Research Article

Psychological Science
1–17

DOI: 10.1177/0956797618807674
www.psychologicalscience.org/PS
SAGE

Tracking Colisteners' Knowledge States During Language Comprehension

Olessia Jouravlev[1,2], Rachael Schwartz[3], Dima Ayyash[1], Zachary Mineroff[1], Edward Gibson[1], and Evelina Fedorenko[1,4,5]
[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology; [2]Institute of Cognitive Science, Carleton University; [3]Cognitive and Linguistic Sciences, Wellesley College; [4]Department of Psychiatry, Harvard Medical School; and [5]Department of Psychiatry, Massachusetts General Hospital, Boston, Massachusetts

Abstract
When we receive information in the presence of other people, are we sensitive to what they do or do not understand? In two event-related-potential experiments, participants read implausible sentences (e.g., "The girl had a little beak") in contexts that rendered them plausible (e.g., "The girl dressed up as a canary for Halloween"). No semantic-processing difficulty (no N400 effect) ensued when they read the sentences while alone in the room. However, when a confederate was present who did not receive the contexts so that the critical sentences were implausible for him or her, participants exhibited processing difficulty: the social-N400 effect. This effect was obtained when participants were instructed to adopt the confederate's perspective—and critically, even without such instructions—but not when performing a demanding comprehension task. Thus, unless mental resources are limited, comprehenders engage in modeling the minds not only of those individuals with whom they directly interact but also of those individuals who are merely present during the linguistic exchange.

Keywords
communication, perspective taking, joint actions, social cognition, ERPs, N400, social N400, open data, open materials

Received 11/30/17; Revision accepted 8/2/18

Communication requires coordination of linguistic and nonlinguistic behavior between conversation partners. We keep track of what information is in the *common ground*, that is, what knowledge, beliefs, and attitudes are shared between us and our conversation partner, and what is in the *privileged ground*, that is, what information may not be available to the partner (Clark, 1993; Levinson, 2000). Producers consider their comprehenders' perspectives when planning utterances (Brennan, Galati, & Kuhlen, 2010; Fussell & Krauss, 1992), and comprehenders take into account producers' mental states when interpreting their utterances (Brown-Schmidt, Gunlogson, & Tanenhaus, 2008; Hanna, Tanenhaus, & Trueswell, 2003; Heller, Grodner, & Tanenhaus, 2008). However, communicative situations often involve more than two individuals. For example, we often receive information in the presence of other people, with whom we may not be directly interacting. Here, we asked whether a comprehender is sensitive to what the colisteners understand.

A priori, we might hypothesize that comprehenders do not model the minds of people around them except when directly interacting with them. After all, mentalizing is costly. Indeed, some researchers have argued that we may not even always model the mind of our conversation partners and, at least initially, adopt an egocentric perspective in interpreting and formulating utterances (Keysar, Barr, Balin, & Brauner, 2000; Keysar, Lin, & Barr, 2003; Lane & Ferreira, 2008). However, mentalizing is such a core part of building successful relationships that it is also easy to imagine that we track the perspectives of anyone present during a conversation (Clark & Carlson, 1982).

Corresponding Author:
Olessia Jouravlev, Massachusetts Institute of Technology, Department of Brain and Cognitive Sciences, 43 Vassar St., Building 46, Room 3037, Cambridge, MA 02139
E-mail: olessiaj@mit.edu

Some evidence suggesting that people represent mental states of all physically present individuals comes from studies of nonlinguistic actions. Individuals performing tasks alongside each other appear to track task requirements and action alternatives of other individuals, even when this compromises performance (Sebanz, Knoblich, & Prinz, 2003; Sebanz, Knoblich, Prinz, & Wascher, 2006). In the language domain, Wilkes-Gibbs and Clark (1992) used a referential communication task—in which a speaker (the director) gives a listener (the matcher) instructions for rearranging images of non-nameable objects—and showed that the director assumed that a passive colistener had established the same common ground with him or her as the actively participating matcher. When the colistener later became the matcher, the director kept using the names that were established in communication with the original matcher.

More recently, Rueschemeyer, Gardner, and Stoner (2015) used event-related potentials (ERPs) to ask whether comprehenders are sensitive to the knowledge states of their colisteners. Participants read implausible sentences (e.g., "The boy had gills") in contexts that rendered them plausible (e.g., "In the boy's dream, he could breathe under water"; Example 1c), along with control plausible sentences in supportive contexts (Example 1a), and implausible sentences in contexts that did not make them plausible (Example 1b):

(1a) "The fishmonger prepared the fish. The fish had gills."

(1b) "The boy woke up at dawn. The boy had gills."

(1c) "In the boy's dream, he could breathe under water. The boy had gills."

The critical manipulation was whether participants were alone or were told to take the perspective of a confederate sitting next to them in front of the same screen. Because the context sentences were presented over headphones and only the participants had headphones, the target sentence—presented visually—in the critical condition (Example 1c) made sense to the participants but not the confederates.

The presence of a confederate did not affect the processing of the target sentence in Examples 1a and 1b: The word "gills" elicited a larger N400 in the latter condition (Kutas & Hillyard, 1980). In the critical condition (Example 1c), when participants were alone, they experienced no processing difficulty at "gills," in line with prior work (Nieuwland & Van Berkum, 2006b; Van Berkum, Koornneef, Otten, & Nieuwland, 2007). Critically, in the presence of a confederate, an N400 effect was observed. The authors took this effect as evidence that participants model the knowledge states of their colisteners and thus experience empathetic confusion, and they termed this effect the *social N400*.

The social-N400 effect is a promising implicit marker of representing other people's minds. However, Rueschemeyer et al. (2015) explicitly instructed participants to adopt the confederate's perspective. It is therefore unclear whether the effect would be obtained without explicit instruction. To illuminate the conditions under which people model the knowledge states of colisteners, we conducted two ERP experiments that collectively had four task conditions: (a) explicit instructions to consider the confederate's perspective (as in Rueschemeyer et al.'s, 2015, study, in line with current emphasis on replication; Aarts et al., 2015), (b) a sensibility-judgment task in which participants were not asked to consider the confederate's perspective, (c) a passive-reading task, and (d) a challenging comprehension task. These diverse tasks allowed us to assess the degree to which the social N400 is obtained spontaneously, as well as how it may be affected by cognitive load.

An additional, more exploratory, goal was to investigate individual differences in perspective taking, which may be affected by linguistic skills (Farrant, Fletcher, & Maybery, 2006), executive abilities (Brown-Schmidt, 2009; Ryskin, Benjamin, Tullis, & Brown-Schmidt, 2015), and social competence (Baron-Cohen et al., 1985; Dawson & Fernald, 1987). Here, we focused on social competence and tested whether better social skills are associated with better perspective taking.

Experiment 1

Participants performed two comprehension tasks while electroencephalographic (EEG) activity was recorded: a sensibility-judgment task and a passive-reading task. In both tasks, they listened to the context sentences over headphones and then read the target sentences (see Fig. 1). For each of the tasks, participants performed two sessions: one in which they were alone (alone session) and another in which a confederate was present (joint session). Importantly, the confederate did not have headphones, and thus the target sentences were plausible for the participants but implausible for the confederates in the critical, context-dependent condition, as discussed in the Materials section. For the sensibility-judgment task, during the alone session, participants decided whether the target sentences made sense to them, and during the joint session, they decided whether the target sentences made sense to the confederate (referred to as "the other person"). The passive-reading instructions were identical in the alone and joint sessions.

The first goal of the experiment was to test the robustness of the social N400 (Aarts et al., 2015).

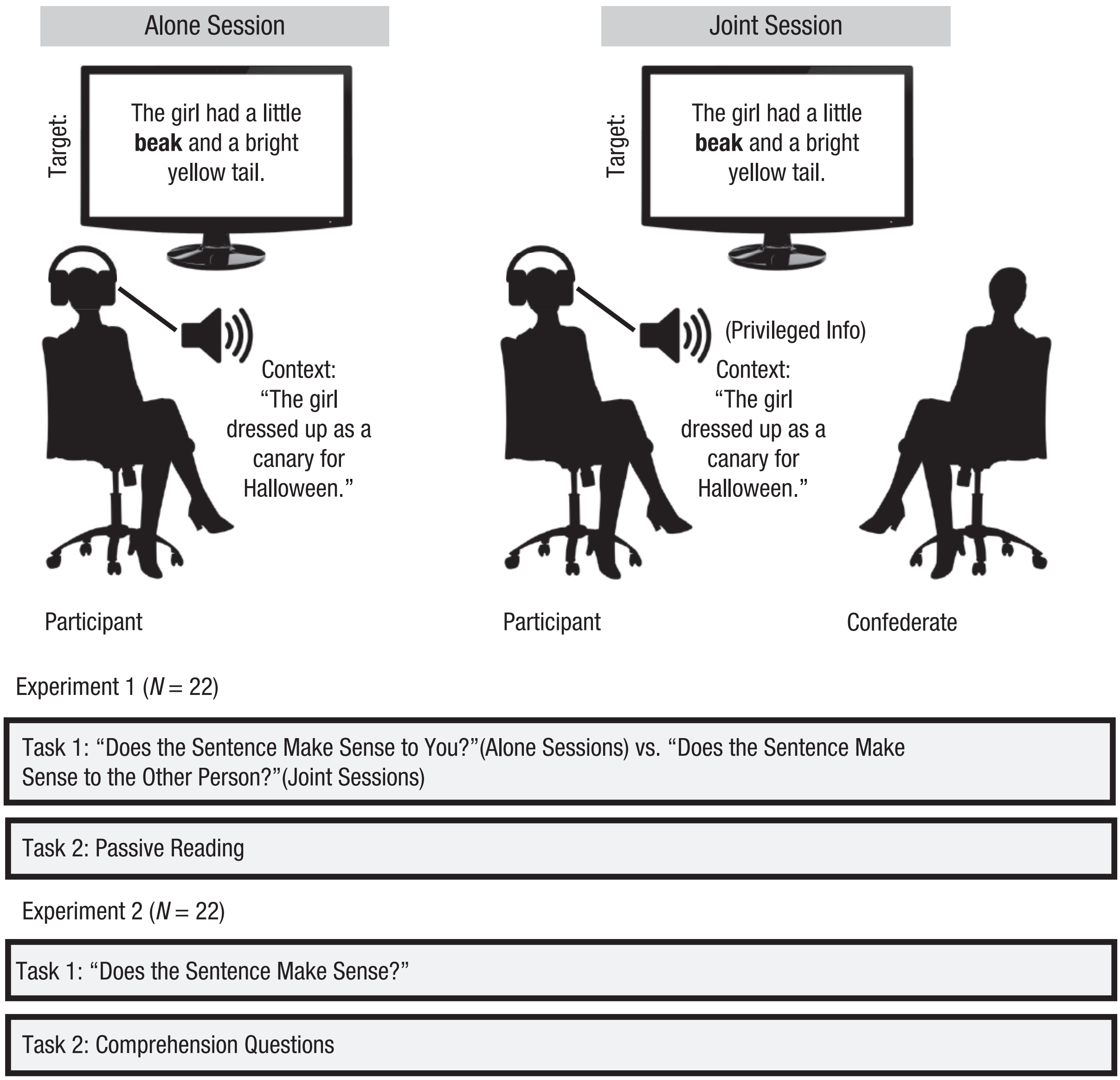

Fig. 1. The experimental setup in Experiments 1 and 2. Participants listened to the context sentences presented to them over headphones and then read the target sentences presented on a computer screen (the critical word is presented here in boldface for illustrative purposes only; during the experiment, the target sentences were presented one word at a time with no boldface). In the alone session, participants were alone in the testing room. In the joint session, they were accompanied by another participant (a confederate). The confederate could see the target sentences on the computer screen but did not have access to the context sentences.

Rueschemeyer et al. (2015) used a between-subjects design with different participants in the alone and joint sessions, with no discussion of matching the groups on linguistic, social, or executive abilities, which have been shown to affect language processing, including in ERP paradigms (Nieuwland & Van Berkum, 2006a; Tanner & Van Hell, 2014; van den Brink et al., 2012). Furthermore, the critical sentences described fantasy worlds. Such sentences require comprehenders to construct an alternate reality and have been shown to be costly even in supportive contexts (Ferguson & Cane, 2015; Hald, Steenbeek-Planting, & Hagoort, 2007; cf. Nieuwland & Van Berkum, 2006b). We used a within-subjects design, and the critical materials described implausible but physically possible events. If robust, the social-N400 effect should be replicated in a within-subjects design and generalize beyond the kinds of materials used in the original study. The second, critical, goal was to test whether the social N400 is obtained without the explicit instruction to adopt the confederate's perspective.

Method

Participants. Twenty-four participants (12 men; age: M = 24.8 years, SD = 3.9, range = 19–32) from the Massachusetts Institute of Technology (MIT) and the surrounding Boston community participated for payment. The sample size was determined on the basis of prior research on electrophysiological correlates of sentence processing (Nieuwland & Van Berkum, 2006b; Rueschemeyer et al., 2015; Van Berkum et al., 2007). Data collection stopped when we reached the enrollment goal. All participants were right-handed (by self-report) native speakers of English with normal or corrected-to-normal vision and hearing. None of the participants reported any neurodevelopmental disorders, psychiatric disorders, or language impairments. All participants gave written informed consent in accordance with the requirement of MIT's Committee on the Use of Humans as Experimental Subjects. Data from 2 participants were excluded (1 because of technical errors that resulted in data loss and 1 because of an excessive number of artifacts in the EEG signal, with more than 25% of trials affected), leaving 22 participants for the analysis.

Materials. The context sentences were recorded by a female native speaker of English. Each recording lasted for a maximum of 4 s, with shorter sentences padded with silence at the end. One hundred items, exemplified in Examples 2 through 4, were constructed with three conditions each: plausible (Examples 2a, 3a, and 4a), implausible (Examples 2b, 3b, and 4b), and context dependent (Examples 2c, 3c, and 4c). (Target words are presented here in boldface for illustrative purposes.)

(2a) Plausible: "The kids were looking at a canary in the pet store with great interest. The bird had a little **beak** and a bright yellow tail."

(2b) Implausible: "Anna was definitely a very cute child. The girl had a little **beak** and a bright yellow tail."

(2c) Context dependent: "The girl dressed up as a canary for Halloween. The girl had a little **beak** and a bright yellow tail."

(3a) Plausible: "Amanda is a renowned lawyer in her city. Amanda wears a suit to **work** every day."

(3b) Implausible: "Amanda works as a secretary at a law company. Amanda wears a swimsuit to **work** every day."

(3c) Context dependent: "Amanda is a swimming instructor at the local pool. Amanda wears a swimsuit to **work** every day."

(4a) Plausible: "John, a builder, is on his way to work. The builder is heading to the **construction** site."

(4b) Implausible: "John, a librarian, is on his way to work. The librarian is heading to the **construction** site."

(4c) Context dependent: "A new library is being erected in downtown Boston. The librarian is heading to the **construction** site."

Each trial consisted of two sentences. The first sentence (length: M = 10 words, range = 4–19) varied across the three conditions and served to establish the appropriate discourse context. The second, critical, sentence (length: M = 11 words, range = 5–17) was identical in the implausible and context-dependent conditions and minimally different (by 1 word) in the plausible condition. The target word was embedded in the second sentence. Its position varied between Words 3 and 12, and it never appeared in the sentence's final position, to minimize response preparation and wrap-up effects (Hagoort, 2003).

The materials were constructed so that the target word in the plausible condition was semantically plausible and highly predictable in the context of the second sentence alone (i.e., the first sentence was not necessary; it merely provided additional information). In the implausible condition, the target word was semantically implausible and unpredictable in the context of the second sentence, and the first sentence did not make the target word more plausible or predictable. Finally, in the context-dependent condition, the target word was semantically implausible and unpredictable in the context of the second sentence alone, but the first sentence rendered it plausible (the full set of materials is available on the Open Science Framework at https://osf.io/x7kma/).

Prior to the ERP study, the materials were normed in two sentence-completion studies. In the first study, participants were shown the first sentence and the second sentence up to but not including the target word (e.g., "The kids were looking at a canary in the pet store with great interest. The bird had a little . . .") and asked to complete the sentence so that it would make sense. The second study was the same except that the first context sentence was not included (e.g., "The bird had a little . . ."). We posted surveys for 150 workers on Amazon's Mechanical Turk. All workers were paid for their participation. Participants were asked to indicate their native language, but payment was not contingent on their responses, and only native English speakers were included in the analyses. For each study, three experimental lists were created so that each list contained only one version of an item. Each list was presented to 25 participants (with trials in random order for each participant). The first word in the completions was used to calculate the cloze probability of the target word.

These norming studies confirmed that we had succeeded in creating the desired manipulations (see Table 1). In particular, the target word was highly expected in the plausible condition, either with or without the first context sentence (cloze probabilities: .57 and .56, respectively), and highly unexpected in the implausible condition, either with or without the first context sentence (cloze probabilities: .01 in both studies). Importantly, in the context-dependent condition, the target word was quite expected when the context sentence was included (cloze probability: .35) but not when only

Table 1. Cloze Probability Values for Target Words in the Two Sentence Contexts Presented in the Norming Studies

	Sentence fragments		
Condition	First and second sentences	Second sentence only	Comparison
Plausible	.57	.56	$t(198) = 0.22, p = .82$
Implausible	.01	.01	$t(198) = 0.76, p = .45$
Context dependent	.35	.01	$\boldsymbol{t(198) = 13.86, p < .001}$

Note: Target words were presented in the context of (a) the first and second sentences (e.g., "The girl dressed up as a canary for Halloween. The girl had a little . . .") and (b) the second sentence only (e.g., "The girl had a little . . ."). Boldface indicates a significant comparison.

the second sentence was included (cloze probability: .01).

Procedure. At the beginning of the study, participants were introduced to another participant (a confederate) and told that they would complete two sentence-comprehension tasks: Each would consist of two sessions (one in which they are in the room by themselves and one in which the other participant is in the room with them). They were fitted with the electroencephalogram cap and headphones and informed that the other participant would not be privy to any information that they receive over the headphones. Next, participants were invited to a sound-attenuated and electrically shielded booth where stimuli were presented to them over the headphones (the context sentences) and on the computer monitor (the target sentences). The confederate joined for two of the sessions, as detailed below.

The 300 trials were distributed across four experimental lists following a Latin square design; each list consisted of 75 trials and contained only one version of an item (plausible, implausible, or context dependent), and there were 25 trials per condition. Each participant saw all four lists across the four combinations of task and session (sensibility judgment and alone, sensibility judgment and joint, passive reading and alone, and passive reading and joint). The pairing between lists and task-session combinations and the task-session order varied across participants. The trials within each list were presented in random order for each participant.

Across the four task-session combinations, each trial started with a simultaneous presentation of (a) the fixation cross on the computer screen and (b) the context sentence over the headphones (for 4,000 ms). Next, the target sentence was presented on the screen word by word at a rate of 450 ms per word. Each word was followed by a 100-ms interstimulus interval, with an additional 400 ms after the last word of the sentence. Further, at the end of each trial in the sensibility-judgment task, a question was presented for 2,000 ms—"Does it make sense to you?" during the alone session or "Does it make sense to the other person?" during the joint session—and participants were instructed to answer by pressing one of two buttons on the keyboard. If participants did not respond within 2,000 ms, the next trial began. In the passive-reading task, to help participants stay awake and alert, we presented an image of a finger pressing a button for 400 ms at the end of each trial, and participants were instructed to press a button on the button box when the image appeared. During the joint sessions, the confederate was seated next to the participant, facing the same computer screen, and was provided with a button box. The confederate was instructed, in the presence of the participant, to perform the same task as the participant (i.e., to answer the question in the sensibility-judgment task or to press a button in the passive-reading task). Each task lasted approximately 15 min, and participants were given breaks between sessions.

After the ERP experiment, participants completed a general background and language history questionnaire, as well as three standardized tests aimed at assessing social competence: (a) the Autism-Spectrum Quotient (AQ; Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001), (b) the "Reading the Mind in the Eyes" Test (RMET; Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001), and (c) the Empathy Quotient (EQ; Baron-Cohen & Wheelwright, 2004). The entire experiment took approximately 2 hr.

EEG recording. EEG activity was recorded from 32 scalp sites (10-20 system positioning), a vertical eye channel for detecting blinks, a horizontal eye channel to monitor for saccades, and two additional electrodes affixed to the skin directly above the mastoid bone for use as reference channels. The ActiveTwo system (BioSemi, Amsterdam, The Netherlands) with active Ag-AgCl electrodes mounted on an elastic cap (Electro-Cap, Eaton, OH) was used. All channels were referenced off-line to an average of the mastoid channels. EEG activity was recorded at a sampling rate of 512 Hz. Following standard procedures in ERP research, we then filtered the signal off-line (bandpass = 0.1–40 Hz), and excluded trials with blinks, eye movements, muscle artifacts, and skin potentials prior to averaging and analyses. Across participants, an average of 7.7% of trials (*SD* = 4.7, range = 1.3–14.7) was excluded.

Behavioral analyses. For the behavioral responses in the sensibility-judgment task, the type of response (yes/no) was submitted as a dependent variable to a generalized linear mixed-effects model (glmer), and reaction time (RT) was submitted as a dependent variable to a linear mixed-effects model (lmer). Analyses were performed with the *lme4* package (Bates, Maechler, Bolker, & Walker, 2018) in the R programming environment (R Core Team, 2008). Each model included experimental manipulations—session (alone vs. joint) and condition (plausible vs. implausible vs. context dependent)—as fixed effects, and participants and items as random effects (the intercepts were always included, and the slopes were included unless their inclusion prevented model convergence). Significance of main and interaction effects was assessed using the likelihood-ratio tests (i.e., models with the target effects included were compared with models without those effects). Significant effects were followed up by planned comparisons performed with the *multcomp* package (Hothorn et al., 2017) in R. Bonferroni correction was used to account for the number of comparisons (n = 3). The button-press responses in the passive-reading task were examined (to ensure that participants were awake and alert) but not analyzed.

EEG and ERP analyses. The continuous EEG signal was divided into epochs over a window from 200 ms prior to the target word onset to 800 ms after target onset. The 200-ms window prior to the target onset was used as the prestimulus baseline. To obtain ERPs, we averaged epochs across trials within a condition for each target electrode (see below) and participant. For visualization purposes, the responses were further averaged across participants (see Figs. 2 and 3).

The ERP component of interest was the N400 (Kutas & Hillyard, 1980), a negative deflection observed at centroparietal locations on the scalp 300 ms to 600 ms after stimulus onset, typically peaking around 400 ms. Given the typical scalp distribution of the N400 (Curran, Tucker, Kutas, & Posner, 1993), we restricted the analyses to the eight central and parietal sites (C3, Cz, C4, CP1, CP2, P3, Pz, P4). Further, given the typical time course of the N400, we used a 200-ms window of interest for analysis (350–550 ms after word onset). The amplitudes within this window were averaged for each condition, session, electrode, and participant and used as dependent measures in the repeated measures analyses of variance (ANOVAs). We used ANOVAs rather than linear mixed-effects models to analyze the ERP data (a) to make the results comparable with those of the Rueschemeyer et al. (2015) study and (b) because single-trial-level data were not readily available.

Following Rueschemeyer et al. (2015), we first conducted a 2 × 3 × 8 ANOVA for each of the tasks (sensibility judgment and passive reading), with session (alone vs. joint), condition (plausible vs. implausible vs. context dependent), and electrode (C3 vs. Cz vs. C4 vs. CP1 vs. CP2 vs. P3 vs. Pz vs. P4) as within-subjects factors (using the Greenhouse-Geisser correction). Significant interactions between session and condition were followed up with planned comparisons to examine ERP magnitudes in the plausible versus implausible versus context-dependent conditions separately in the alone and joint sessions. Significance values were Bonferroni corrected for the number of comparisons within each session (n = 3).

Results

Behavioral results. Average proportions of "yes" responses and average RTs in the sensibility-judgment task are reported in Table 2. Linear mixed-effects models revealed a significant interaction between the experimental manipulations—session (alone vs. joint) and condition (plausible vs. implausible vs. context dependent)—for both dependent measures—"yes" responses: $\chi^2(2)$ = 221.98, p < .001; RTs: $\chi^2(2)$ = 10.94, p = .004. Planned comparisons revealed that during the alone session, the proportion of "yes" responses differed across all three condition pairs (plausible vs. context dependent: z = 5.16, p = .001, corrected throughout the present article when multiple comparisons were performed; plausible vs. implausible: z = 18.04, p < .001; context dependent vs. implausible: z = 17.85, p < .001), with the largest proportion of "yes" responses being given in the plausible condition (.95), followed by the context-dependent condition (.85), and finally, by the implausible condition (.25).

During the joint session, the proportion of "yes" responses was significantly higher in the plausible condition (.94) than in the implausible condition (.22; z = 18.95, p < .001) or the context-dependent condition (.27; z = 18.01, p < .001). The latter two conditions did not differ significantly (z = 2.37, p = .06). Thus, as expected, participants judged sentences in the plausible condition as making sense (to them and to the confederate) and sentences in the implausible condition as not making sense (to them or to the confederate). Most importantly, responses in the context-dependent condition varied between sessions: During the alone session, participants judged the sentences as making sense to them, and during the joint session, they judged the sentences as not making sense to the confederate.

With respect to RTs, during the alone session, participants took longer to decide on the sensibility of sentences in the implausible condition (M = 924 ms)

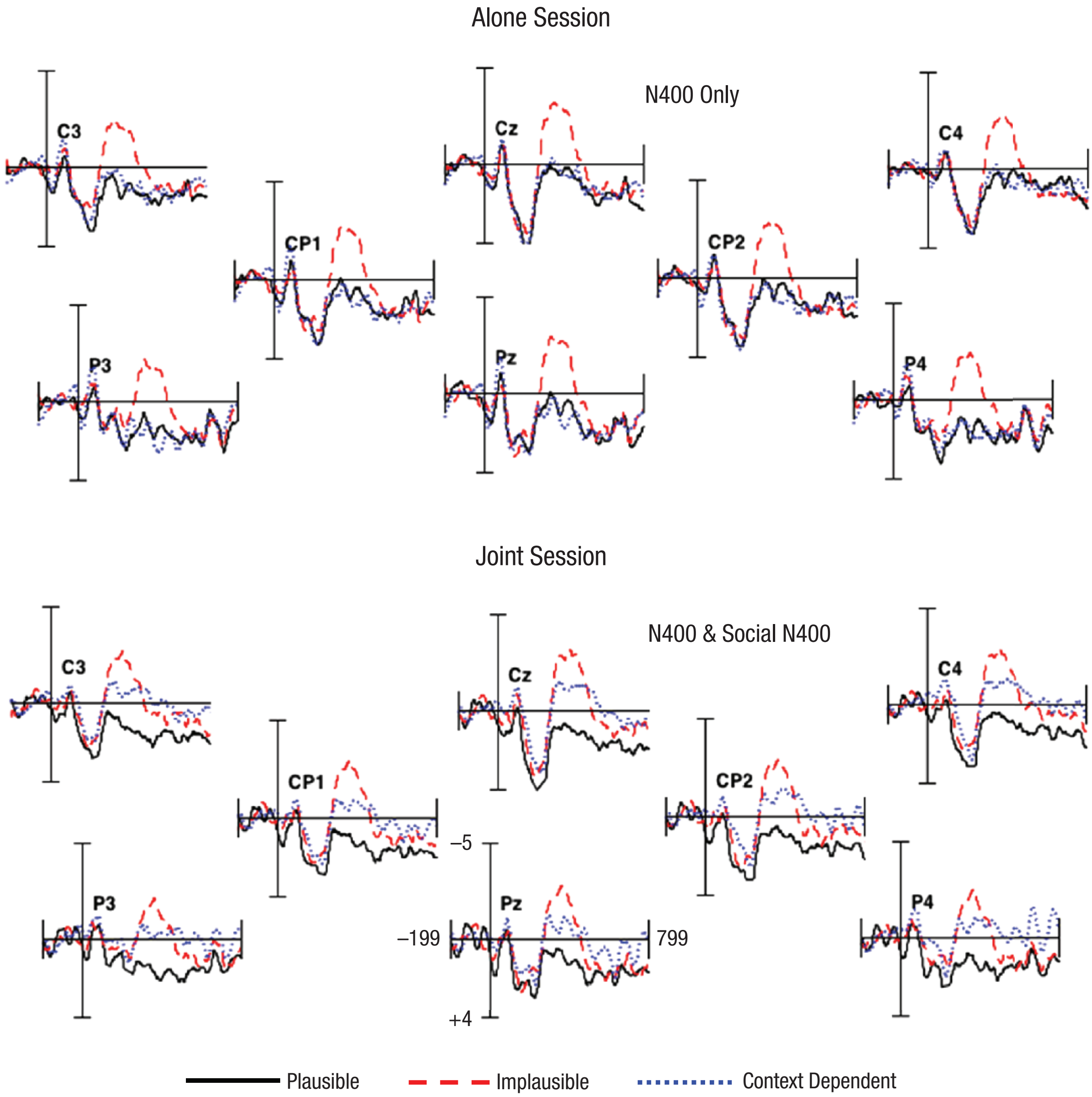

Fig. 2. Event-related-potential waveforms evoked by the target words in the plausible, implausible, and context-dependent conditions in the sensibility-judgment task in Experiment 1 ("Does the sentence make sense to you/Does the sentence make sense to the other person?"). Results are shown separately for the alone and joint sessions and for each of the eight target electrodes. In each graph, the *x*-axis shows time in milliseconds, and the *y*-axis shows event-related-potential amplitudes in microvolts. Key effects are indicated in each session. The N400 label refers to the increased negativity in the 350-ms to 550-ms time window in response to the implausible compared with the plausible condition. The social N400 refers to the increased negativity in the same time window in response to the context-dependent compared with the plausible condition.

than in the plausible condition (M = 746 ms; z = 7.24, p < .001) or the context-dependent condition (M = 789 ms; z = 5.69, p < .001). The latter two conditions did not differ significantly (z = 1.55, p = .32). During the joint session, RTs were significantly longer in the implausible condition (M = 903 ms) and context-dependent condition (M = 855 ms) than in the plausible condition (M = 740 ms; implausible vs. plausible: z = 7.51, p < .001; context dependent vs. plausible: z = 5.72, p < .001). The two former conditions did not differ significantly (z = 1.78, p = .22). Thus, there was a processing cost for sentences in the context-dependent condition during the joint session (the data are available at https://osf.io/x7kma/).

ERP results. The waveforms evoked by the target words in the three conditions (plausible, implausible, context dependent) during the alone and joint sessions are shown in Figure 2 (for the sensibility-judgment task) and Figure 3 (for the passive-reading task). Mean ERP

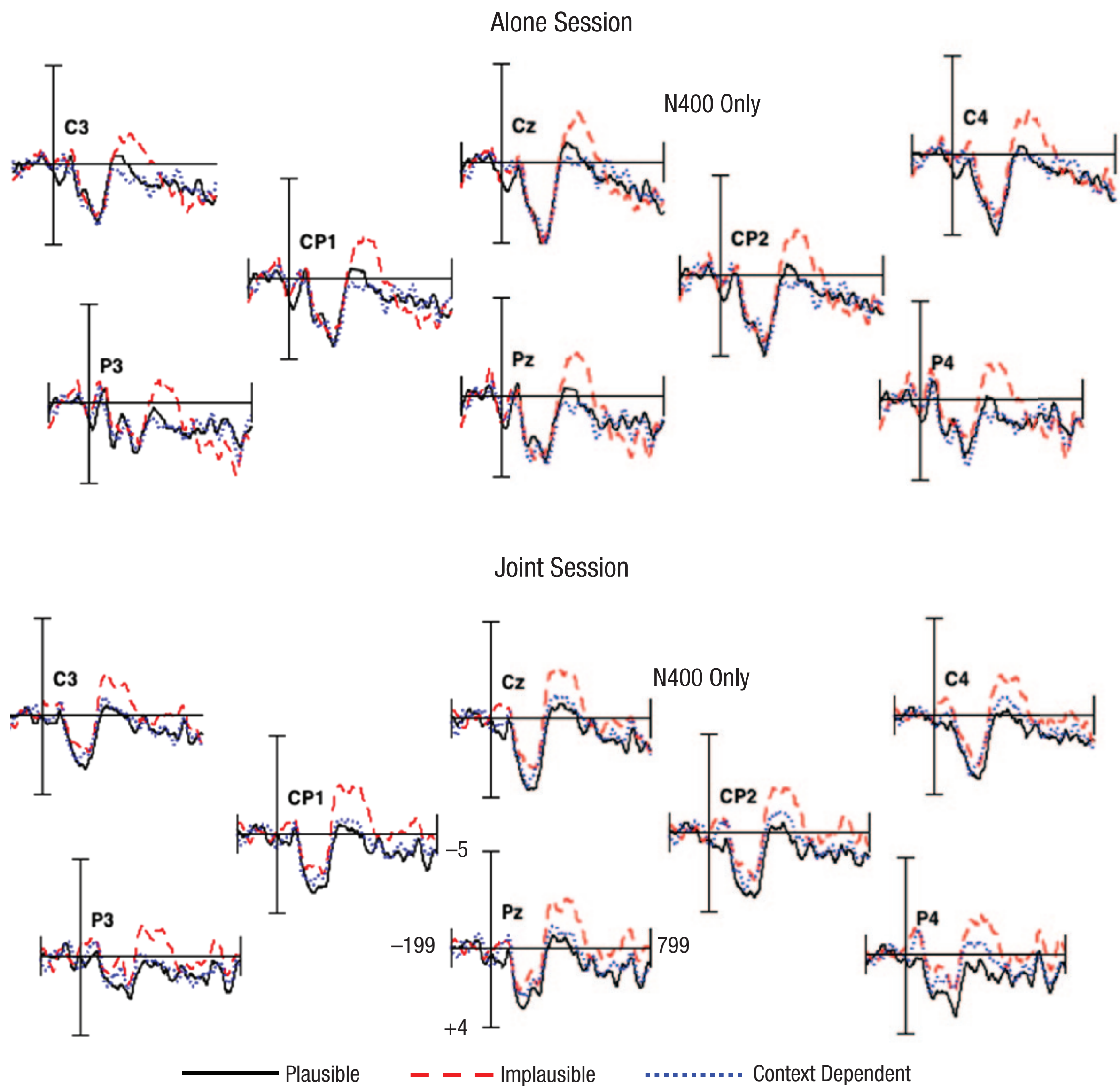

Fig. 3. Event-related-potential waveforms evoked by the target words in the plausible, implausible, and context-dependent conditions in the passive-reading task in Experiment 1. Results are shown separately for the alone and joint sessions and for each of the eight target electrodes. In each graph, the *x*-axis shows time in milliseconds, and the *y*-axis shows event-related-potential amplitudes in microvolts. Key effects are indicated in each session. The N400 label refers to the increased negativity in the 350-ms to 550-ms time window in response to the implausible compared with the plausible condition.

amplitudes in the N400 time window are provided in Table 3.

For the sensibility-judgment task, we observed a main effect of condition, $F(2, 42) = 26.93$, $p < .001$, $\eta^2 = .56$, with average ERPs being significantly more negative in the implausible condition than in the plausible condition (−1.54 vs. 1.05, respectively), $t(21) = 5.29$, $p < .001$, or context-dependent condition (−1.54 vs. 0.12, respectively), $t(21) = 3.39$, $p = .004$. The latter two conditions did not differ significantly, $t(21) = 1.91$, $p = .18$. Critically, we observed a significant interaction between condition and session, $F(2, 42) = 7.57$, $p = .002$, $\eta^2 = .27$. Planned comparisons revealed that during the alone session, the average magnitude of the N400 was reduced in the plausible condition (0.87) and context-dependent condition (1.00) compared with the implausible condition (−1.60)—plausible vs. implausible: $t(21) = 4.24$, $p < .001$; context dependent vs. implausible: $t(21) = 4.01$, $p < .001$. The plausible and context-dependent conditions did not differ significantly, $t(21) = 0.21$, $p = .99$. The fact that the pattern observed in the context-dependent condition was similar to that in the plausible

Table 2. Average Proportion of "Yes" Responses and Average Reaction Time (RT) in the Sensibility-Judgment Task in Experiment 1

	Alone session		Joint session	
Condition	"Yes"	RT	"Yes"	RT
Plausible	.95 (.05)	746 (73)	.94 (.05)	740 (78)
Implausible	.25 (.05)	924 (83)	.22 (.05)	903 (82)
Context dependent	.85 (.08)	789 (84)	.27 (.05)	855 (82)

Note: RTs are given in milliseconds. Standard errors of the mean by participants are provided in parentheses.

condition is in line with prior work that established that contextual information can alleviate the processing difficulty of sentences that are implausible out of context (Nieuwland & Van Berkum, 2006b; Van Berkum et al., 2007).

When a confederate was present, the average magnitude of the N400 in the plausible condition (1.22), but not the context-dependent condition (−0.77), was significantly reduced compared with the implausible condition (−1.47)—plausible vs. implausible: $t(21) = 5.03$, $p < .001$; context dependent vs. implausible: $t(21) = 1.30$, $p = .58$. ERPs in the context-dependent condition were significantly more negative than in the plausible condition, $t(21) = 3.72$, $p = .001$. Thus, patterns of ERPs observed during the joint session suggest that participants experienced difficulty in processing the sentences in the context-dependent condition.

For the passive-reading task, we observed a main effect of condition, $F(2, 42) = 14.44$, $p < .001$, $\eta^2 = .41$, with average ERPs being significantly more negative in the implausible condition than the plausible condition (−1.25 vs. 0.32, respectively), $t(21) = 3.15$, $p = .008$, or the context-dependent condition (−1.25 vs. 0.26, respectively), $t(21) = 3.02$, $p = .01$. The latter two conditions did not differ significantly, $t(21) = 0.12$, $p = .99$. We found no evidence of a significant interaction between condition and session, $F(2, 42) = 0.72$, $p = .49$, $\eta^2 = .03$. The data for this experiment are available at https://osf.io/x7kma/. To ensure that the difference in the N400 time window was not driven by differences emerging from earlier ERP effects (e.g., a P200), we performed a peak-to-peak analysis (see the Supplemental Material). The results of this analysis replicated the results of the mean-amplitude analysis.

Experiment 2

In the sensibility-judgment task of Experiment 1, in which participants were explicitly instructed to adopt the confederate's perspective, we replicated the social-N400 effect (Rueschemeyer et al., 2015) in a within-subjects design with new materials. We failed to observe the social N400 in the passive-reading task, possibly because participants were not engaged deeply enough with the task.

In Experiment 2, we modified the tasks to shed further light on the conditions under which the social N400 is obtained. The materials and basic setup were the same, except for two changes. First, in the sensibility-judgment task, participants were not explicitly instructed whose perspective to adopt. The question simply asked, "Does the sentence make sense?" and they could decide for themselves whose perspective to take. And second, the passive-reading task was replaced with a demanding comprehension-question task, to evaluate the effect of cognitive load on the social N400.

Method

Participants. Twenty-three participants (10 males; age: $M = 26.1$ years, $SD = 5.4$, range = 20–40) from MIT and the surrounding Boston community participated for payment. All were right-handed (by self-report) native speakers

Table 3. Average Event-Related-Potential Amplitudes (in Microvolts) in the N400 Time Window Evoked by the Target Words in the Sensibility-Judgment Task and in the Passive-Reading Task of Experiment 1

	Sensibility-judgment task		Passive-reading task	
Condition	Alone session	Joint session	Alone session	Joint session
Plausible	0.87 (0.41)	1.22 (0.37)	0.38 (0.52)	0.27 (0.49)
Implausible	−1.60 (0.40)	−1.47 (0.40)	−1.07 (0.36)	−1.43 (0.29)
Context dependent	1.00 (0.08)	−0.77 (0.36)	0.72 (0.39)	−0.21 (0.51)

Note: Standard errors of the mean by participants are provided in parentheses.

of English with normal or corrected-to-normal vision and hearing. None participated in Experiment 1. All participants gave written informed consent in accordance with the requirement of MIT's Committee on the Use of Humans as Experimental Subjects. Data from 1 participant were excluded because of an excessive number of artifacts in the EEG signal, with more than 25% of trials affected, leaving 22 participants for the analysis.

Materials. The materials were identical to those used in Experiment 1. For the comprehension-question task, a yes/no question was written for each condition of each item. The questions were constructed to encourage deep engagement with the materials: Answering them correctly required both (a) keeping the context and the target sentences active in working memory and (b) reasoning about the content of the sentences. For example, for the sentence "The kids were looking at a canary in the pet store with great interest. The bird had a little beak and a bright yellow tail," the question asked, "Was the bird for sale?" and for the sentence "Mary is making an unusual dessert from bacon. Mary sprinkled the bacon with sugar and nutmeg," the question asked, "Is Mary a vegetarian chef?" (All materials are available at https://osf.io/x7kma/.)

Procedure and analyses. The procedure was identical to that used in Experiment 1, except for the changes noted above. In particular, in the sensibility-judgment task, participants were not explicitly instructed whether they should take their own perspective or the perspective of the confederate when making the judgment: During both the alone and joint sessions, the question simply asked, "Does the sentence make sense?" As in Experiment 1, the question was presented for 2,000 ms, and participants were instructed to answer by pressing one of two buttons on the keyboard. If participants did not respond within the 2,000-ms window, the next trial began. The passive-reading task was replaced with a comprehension task with yes/no questions about the content of the materials. The question was presented for 3,000 ms, and participants were instructed to answer by pressing one of two buttons on the keyboard. If participants did not respond within the 3,000-ms window, the next trial began.

As in Experiment 1, in the joint sessions, the confederate was seated next to the participant, facing the same computer screen, and was provided with a button box. The confederate was instructed, in the presence of the participant, to perform the same task as the participant. As in Experiment 1, participants completed three standardized tests aimed at assessing social competence, and the entire experiment took approximately 2 hr.

The EEG recording procedure was identical to that in Experiment 1. Across participants, an average of 6.2% of trials (SD = 5.6, range = 0.7–17.7) was excluded because of the presence of artifacts. The analyses were identical to those in Experiment 1, except that for the behavioral analyses, both the sensibility-judgment task and the comprehension-question task were analyzed.

Results

Behavioral results. Average proportions of "yes" responses and average RTs in the sensibility-judgment task are reported in Table 4, and average accuracies and average RTs in the comprehension-question task are reported in Table 5. In the sensibility-judgment task, results were similar to those in Experiment 1: Linear mixed-effects models revealed a significant interaction between session (alone vs. joint) and condition (plausible vs. implausible vs. context dependent), although in this experiment, the interaction was present in only the response data, $\chi^2(2) = 14.54$, $p < .001$, not in RTs, $\chi^2(2) = 1.73$, $p = .42$. Planned comparisons revealed that during the alone session, average proportions of "yes" responses were higher in the plausible condition (.95) and context-dependent condition (.94) than in the implausible condition (.22)—plausible vs. implausible: $z = 17.3$, $p < .001$; context dependent vs. implausible: $z = 17.05$, $p < .001$. The plausible and context-dependent conditions did not differ significantly ($z = 1.04$, $p = .64$). Thus, participants made use of the information provided in the context sentences to make sense of the target sentences in the context-dependent condition. During the

Table 4. Average Proportion of "Yes" Responses and Average Reaction Time (RT) in the Sensibility-Judgment Task in Experiment 2

	Alone session		Joint session	
Condition	"Yes"	RT	"Yes"	RT
Plausible	.95 (.05)	823 (40)	.98 (.04)	849 (44)
Implausible	.22 (.06)	967 (45)	.24 (.06)	987 (45)
Context dependent	.94 (.05)	832 (45)	.89 (.04)	894 (45)

Note: RTs are given in milliseconds. Standard errors of the mean by participants are provided in parentheses.

Table 5. Average Accuracy (Proportion Correct) and Average Reaction Time (RT) in the Comprehension-Question Task in Experiment 2

	Alone session		Joint session	
Condition	Accuracy	RTs	Accuracy	RTs
Plausible	.96 (.04)	1,780 (109)	.95 (.05)	1,791 (115)
Implausible	.91 (.06)	1,745 (114)	.92 (.07)	1,773 (109)
Context dependent	.91 (.06)	1,774 (112)	.92 (.07)	1,820 (112)

Note: RTs are given in milliseconds. Standard errors of the mean by participants are provided in parentheses.

joint session, proportions of "yes" responses differed across all three condition pairs (plausible vs. context dependent: $z = 5.41$, $p < .001$; plausible vs. implausible: $z = 17.21$, $p < .001$; context dependent vs. implausible: $z = 16.29$, $p < .001$), with the largest average proportion of "yes" responses being given in the plausible condition (.98), followed by the context-dependent condition (.89), and finally, by the implausible condition (.24). This pattern suggests that when accompanied by a confederate, at least some of the participants adopted the confederate's perspective some of the time when deciding whether the sentence made sense.

In the comprehension task, participants were highly accurate across conditions (range = 0.91–0.96), with no evidence of an interaction between session (alone vs. joint) and condition (plausible vs. implausible vs. context dependent) in either the accuracies, $\chi^2(2) = 0.69$, $p = .71$, or the RTs, $\chi^2(2) = 0.63$, $p = .73$.

ERP results. The waveforms evoked by the target words in the three conditions (plausible, implausible, context dependent) during the alone and joint sessions are shown in Figure 4 (for the sensibility-judgment task) and Figure 5 (for the comprehension-question task). Mean ERP amplitudes in the N400 time window are provided in Table 6.

For the sensibility-judgment task, we observed a main effect of condition, $F(2, 42) = 20.02$, $p < .001$, $\eta^2 = .49$, with average ERPs being significantly more negative in the implausible condition than in the plausible condition (−1.12 vs. 0.70, respectively), $t(21) = 4.04$, $p < .001$, or the context-dependent condition (−1.12 vs. 0.43, respectively), $t(21) = 3.45$, $p = .003$. The latter two conditions did not differ significantly, $t(21) = 0.59$, $p = .99$. Critically, we observed a marginally significant interaction between condition and session, $F(2, 42) = 3.33$, $p = .05$, $\eta^2 = .14$. Planned comparisons revealed that during the alone session, the average magnitude of the N400 was reduced in the plausible condition (0.46) and context-dependent condition (0.78) compared with the implausible condition (−0.92)—plausible vs. implausible: $t(21) = 3.36$, $p = .01$; context dependent vs. implausible: $t(21) = 3.95$, $p = .003$. The plausible and context-dependent conditions did not differ significantly, $t(21) = 0.81$, $p = .81$. Thus, similar to participants in Experiment 1, participants appeared to have no difficulty understanding the sentences in the context-dependent condition when they processed these sentences alone.

When a confederate was present, the average magnitude of the N400 in the plausible condition (0.93) and context-dependent condition (0.07) was significantly reduced compared with the implausible condition (−1.32)—plausible vs. implausible: $t(21) = 4.86$, $p < .001$; context dependent vs. implausible: $t(21) = 3.85$, $p = .003$. Further, ERPs in the context-dependent condition were significantly more negative than in the plausible condition, $t(21) = 2.98$, $p = .02$. Thus, as in Experiment 1, participants experienced difficulty in processing the sentences in the context-dependent condition when a confederate was present.

For the comprehension-question task, we observed a main effect of condition, $F(2, 42) = 12.48$, $p < .001$, $\eta^2 = .37$, with average ERPs being significantly more negative in the implausible condition than the plausible condition (−1.14 vs. 0.08, respectively), $t(21) = 3.35$, $p = .004$, or the context-dependent condition (−1.14 vs. 0.42, respectively), $t(21) = 4.29$, $p < .001$. The latter two conditions did not differ significantly, $t(21) = 0.93$, $p = .74$. The interaction between condition and session was not significant, $F(2, 42) = 0.02$, $p = .98$, $\eta^2 = .01$. The data for this experiment are available at https://osf.io/x7kma/.

An Exploratory Analysis: The Effect of Social Competence on Perspective Taking

Method

For this analysis, we combined the ERP data from the sensibility-judgment task performed in the presence of a confederate in Experiments 1 and 2, which yielded data for a total of 44 participants. For each participant,

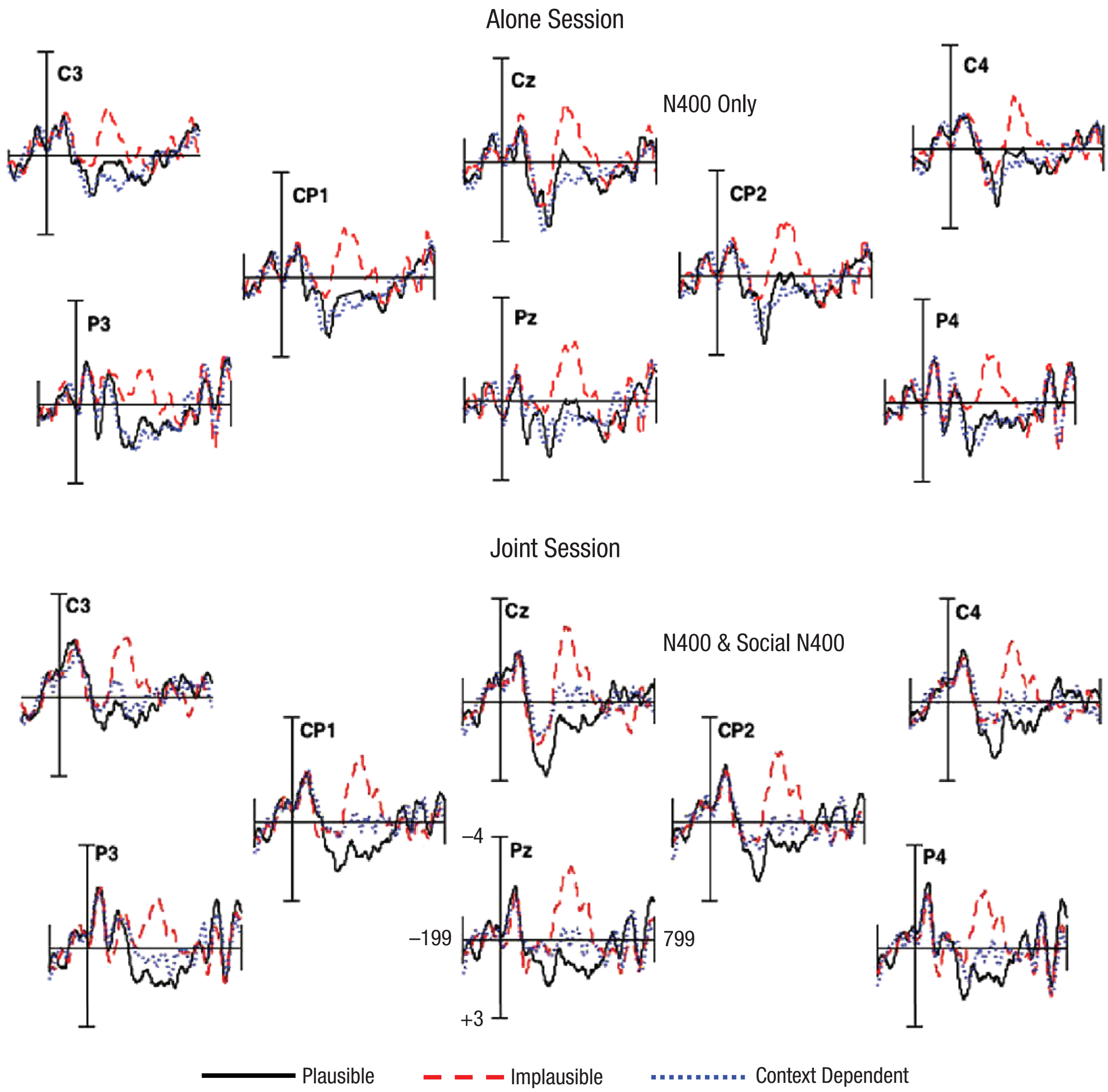

Fig. 4. Event-related-potential waveforms evoked by the target words in the plausible, implausible, and context-dependent conditions in the sensibility-judgment task in Experiment 2 ("Does the sentence make sense?"). Results are shown separately for the alone and joint sessions and for each of the eight target electrodes. In each graph, the *x*-axis shows time in milliseconds, and the *y*-axis shows event-related-potential amplitudes in microvolts. Key effects are indicated in each session. The N400 label refers to the increased negativity in the 350-ms to 550-ms time window in response to the implausible compared with the plausible condition. The social N400 refers to the increased negativity in the same time window in response to the context-dependent compared with the plausible condition.

we computed the average magnitude of the social-N400 effect (plausible – context dependent) and the average magnitude of the classic N400 effect (plausible – implausible). We next performed three regressions predicting the size of the social N400 from each of the behavioral measures (the AQ, Baron-Cohen, Wheelwright, Skinner, et al., 2001; the RMET, Baron-Cohen, Wheelwright, Hill, et al., 2001; and the EQ, Baron-Cohen & Wheelwright, 2004), controlling for the size of the N400 effect. The results were Bonferroni corrected for the number of comparisons ($n = 3$).

Results

Descriptive statistics for the three tests of social competence (see Table 7) suggest that our participants varied substantially in their social-skill level, and this variability can thus be related to the size of the

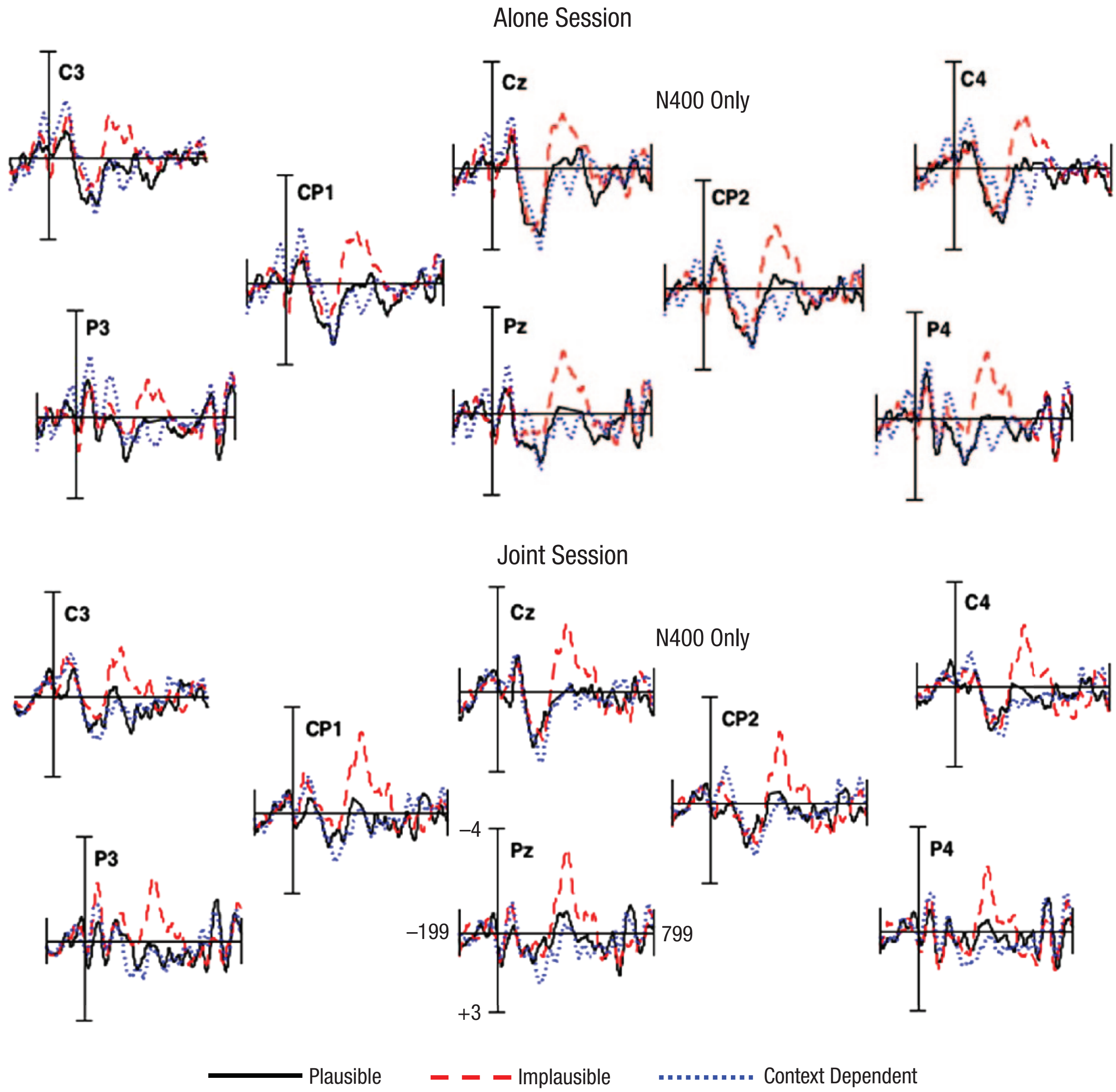

Fig. 5. Event-related-potential waveforms evoked by the target words in the plausible, implausible, and context-dependent conditions in the comprehension-question task in Experiment 2. Results are shown separately for the alone and joint sessions and for each of the eight target electrodes. In each graph, the *x*-axis shows time in milliseconds, and the *y*-axis shows event-related-potential amplitudes in microvolts. Key effects are indicated in each session. The N400 label refers to the increased negativity in the 350-ms to 550-ms time window in response to the implausible compared with the plausible condition.

social-N400 effect. In the critical correlation analyses, the size of the social-N400 effect was correlated with the AQ scores—although this effect did not survive the Bonferroni correction, $r(42) = -.31$, $p = .04$ uncorrected (see Fig. 6)—but not with the RMET or the EQ scores, rs(42) < .03, ps > .84. The relationship between the AQ scores and the size of the social-N400 effect is suggestive: Neurotypical individuals with higher autistic trait load appear to be less likely to engage in adopting the perspective of their colisteners.

General Discussion

Endowed with powerful social skills, humans can extract rich information about the mental states of other people. We asked whether comprehenders track the knowledge states of individuals who are present during a linguistic exchange but with whom they do not interact. In two ERP experiments, participants read implausible sentences ("The girl had a little beak"), preceded by spoken contexts that rendered them plausible ("The

Table 6. Average Event-Related-Potential Amplitudes (in Microvolts) in the N400 Time Window Evoked by the Target Words in the Sensibility-Judgment Task and in the Comprehension-Question Task of Experiment 2

	Sensibility-judgment task		Comprehension-question task	
Condition	Alone session	Joint session	Alone session	Joint session
Plausible	0.46 (0.34)	0.93 (0.38)	0.03 (0.30)	0.13 (0.28)
Implausible	−0.92 (0.29)	−1.32 (0.40)	−1.15 (0.32)	−1.14 (0.34)
Context dependent	0.78 (0.38)	0.07 (0.40)	0.40 (0.28)	0.44 (0.34)

Note: Standard errors of the mean by participants are provided in parentheses.

girl dressed up as a canary for Halloween"). Results showed that no semantic difficulty ensued when participants were reading the critical sentences while alone in the room, in line with prior work (Nieuwland & Van Berkum, 2006b; Van Berkum et al., 2007). However, when another individual was present for whom the critical sentences were implausible (because they had no access to the context sentence), participants showed an ERP marker of processing difficulty (the N400). Given the evidence for the automaticity of speech processing (e.g., Hugdahl, Thomsen, Ersland, Rimol, & Niemi, 2003; Scott, Gallée, & Fedorenko, 2017), it is unlikely that participants strategically ignored the context sentences when accompanied by colisteners. Thus, we argue that processing difficulty resulted because participants experienced empathetic confusion for their colisteners because they knew that the target sentence would not make sense to them.

This social-N400 effect was reported by Rueschemeyer et al. (2015; see also Westley, Kohút, & Rueschemeyer, 2017). We conceptually replicated this effect and established its robustness to changes in design (within-subjects vs. between-subjects) and materials. Critically, in addition to replicating the social N400 under the explicit instruction to the participants to adopt the confederate's perspective (Experiment 1), we found that such instructions were not needed for the social N400 to emerge. In Experiment 2, participants exhibited the social N400 when the task was to simply decide whether the target sentence makes sense. We did not find evidence of the social N400 when participants read the sentences passively, plausibly because they failed to engage deeply with the materials under those conditions. Finally, no social N400 was observed when the task was a demanding comprehension-question task, suggesting that cognitive load may limit our mentalizing capacity (Epley, Morewedge, & Keysar, 2004; Lin, Keysar, & Epley, 2010).

A number of questions remain about the nature and scope of the social-N400 effect. Is this effect limited to situations in which a colistener is physically present, or would it emerge if a colistener were present via a video conference or phone call? How does the nature of our relationship with the colistener affect the likelihood of us adopting their perspective? Does it matter if the colistener is someone whose opinion we care about? And how do these differences in our relationships with the colisteners affect the nature and dynamics of our mentalizing in situations with multiple colisteners? Results from our demanding comprehension task suggest that people have limited resources for perspective taking, so how do they distribute these

Table 7. Descriptive Statistics for the Three Tests of Social Competence in the Exploratory Analysis

Test of social competence	*M*	*SD*	Range	Number of participants with clinically significant ASD traits
Autism-Spectrum Quotient	17.39	6.29	9–36	2
"Reading the Mind in the Eyes" Test	27.66	3.54	19–33	
Empathy Quotient	43.30	12.93	22–75	10

Note: Eighty percent of individuals with a clinical diagnosis of an autism spectrum disorder (ASD) have a total score of 32 or higher on the Autism-Spectrum Quotient (compared with only 2% of individuals without an ASD diagnosis). Thus, a score of 32 or higher is considered to indicate clinically significant levels of autistic traits (Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001). Similarly, a score of 30 or lower on the Empathy Quotient is considered to indicate clinically significant levels of lack of empathy (Baron-Cohen & Wheelwright, 2004).

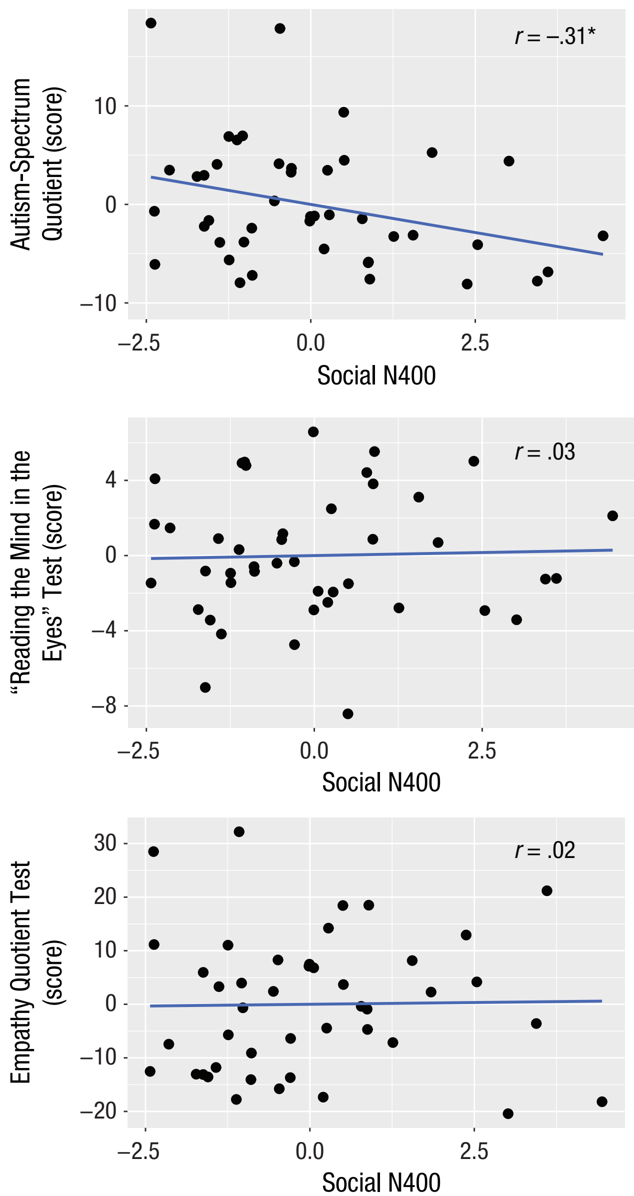

Fig. 6. Scatterplots (with best-fitting regression lines) showing correlations between the magnitude of the social-N400 effect and the behavioral score on each of three tests of social competence: the Autism-Spectrum Quotient, “Reading the Mind in the Eyes” Test, and Empathy Quotient. The social-N400 magnitudes and the behavioral scores are unstandardized residuals, controlling for the magnitude of the classic N400 effect. The asterisk indicates statistical significance at the $p < .05$, uncorrected, level (the corrected level is $p < .017$).

resources across multiple colisteners? Do people select and track one colistener at a time, or do they track multiple colisteners but in a less detailed manner?

Finally, in an exploratory component of the study, we found that social competence, measured by the AQ (Baron-Cohen, Wheelwright, Skinner, et al., 2001), explains some variance in the size of the social N400 across individuals: Individuals with higher autistic trait load showed smaller social N400s (controlling for the size of the regular N400 effect). This relationship suggests that the social-N400 effect may be reduced or absent in individuals with autism spectrum disorders, a population characterized by deficits in social interaction (Tager-Flusberg, Paul, & Lord, 2013). To the extent that the social-N400 effect proves to be stable and reliable within individuals over time, it might be a candidate marker of autism or communicative difficulties more generally.

Action Editor

Eddie Harmon-Jones served as action editor for this article.

Author Contributions

O. Jouravlev, E. Gibson, and E. Fedorenko developed the study concept. All the authors contributed to the study design. Testing and data collection were performed by O. Jouravlev, R. Schwartz, D. Ayyash, and Z. Mineroff. O. Jouravlev analyzed and interpreted the data and drafted the manuscript, and E. Gibson and E. Fedorenko provided critical revisions. All the authors approved the final manuscript for submission.

Acknowledgments

We thank Rashida Khudiyeva and Catherine Looby for help with data collection.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This research was supported by a grant from the Simons Foundation to the Simons Center for the Social Brain at Massachusetts Institute of Technology. E. Fedorenko was also supported by Award No. HD057522 from the National Institute of Child Health and Human Development and by Award No. DC016607 from the National Institute on Deafness and Other Communication Disorders, and E. Gibson was supported by Linguistics Program Award No. 1534318 from the National Science Foundation.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797618807674

Open Practices

All data and materials have been made publicly available via the Open Science Framework and can be accessed at osf.io/x7kma. The design and analysis plans for the experiments were not preregistered. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797618807674. This article has received the badges for Open Data and Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

References

Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., . . . Zuni, K. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, 253–267.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*, 37–46.

Baron-Cohen, S., & Wheelwright, S. (2004). The Empathy Quotient: An investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders*, *34*, 163–175.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry, 42*, 241–251.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, *31*, 5–17.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2018). lme4: Linear mixed-effects models using 'Eigen' and S4 (R package version 1.1-7) [Computer software]. Retrieved from https://cran.r-project.org/web/packages/lme4/index.html

Brennan, S. E., Galati, A., & Kuhlen, A. K. (2010). Two minds, one dialog: Coordinating speaking and understanding. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 53, pp. 301–344). Cambridge, MA: Academic Press.

Brown-Schmidt, S. (2009). The role of executive function in perspective taking during online language comprehension. *Psychonomic Bulletin & Review*, *16*, 893–900.

Brown-Schmidt, S., Gunlogson, C., & Tanenhaus, M. K. (2008). Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition*, *107*, 1122–1134.

Clark, H. H. (1993). *Arenas of language use*. Chicago, IL: University of Chicago Press.

Clark, H. H., & Carlson, T. B. (1982). Hearers and speech acts. *Language*, *58*, 332–373.

Curran, T., Tucker, D. M., Kutas, M., & Posner, M. I. (1993). Topography of the N400: Brain electrical activity reflecting semantic expectancy. *Electroencephalography and Clinical Neurophysiology*, *88*, 188–209.

Dawson, G., & Fernald, M. (1987). Perspective-taking ability and its relationship to the social behavior of autistic children. *Journal of Autism and Developmental Disorders*, *17*, 487–498.

Epley, N., Morewedge, C. K., & Keysar, B. (2004). Perspective taking in children and adults: Equivalent egocentrism but differential correction. *Journal of Experimental Social Psychology*, *40*, 760–768.

Farrant, B. M., Fletcher, J., & Maybery, M. T. (2006). Specific language impairment, theory of mind, and visual perspective taking: Evidence for simulation theory and the developmental role of language. *Child Development*, *77*, 1842–1853.

Ferguson, H. J., & Cane, J. E. (2015). Examining the cognitive costs of counterfactual language comprehension: Evidence from ERPs. *Brain Research*, *1622*, 252–269.

Fussell, S. R., & Krauss, R. M. (1992). Coordination of knowledge in communication: Effects of speakers' assumptions about what others know. *Journal of Personality and Social Psychology*, *62*, 378–391.

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

Hald, L. A., Steenbeek-Planting, E. G., & Hagoort, P. (2007). The interaction of discourse context and world knowledge in online sentence comprehension. Evidence from the N400. *Brain Research*, *1146*, 210–218.

Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, *49*, 43–61.

Heller, D., Grodner, D., & Tanenhaus, M. K. (2008). The role of perspective in identifying domains of reference. *Cognition*, *108*, 831–836.

Hothorn, T., Bretz, F., Westfall, P., Heiberger, R. M., Schuetzenmeister, A., Scheibe, S., & Hothorn, M. T. (2017). Package 'multcomp.' Retrieved from http://cran.statsfu.ca/web/packages/multcomp/multcomp

Hugdahl, K., Thomsen, T., Ersland, L., Rimol, L. M., & Niemi, J. (2003). The effects of attention on speech perception: An fMRI study. *Brain and Language*, *85*, 37–48.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*, 32–38.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*, 25–41.

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

Lane, L. W., & Ferreira, V. S. (2008). Speaker-external versus speaker-internal forces on utterance form: Do cognitive demands override threats to referential success? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1466–1481.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press.

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, *46*, 551–556.

Nieuwland, M. S., & Van Berkum, J. J. A. (2006a). Individual differences and contextual bias in pronoun resolution: Evidence from ERPs. *Brain Research*, *1118*, 155–167.

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

R Core Team. (2008). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rueschemeyer, S. A., Gardner, T., & Stoner, C. (2015). The social N400 effect: How the presence of other listeners affects language comprehension. *Psychonomic Bulletin & Review*, *22*, 128–134.

Ryskin, R. A., Benjamin, A. S., Tullis, J., & Brown-Schmidt, S. (2015). Perspective-taking in comprehension, production, and memory: An individual differences approach. *Journal of Experimental Psychology: General*, *144*, 898–915.

Scott, T. L., Gallée, J., & Fedorenko, E. (2017). A new fun and robust version of an fMRI localizer for the frontotemporal language system. *Cognitive Neuroscience*, *8*, 167–176.

Sebanz, N., Knoblich, G., & Prinz, W. (2003). Representing others' actions: Just like one's own? *Cognition*, *88*(3), B11–B21.

Sebanz, N., Knoblich, G., Prinz, W., & Wascher, E. (2006). Twin peaks: An ERP study of action planning and control in coacting individuals. *Journal of Cognitive Neuroscience*, *18*, 859–870.

Tager-Flusberg, H., Paul, R., & Lord, C. (2013). Language and communication in autism. In F. R. Volkmar, R. Paul, A. Klin, & D. Cohen (Eds.), *Handbook of autism and pervasive developmental disorders* (pp. 335–364). Hoboken, NJ: Wiley.

Tanner, D., & Van Hell, J. G. (2014). ERPs reveal individual differences in morphosyntactic processing. *Neuropsychologia*, *56*, 289–301.

Van Berkum, J. J. A., Koornneef, A. W., Otten, M., & Nieuwland, M. S. (2007). Establishing reference in language comprehension: An electrophysiological perspective. *Brain Research*, *1146*, 158–171.

van den Brink, D., Van Berkum, J. J. A., Bastiaansen, M. C. M., Tesink, C. M. J. Y., Kos, M., Buitelaar, J. K., & Hagoort, P. (2012). Empathy matters: ERP evidence for inter-individual differences in social language processing. *Social Cognitive and Affective Neuroscience*, *7*, 173–183.

Westley, A., Kohút, Z., & Rueschemeyer, S. A. (2017). "I know something you don't know": Discourse and social context effects on the N400 in adolescents. *Journal of Experimental Child Psychology*, *164*, 45–54.

Wilkes-Gibbs, D., & Clark, H. H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language*, *31*, 183–194.