandra used a fireplace to heat the frozen pie that her grandma made for her. Her grandma always made food for her to take to the cabin by the ski hill.	atypical	0
David used a bat to hit the dirty baseball really hard while playing in the backyard. Unfortunately, he broke the neighbour's window.	typical	44
David used a hammer to hit the dirty baseball really hard while holding a beer in his other hand. They were playing the game that his brother had invented, and that they called “hammerball.”	atypical	0
David used a hammer to hit the dirty spike really hard while holding a beer in his other hand. He and his friends were pretty irresponsible.	typical	89
David used a bat to hit the dirty spike really hard while playing in the backyard. He stupidly put a large dent in his brand new bat.	atypical	0
Brian used a frame to hold the antique photograph during his art class. They were having an exhibition at his high school.	typical	94
Brian used a clamp to hold the antique photograph during his	atypical	0

Sentences	Condition	Weighted Score
art class. They were having an exhibition at his high school.		
Brian used a clamp to hold the antique wood during his art class. He had almost finished making his picture frame, and he was now painting it.	typical	50
Brian used a frame to hold the antique wood during his art class. For his project, he had painted a nature scene on barn board.	atypical	0
Joseph used a rifle to kill the unfortunate deer that they had been pursuing for an hour. It's the third deer that he and his father had caught so far.	typical	50
Joseph used a harpoon to kill the unfortunate deer that they had been pursuing for an hour. The bullets had not killed it, and they wanted to put it out of its misery.	atypical	11
Joseph used a harpoon to kill the unfortunate whale that they had been pursuing for three hours. It was his first catch since he had become the captain of an Inuit whaling crew.	typical	66
Joseph used a rifle to kill the unfortunate whale that they had been pursuing for three hours. It was his first catch since he had become the captain of an Inuit whaling crew.	atypical	0
Jamie used a match to light the cheap cigarette in the motel near the airport. She had just flown to Mexico from Detroit, and the US airport security had taken her lighter away.	typical	41
Jamie used a lantern to light the cheap cigarette in the motel near the airport. She had just flown to Mexico from Detroit, and the US airport security had taken her lighter away.	atypical	0
Jamie used a lantern to light the cheap room in the motel near the airport. The power had gone out, and the manager had brought everyone a lantern.	typical	34
Jamie used a match to light the cheap room in the motel near the airport. The power had just gone out, and she was trying to find her way around.	atypical	0
Willie used the scissors to open the old package that he found in the basement. After opening it, he realized it was a present that his parents bought for him for this coming Christmas.	typical	87
Willie used the can-opener to open the old package that he found in the basement. The can-opener was the only sharp thing he could find.	atypical	0
Willie used the can-opener to open the old soup that he found in the basement. It was one of those types without an easy-to-open pull tab on the lid.	typical	50
Willie used the scissors to open the old soup that he found in the cabin. He had been lost for two days and had just happened to see the cabin from the top of a nearby hill.	atypical	0
Casey used an alarm to protect the precious car that she purchased a month ago. She had learned recently that she could get a discount on her auto insurance by installing an anti-theft device.	typical	48
Casey used a fence to protect the precious car that she purchased a month ago. Some one had key-scratched her car in her open front yard a week ago, and it wasn't going to happen again.	atypical	0
Casey used a fence to protect the precious property that she purchased a month ago. Given her recent celebrity status, she was worried about stalkers and people coming onto the property.	typical	49
Casey used an alarm to protect the precious property that she purchased a month ago. She had heard about a number of recent burglaries in the neighbourhood.	atypical	5
Thomas used a horse to pull the old-fashioned carriage from	typical	64

Sentences	Condition	Weighted Score
the barn to the park. He felt privileged to chauffeur the Mayor in the annual parade.		
Thomas used a pick-up truck to pull the old-fashioned carriage from the barn to the park. The city asked him to display the 100-year old carriage in the annual parade.	atypical	0
Thomas used a pick-up truck to pull the old-fashioned trailer from the barn to the market. He's selling pumpkins and apples at the farmers' market.	typical	43
Thomas used a horse to pull the old-fashioned trailer from the barn to the park. Being Amish, he wasn't allowed to use gasoline powered vehicles.	atypical	0
Rene used the coins to purchase the hand-made candy at the farmers' market. She had a pocket full of dimes that her mother had given to her.	typical	29
Rene used the credit card to purchase the hand-made candy at the farmers' market. She couldn't believe how many chocolate bars she was buying for Halloween.	atypical	0
Rene used the credit card to purchase the hand-made clothes at the farmers' market. She got her first ever credit card this morning and went directly for the dress she's been keeping her eyes on.	typical	68
Rene used the coins to purchase the hand-made clothes at the farmers' market. She had a pocket full of townies and wanted to get rid them.	atypical	5
Rick used a rope to secure the large boat properly so that no strong winds would blow it away from the dock. The weather report called for an overnight storm.	typical	32
Rick used a lock to secure the large boat properly so that no one would break into his house boat while he walked around town. He had a lot of booze, plus his laptop, along with him on his week long trip.	atypical	0
Rick used a lock to secure the large door properly so that no one would break into his shed. He had stored both his and his wife's new racing bikes in there for the winter.	typical	61
Rick used a rope to secure the large door properly so that it wouldn't fall over in his truck on his way home. There were two large windows in it, and he didn't want any broken glass.	atypical	6
Jimmie used a dish to serve the fabulous dessert following the main course last night. He had artfully decorated a tiramisu, garnished it with a dusting of cocoa powder and with shaved semisweet chocolate and raspberries.	typical	17
Jimmie used a mug to serve the fabulous dessert following the main course last night. His kids always loved chocolate pudding.	atypical	0
Jimmie used a mug to serve the fabulous tea following the main course last night. He used a special blend of spices to make a delicious chai tea.	typical	71
Jimmie used a dish to serve the fabulous tea following the main course last night. His Chinese platter could hold four cups and was a nice touch when he was having company.	atypical	0
Terry used a shovel to spread the fresh dirt all around the flower bed so that it made a nice mound. This was the year that he was finally going to plant perennials in the front yard.	typical	65
Terry used a knife to spread the fresh dirt all around his terrarium. His lizards liked it when he made a little hill in the middle.	atypical	0
Terry used a knife to spread the fresh jam all around his toast so that it covered the whole thing. He also always made sure that there was at least one strawberry in each quadrant.	typical	53

Sentences	Condition	Weighted Score
Terry used a shovel to spread the fresh jam all around the world's largest loaf of bread. The Guinness representatives had already come and gone, and they were celebrating their new world record.	atypical	0
Donna used the hose to wash her filthy car after she came back from the beach. The strong and humid wind had covered her car with sand and salt.	typical	91
Donna used the shampoo to wash her filthy car after she came back from the beach. The strong and humid wind had covered her car with sand and salt.	atypical	8
Donna used the shampoo to wash her filthy hair after she came back from the beach. Her hair was covered with sand, and she wanted to wash it out.	typical	99
Donna used the hose to wash her filthy hair after she came back from the beach. Her hair was covered with sand, and she wanted to wash it out.	atypical	0
Kelly used a rag to wipe her greasy counter which became really dirty while she was making pancakes. It was first time cooking by herself, and she had a hard time mixing the batter.	typical	49
Kelly used a kleenex to wipe her greasy counter which became really dirty while she was making pancakes. It was first time cooking by herself, and she had a hard time mixing the batter.	atypical	10
Kelly used a kleenex to wipe her greasy nose which became really dirty while she was cleaning the garage. It was extremely dusty in there.	typical	91
Kelly used a rag to wipe her greasy nose which became really dirty while she was cleaning the garage. It was extremely dusty in there.	atypical	0
Mary used the paper to wrap the wonderful gift for her daughter who was coming to dinner that night. She was excited about her kid's twenty-fifth birthday.	typical	44
Mary used the tinfoil to wrap the wonderful gift for her daughter who was coming to dinner that night. She was excited about her kid's twenty-fifth birthday.	atypical	0
Mary used the tinfoil to wrap the wonderful leftovers for her daughter who was going back to her apartment. This was a Sunday ritual and Mary always tried to help her daughter as much as possible.	typical	31
Mary used the paper to wrap the wonderful leftovers for her daughter who was going back to her apartment. This was a Sunday ritual and Mary always tried to help her daughter as much as possible.	atypical	4

References

Altmann G, Kamide Y. Incremental interpretation at verbs: Restricting the domain of subsequent reference. Cognition. 1999; 73:247–264. [PubMed: 10585516]

Bicknell K, Elman JL, Hare M, McRae K, Kutas M. Effects of event knowledge in processing verbal arguments. Journal of Memory & Language. 2010; 63:489–505. [PubMed: 21076629]

Bornkessel I, Schlesewsky M. The extended argument dependency model: a neurocognitive approach to sentence comprehension across languages. Psychological Review. 2006; 113:787–821. [PubMed: 17014303]

Caplan D, Hildebrandt N, Waters GS. Interaction of verb selectional restrictions, noun animacy, and syntactic form in sentence processing. Language and Cognitive Processes. 1994; 9:549–585.

Chomsky, N. Aspects of the theory of syntax. MIT Press; Cambridge: 1965.

Chomsky, N. Reflections on language. Pantheon; New York: 1975.

Clifton C, Staub A. Parallelism and competition in syntactic ambiguity resolution. Language and Linguistics Compass. 2008; 2:234–250.

Cottrell, GW. A connectionist approach to word sense disambiguation. Morgan Kaufmann; San Mateo, CA: 1988.

DeLong KA, Urbach TP, Kutas M. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. Nature Neuroscience. 2005; 8:1117–1121.

Frazier L. Constraint satisfaction as a theory of sentence processing. Journal of Psycholinguistic Research. 1995; 24:437–469. [PubMed: 8531169]

Frazier L, Clifton C, Randall J. Filling gaps: Decision principles and structure in sentence comprehension. Cognition. 1983; 13:187–222.

Frisson S, Rayner K, Pickering M. Effects of contextual predictability and transitional probability on eye movements during reading. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2005; 31:862–877.

Garnsey SM, Pearlmutter N, Meyers E, Lotocky MA. The contribution of verb-bias and plausibility to the comprehension of temporarily ambiguous sentences. Journal of Memory and Language. 1997; 37:58–93.

Hardin, J.; Hilbe, J. Generalized Estimating Equations. Chapman and Hall/CRC; London: 2003.

Hare M, Elman JL, Tabaczynski T, McRae K. The wind chilled the spectators, but the wine just chilled: Sense, structure, and sentence comprehension. Cognitive Science. 2009:610–628. [PubMed: 19750146]

Jackendoff, R. Foundations of language: Brain, meaning, grammar, evolution. Oxford University Press; Oxford, UK: 2002.

Jaeger FT. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. Journal of Memory and Language. 2008; 59:434–446. [PubMed: 19884961]

Joseph HSSL, Liversedge SP, Blythe HI, White SJ, Gathercole SE, Rayner K. Childrens' and adults' processing of anomaly and implausibility during reading: Evidence from eye movements. Quarterly Journal of Experimental Psychology. 2008; 61:708–723.

Just M, Carpenter P, Wooley J. Paradigms and processes in reading comprehension. Journal of Experimental Psychology: General. 1982; 111:228–238. [PubMed: 6213735]

Kamide Y, Altmann GTM, Haywood SL. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. Journal of Memory and Language. 2003; 49:133–156.

Katz, JJ. Semantic theory. Harper & Row; New York: 1972.

Knoeferle P, Crocker MW, Scheepers C, Pickering MJ. The influence of the immediate visual context on incremental thematic role-assignment: evidence from eye-movements in depicted events. Cognition. 2005; 95:95–127. [PubMed: 15629475]

McDonald SA, Shillcock RC. Eye movements reveal the on-line computation of lexical probabilities. Psychological Science. 2003; 14:648–652. [PubMed: 14629701]

McRae K, Boisvert S. Automatic semantic similarity priming. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1998; 24:558–572.

McRae K, Ferretti TR, Amyote L. Thematic roles as verb-specific concepts. Language and Cognitive Processes. 1997; 12:137–176.

McRae K, Spivey-Knowlton MJ, Tanenhaus MK. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. Journal of Memory and Language. 1998; 38:283–312.

Morris RK. Lexical and message-level sentence context effects on fixation times in reading. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1994; 20:92–103.

Murray WS. The nature and time course of pragmatic plausibility effects. Journal of Psycholinguistic Research. 2006; 35:79–99. [PubMed: 16502145]

Myers EB, Blumstein SE. Selectional restriction and semantic priming effects in normals and Broca's aphasics. Journal of Neurolinguistics. 2005; 18:277–296.

Nelson, DL.; McEvoy, CL.; Schreiber, TA. The University of South Florida word association, rhyme, and word fragment norms. 1998. http://www.usf.edu/freeassociation/

Nirenburg, S.; Raskin, V. Ontological Semantics. The MIT Press; Cambridge, MA: 2004.

Patson ND, Warren T. Eye movements when reading implausible sentences: Investigating potential structural influences on semantic integration. Quarterly Journal of Experimental Psychology. 2010; 63:1516–1532.

Pollatsek A, Well AD. On the use of counterbalanced designs in cognitive research: A suggestion for a better and more powerful analysis. Journal of Experimental Psychology: Learning, Memory & Cognition. 1995; 21:785–794.

Rayner K. Eye movements in reading and information processing: 20 years of research. Psychological Bulletin. 1998; 124:372–422. [PubMed: 9849112]

Rayner K, Warren T, Juhasz BJ, Liversedge SP. The effect of plausibility on eye movements in reading. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2004; 30:1290–1301.

Rayner K, Well AD. Effects of contextual constraint on eye movements in reading: A further examination. Psychonomic Bulletin & Review. 1996; 3:504–509.

Richardson DC, Spivey MJ, Barsalou LW, McRae K. Spatial representations activated during real-time comprehension of verbs. Cognitive Science. 2003; 27:767–780.

Sanford A, Garrod S. Memory-based approaches and beyond. Discourse Processes. 2005; 39:205–224.

Schlesinger, IM. Cognitive space and linguistic case. Cambridge University Press; Cambridge, NY: 1995.

Sperber, D.; Wilson, D. Relevance: communication and cognition. Blackwell; Oxford, UK: 1986.

Staub A, Rayner R, Pollatsek A, Hyönä J, Majewski H. The time course of plausibility effects on eye movements in reading: Evidence from noun-noun compounds. Journal of Experimental Psychology. 2007; 33:1162–1169. [PubMed: 17983320]

Thornton R, MacDonald MC. Plausibility and grammatical agreement. Journal of Memory and Language. 2003; 48:740–759.

Tomasello, M. First verbs: A case study of early grammatical development. Cambridge University Press; Cambridge: 1992.

Trueswell JC, Tanenhaus MK, Garnsey SM. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. Journal of Memory and Language. 1994; 33:285–318.

Van Berkum JJA, Brown CM, Zwitserlood P, Kooijman V, Hagoort P. Anticipating upcoming words in discourse: Evidence from ERPs and reading times. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2005; 31:443–467.

Van Gompel RPG, Pickering MJ, Pearson J, Liversedge SP. Evidence against competition during syntactic ambiguity resolution. Journal of Memory and Language. 2005; 52:284–307.

Vu H, Kellas G, Petersen E, Metcalf K. Situation-evoking stimuli, domain of reference, and the incremental interpretation of lexical ambiguity. Memory & Cognition. 2003; 31:1302–1315.

Warren T, McConnell K. Investigating effects of selectional restriction violations and plausibility violation severity on eye-movements in reading. Psychonomic Bulletin & Review. 2007; 14:770–775. [PubMed: 17972747]

Warren T, McConnell K, Rayner K. Effects of context on eye movements when reading about plausible and impossible events. Journal of Experimental Psychology: Learning, Memory and Cognition. 2008; 34:1001–1010.

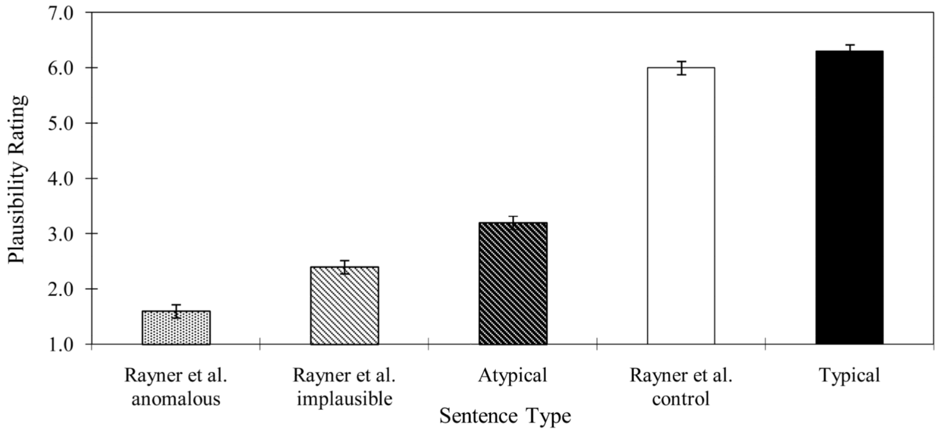

Figure 1.

Plausibility ratings from Norming Study 2.

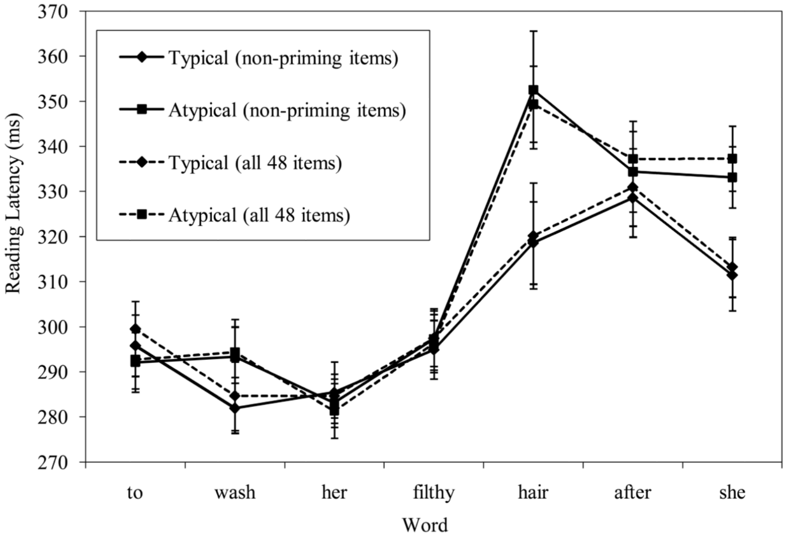

Figure 2.

Comparing Experiment 1 reading latencies (dotted lines) with those resulting from removing the items that showed priming in Experiment 2 (solid lines).

Table 1

Reading Latency (in ms) and *F*-statistics for Experiment 1, Self-paced Reading.

	Word						
	to	***wash***	***her***	***filthy***	***hair***	***after***	***she***
Latency							
Atypical	293	294	281	296	349	337	337
Typical	299	285	285	297	320	331	313
Difference	−6	9	−4	−1	29	6	24
F-test							
$F_1(1, 35)$	1.12	1.40	0.25	0.02	6.68	0.30	4.11
$F_2(1, 47)$	0.62	1.08	0.18	0.02	5.16	0.35	5.07
MSE_1	740	1224	770	987	2295	2409	2542
MSE_2	1787	2126	1417	1461	3963	2758	2746
p_1	.30	.24	.62	.88	.01	.59	.05
p_2	.44	.31	.67	.89	.03	.56	.03

Table 2

First Fixation Duration (in ms) and *F*-statistics for Experiment 3, Eyetracking.

	Region				
	R1 ***to wash***	**R2** ***her filthy***	**R3** ***hair***	**R4** ***after she***	**R5** ***came***
Duration					
Atypical	202	208	217	201	190
Typical	204	210	206	195	190
Difference	−2	−2	11	6	0
F-test					
$F_1(1, 31)$	0.21	0.24	4.08	2.22	0.31
$F_2(1, 47)$	0.09	0.24	6.55	1.15	0.04
MSE_1	3204	3264	3057	4134	3057
MSE_2	3712	4495	3724	4169	6076
p_1	.65	.63	.052	.15	.58
p_2	.77	.63	.014	.29	.85

Table 3

First Pass Reading Time (or Gaze Duration, in ms) and *F*-statistics for Experiment 3, Eyetracking.

	Region				
	R1 ***to wash***	**R2** ***her filthy***	**R3** ***hair***	**R4** ***after she***	**R5** ***came***
Duration					
Atypical	270	313	244	270	212
Typical	270	313	223	264	205
Difference	0	0	21	6	7
F-test					
$F_1(1, 31)$	0.03	0.01	10.42	0.52	1.16
$F_2(1, 47)$	0.01	0.01	11.85	0.15	0.43
MSE_1	14019	8724	5715	17598	4264
MSE_2	20180	26250	7074	14979	12556
p_1	.88	.94	.003	.48	.29
p_2	.92	.95	.001	.70	.51

Table 4

Regression Path Duration (or Go-Past times, in ms) and *F*-statistics for Experiment 3, Eyetracking.

	Region				
	R1 ***to wash***	**R2** ***her filthy***	**R3** ***hair***	**R4** ***after she***	**R5** ***came***
Duration					
Atypical	322	386	284	335	310
Typical	326	372	272	318	259
Difference	−4	14	12	17	51
F-test					
$F_1(1, 31)$	0.32	0.96	1.17	1.16	7.08
$F_2(1, 47)$	0.36	0.34	1.03	0.78	7.23
MSE_1	37200	55208	17062	48223	53112
MSE_2	83779	63403	30870	52952	44659
p_1	.58	.34	.29	.29	.012
p_2	.55	.56	.32	.38	.009

Table 5

Total Reading Time (in ms) and *F*-statistics for Experiment 3, Eyetracking.

	Region				
	R1 ***to wash***	**R2** ***her filthy***	**R3** ***hair***	**R4** ***after she***	**R5** ***came***
Duration					
Atypical	400	484	302	360	242
Typical	356	445	259	335	225
Difference	44	39	43	25	17
F-test					
$F_1(1, 31)$	7.74	15.23	12.94	2.10	2.07
$F_2(1, 47)$	5.28	3.94	12.35	2.07	1.48
MSE_1	45561	19828	27391	60369	29538
MSE_2	73237	65785	28004	58091	24641
p_1	.009	.001	.001	.16	.16
p_2	.026	.053	.001	.16	.23

Table 6

Skipping Percentages for Experiment 3, Eyetracking. The inferential statistics (Wald Chi-Squares) refer to the event typicality effect as established via logit binomial *GEE*s (Hardin & Hilbe, 2003) by participants (WCS_1) and items (WCS_2).

	Region				
	R2 ***to shampoo***	**R3** ***her filthy***	**R4** ***hair***	**R5** ***after she***	**R6** ***came***
Skipping Frequency					
Atypical	7.2%	2.4%	12.3%	14.4%	16.3%
Typical	6.4%	2.4%	16.9%	11.3%	16.2%
Difference	0.8%	0%	−4.6%	3.1%	0.1%
Logit Binomial GEEs					
$WCS_1(1)$	0.277	<0.001	2.872	1.036	.001
$WCS_2(1)$	0.172	<0.001	4.276	2.467	.002
p_1	.60	.99	.09	.31	.97
p_2	.68	1.00	.04	.12	.96

Table 7

Reading Latency (in ms) and *F*-statistics for Experiment 1, Self-paced Reading, with High Predictability Items removed.

	Word						
	to	***wash***	***her***	***filthy***	***hair***	***after***	***she***
Latency							
Atypical	288	295	284	297	359	330	333
Typical	294	283	289	296	322	330	315
Difference	−6	12	−5	1	37	0	18
F-test							
F_1(1, 35)	0.51	1.52	0.32	0.01	6.46	0.00	1.93
F_2(1, 47)	0.59	0.98	0.43	0.00	4.53	0.002	1.73
MSE_1	1198	1880	1568	1355	3907	2457	3134
MSE_2	1746	2259	1728	1708	4796	3071	3279
p_1	.48	.23	.58	.93	.02	.99	.17
p_2	.45	.33	.52	.99	.04	.96	.20

Table 8

First Fixation Duration (in ms) and *F*-statistics for Experiment 3, Eyetracking, with High-predictability Items Excluded.

	Region				
	R1 ***to wash***	**R2** ***her filthy***	**R3** ***hair***	**R4** ***after she***	**R5** ***came***
Duration					
Atypical	204	212	216	201	192
Typical	200	212	205	192	186
Difference	4	0	11	9	6
F-test					
$F_1(1, 31)$	0.43	0.05	2.62	2.91	1.01
$F_2(1, 33)$	0.73	0.02	4.78	2.17	0.87
MSE_1	3288	3674	3459	3626	3211
MSE_2	3422	4803	3770	4010	4974
p_1	.52	.83	.12	.10	.32
p_2	.40	.90	.04	.15	.36

Table 9

First Pass Reading Time (or Gaze Duration, in ms) and *F*-statistics for Experiment 3, Eyetracking, with High-predictability items Excluded.

	Region				
	R1 ***to wash***	**R2** ***her filthy***	**R3** ***hair***	**R4** ***after she***	**R5** ***came***
Duration					
Atypical	267	313	247	270	221
Typical	266	318	222	259	200
Difference	1	−5	25	11	21
F-test					
$F_1(1, 31)$	0.02	0.11	8.55	1.32	4.25
$F_2(1, 33)$	0.02	0.19	9.77	0.41	4.34
MSE_1	9652	12424	6623	17238	9740
MSE_2	12355	28650	7810	16220	9835
p_1	.89	.76	.006	.26	.05
p_2	.89	.67	.003	.52	.04

Table 10

Regression Path Duration (or Go-Past times, in ms) and *F*-statistics for Experiment 3, Eyetracking, with High-predictability items Excluded.

	Region				
	R1 ***to wash***	**R2** ***her filthy***	**R3** ***hair***	**R4** ***after she***	**R5** ***came***
Duration					
Atypical	325	386	285	320	313
Typical	300	380	263	320	258
Difference	25	6	22	0	55
F-test					
F_1(1, 31)	2.43	0.20	2.05	0.00	5.05
F_2(1, 33)	1.47	0.00	1.29	0.02	6.12
MSE_1	25427	51817	28205	47606	58167
MSE_2	57023	66788	38205	41507	42269
p_1	.13	.66	.16	.97	.03
p_2	.23	.95	.26	.91	.02

Table 11

Total Reading Time (in ms) and *F*-statistics per region for Experiment 3, Eyetracking, with High-predictability items Excluded.

	Region				
	R1 ***to wash***	**R2** ***her filthy***	**R3** ***hair***	**R4** ***after she***	**R5** ***came***
Duration					
Atypical	391	475	306	353	258
Typical	347	445	254	330	223
Difference	44	30	52	24	36
F-test					
$F_1(1, 31)$	7.50	4.52	15.09	1.18	3.85
$F_2(1, 33)$	4.55	1.80	11.14	1.53	6.35
MSE_1	29422	24071	22556	62025	41626
MSE_2	65295	67969	33799	47620	23047
p_1	.01	.04	.001	.29	.06
p_2	.04	.19	.002	.22	.02

Table 12

Skipping Percentages for Experiment 3, Eyetracking, with High-predictability items Excluded). The inferential statistics (Wald Chi-Squares) refer to the event typicality effect as established via logit binomial *GEEs* (Hardin & Hilbe, 2003) by participants (WCS_1) and items (WCS_2).

	Region				
	R1 ***to wash***	**R2** ***her filthy***	**R3** ***hair***	**R4** ***after she***	**R5** ***came***
Skipping Frequencies					
Atypical	8.0%	2.3%	13.3%	14.8%	16.5%
Typical	6.5%	3.0%	18.3%	12.2%	16.7%
Difference	1.5%	−0.7%	−5.0%	2.6%	−0.2%
Logit Binomial GEEs					
$WCS_1(1)$	0.558	0.397	2.154	0.755	0.003
$WCS_2(1)$	0.475	0.281	3.013	1.69	0.003
p_1	.46	.53	.14	.39	.96
p_2	.49	.60	.08	.19	.97