Tracking Colisteners' Knowledge States During Language Comprehension

Tracking co-listeners' knowledge states during language comprehension

Short title: Knowledge states of co-listeners

Olessia Jouravlev[1,2], Rachael Schwarz[3], Dima Ayaash[1], Zachary Mineroff[1], Edward Gibson[1], and Evelina Fedorenko[1,4,5]

[1]MIT, [2]Carleton University, [3]Wellesley College, [4]HMS, [5]MGH

Acknowledgements: This research was supported by a grant from the Simons Foundation to the SCSB at MIT. EF was also supported by award HD057522 from NICHD, and EG by the Linguistics Program Award 1534318 from NSF. We thank Rashida Khudiyeva and Catherine Looby for help with data collection.

Address for correspondence:
Olessia Jouravlev, olessiaj@mit.edu
Massachusetts Institute of Technology
Brain & Cognitive Sciences Department
43 Vassar Street, Building 46, Room 3037
Cambridge, MA 02139

Abstract

When we receive information in the presence of others, are we sensitive to what they do or do not understand? In two ERP 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) ensued when they read the sentences alone. However, when a confederate was present who did not receive the contexts so that the critical sentences were implausible for them, participants exhibited processing difficulty: the "Social N400" effect. This effect obtained when participants were instructed to adopt the confederate's perspective, and most 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 they directly interact with but also those merely present during the linguistic exchange.

Keywords: communication, perspective taking, joint actions, social cognition, ERPs, N400

1. Introduction

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

A priori, we might hypothesize that comprehenders do not model the minds of those around them except when directly interacting with them. After all, mentalizing is costly. Indeed, some 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 (Lane & Ferreira, 2008; Keysar et al., 2000, 2003). 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).

Some evidence suggesting that people represent mental states of all physically present individuals comes from studies of non-linguistic actions. Individuals performing tasks alongside each other appear to track task requirements and action alternatives of

others, even when this compromises performance (Sebanz et al., 2003, 2006). In the language domain, Wilkes-Gibbs & Clark (1992) used a referential communication task – where a speaker (the Director) gives a listener (the Matcher) instructions for rearranging images of non-nameable objects – and showed that the Director assumes that a passive co-listener has established the same common ground with him/her as the actively participating Matcher. When the co-listener later became the Matcher, the Director kept using the names that were established in communication with the original Matcher.

More recently, Rueschemeyer et al. (2015) used ERPs to ask whether comprehenders are sensitive to the knowledge states of their co-listeners. 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*; (1c)), along with control plausible sentences in supportive contexts (1a), and implausible sentences where the context did not make them plausible (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 (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 (1a) and (1b): the word *gills* elicited a larger N400 in the latter condition (Kutas &

Hilliyard, 1980). In the critical condition (1c), when participants were alone, they experienced no processing difficulty at *gills*, in line with prior work (Nieuwland & Van Berkum, 2006b; Van Berkum et al., 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 co-listeners and thus experience empathetic confusion, and termed it the *Social N400*.

The Social N400 effect is a promising implicit marker of representing others' minds. However, Rueschemeyer et al. explicitly instructed participants to adopt the confederate's perspective. It is therefore unclear whether the effect would obtain without explicit instruction.

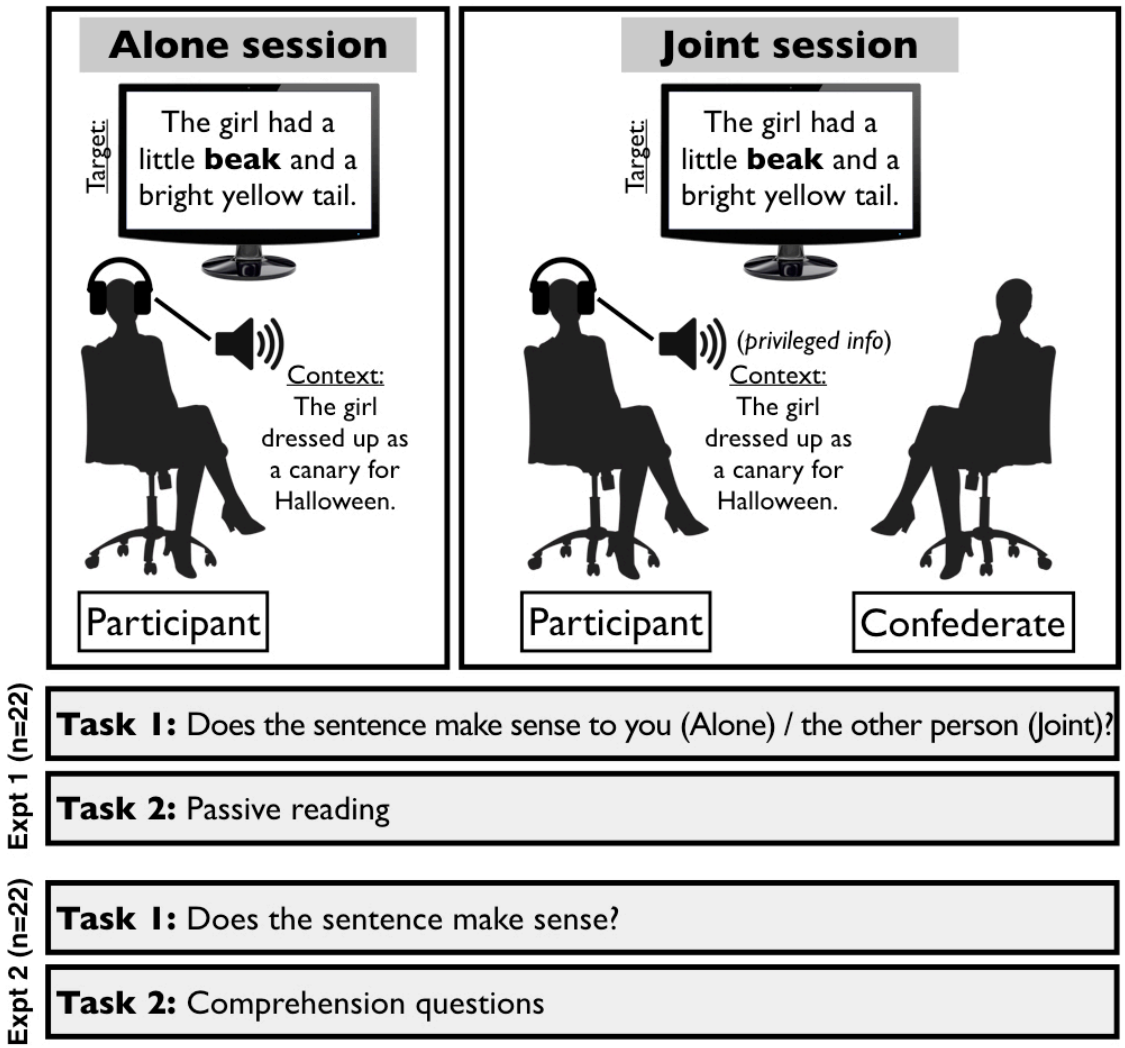
To illuminate the conditions under which we model the knowledge states of co-listeners, across two ERP experiments we examined four task conditions: a) explicit instructions to consider the confederate's perspective (as in Rueschemeyer et al.'s study, in line with current emphasis on replication; Aarts et al., 2015), b) a sensibility judgment task that did not ask participants 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 obtains 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 et al., 2006), executive abilities (Brown-Schmidt, 2009; Ryskin et al., 2015), and/or social competence (Baron-Cohen et al, 1985; Dawson & Fernald, 1987). We here focused on social competence and tested whether better social skills are associated with better perspective taking.

2. Experiment 1

Participants performed two comprehension tasks while 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 (Figure 1). For each of the tasks, participants performed two sessions: one where they were alone (*Alone* sessions), and another where a confederate was present (*Joint* sessions). 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 2.1.2. 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 between the Alone and Joint sessions.

Figure 1. *The experimental setup in Experiments 1 and 2.*



The first goal of the experiment was to test the robustness of the Social N400 (Aarts et al., 2015). Rueschemeyer et al. (2015) used a between-subjects design with different participants in the Alone vs. 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 et al., 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 replicate 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 obtains without the explicit instruction to adopt the confederate's perspective.

2.1. Methods

2.1.1. Participants: Twenty-four participants (12 males; *M*(age) = 24.8, *SD* = 3.9, range 19-32 years) from MIT and the surrounding Boston community participated for payment. The sample size was determined based on 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, psychiatric disorders, or any 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 two participants were excluded (one due to technical errors that resulted in data loss, and one due to an excessive number of artifacts in the EEG signal, with more that 25% of trials affected), leaving 22 participants for the analysis.

2.1.2. Materials: One hundred items, exemplified in (2)-(4), were constructed with three conditions each: Plausible (2a, 3a, 4a), Implausible (2b, 3b, 4b), and Context-dependent (2c, 3c, 4c).

(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 (*M*(length) = 10 words, range: 4-19) varied across the three conditions and served to establish the appropriate discourse context. The second, critical, sentence (*M*(length) = 11 words, range: 5-17) was identical between the implausible and context-dependent conditions, and minimally different (in one word) from the plausible condition. The target word was embedded in the second sentence. Its position varied between word 3 and 12, and it never appeared in the

sentence-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 (see https://osf.io/fnt6v/ for the full set of materials).

Prior to the ERP study, the materials were normed in two sentence completion studies. In the first study, participants were presented with 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.com'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 trial order randomized 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 (Table 1). In particular, the target word was highly expected in the plausible condition, either with or without the first context sentence (cloze probabilities:

0.57 and 0.56, respectively), and highly unexpected in the implausible condition, either with or without the first context sentence (cloze probabilities: 0.01 in both studies). Importantly, in the context-dependent condition, the target word was quite expected when the context sentence was included (cloze probability: 0.35), but not when only the second sentence was included (close probability: 0.01).

Table 1. *Cloze probability values for the target words in the context of (a) the first and second sentence (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___).*

	Sentence Fragments		
Conditions	First & Second Sentences	Second Sentence Only	t-test
Plausible	0.57	0.56	$t(198)=0.22, p=.82$
Implausible	0.01	0.01	$t(198)=0.76, p=.45$
Context-dependent	0.35	0.01	**$t(198)=13.86, p<.001$**

The context sentences were recorded by a female native speaker of English, for auditory presentation. Each recording lasted for a maximum of 4 sec, with shorter sentences padded with silence at the end.

2.1.3. 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 where they are in the room by themselves, and one where the other participant is in the room with them). They were fitted with the EEG cap and headphones, and instructed 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), with the confederate joining for two of the sessions, as detailed below.

The 300 trials were distributed across four experimental lists following a Latin

Square design, so that each list consisted of 75 trials and contained only one version of an item (plausible, implausible, or context-dependent), with 25 trials per condition. Each participant saw all four lists across the four task/session combinations (sensibility judgment / Alone; sensibility judgment / Joint; passive reading / Alone; and passive reading / Joint). The i) pairing between lists and task/session combinations and ii) task/session order varied across participants. The order of trials within each list was randomized 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 the rate of 450 ms per word. Each word was followed by a 100 ms inter-stimulus 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, an image of a finger pressing a button was presented 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 minutes, 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 questionnaire (ASQ; 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 questionnaire (EQ; Baron-Cohen & Wheelwright, 2004). The entire experiment took approximately 2 hours.

2.1.4. 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 to be used as reference channels. The Active Two Biosemi system with active Ag-AgCl electrodes mounted on an elastic cap (Electro-Cap Inc.) was used. All channels were referenced offline to an average of the mastoid channels. EEG activity was recorded at a sampling rate of 512 Hz. Following standard procedures in ERP research, the signal was then filtered offline (bandpass 0.1-40 Hz), and trials with blinks, eye movements, muscle artifacts, and skin potentials were excluded prior to averaging and analyses. Across participants, an average of 7.7 % of trials (SD = 4.7; range 1.3-14.7) were excluded.

2.1.5. 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 times (RTs) – to a linear mixed-effects model (lmer) performed with the lme4 package (Bates, Maechler, Bolker, & Walker, 2014) in R. 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 to

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.

2.1.6. EEG/ERP analyses: Continuous EEG signal was divided into epochs over a window from 200 ms prior to the target word onset to 800 ms post onset. The 200 ms window prior to the target word onset was used as the pre-stimulus baseline. To obtain event-related potentials (ERPs), epochs were averaged across trials within a condition for each target electrode (see below) and participant. For visualization purposes, the responses were further averaged across participants (Figures 2-3).

The ERP component of interest was the N400 (Kutas & Hillyard, 1980), a negative deflection observed at centro-parietal locations on the scalp 300-600 ms post 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 time-window of interest for analysis (350-550 ms post word onset). The amplitudes within this time-window were averaged for each condition, session, electrode, and participant, and used as dependent measures in the repeated measures ANOVAs. We used ANOVAs rather than linear mixed-effects models to analyze the ERP data (i) to make the results comparable with Rueschemeyer et al. study (2015), and (ii) because single-trial-level data were not readily available.

Following Rueschemeyer et al. (2015), for each of the tasks (sensibility judgment and passive reading), we first conducted a 2 x 3 x 8 ANOVA, with session (Alone vs. Joint), condition (Plausible vs. Implausible vs. Context-dependent), and electrode (C3, Cz, C4, CP1, CP2, P3, Pz, P4) as within-subject 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 vs. Implausible vs. 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).

2.2. Results

2.2.1. Behavioral Results: Average proportions of yes responses and RTs in the sensibility judgment task are reported in Table 2.

Table 2. *Average proportions of yes responses (% Yes) and reaction times (RTs, in ms) in the sensibility judgment task in Experiment 1 as a function of the two experimental manipulations (session and condition). Standard errors of the mean by participants are provided in parentheses.*

	Session			
	Alone		Joint	
Condition	% Yes	RTs	% Yes	RTs
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)

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 (responses: χ^2 (2) = 221.98, $p < 0.001$; RTs: χ^2 (2) = 10.94, $p = 0.004$). Planned comparisons revealed that during the Alone session, proportions of yes responses differed across all three condition pairs (Plausible vs. Context-dependent: $z = 5.16$, $p = 0.001$, *corrected* here and elsewhere; Plausible vs. Implausible: $z = 18.04$, $p < 0.001$; Context-dependent vs. Implausible: $z = 17.85$, $p < 0.001$), 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 (.94) than the Implausible (.22; $z =$

18.95, $p < 0.001$) or the Context-dependent condition (.27; $z = 18.01$, $p < 0.001$). The latter two conditions did not differ significantly ($z = 2.37$, $p = 0.06$). 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 reaction times, during the Alone session participants took longer to decide on the sensibility of sentences in the Implausible (924 ms) than the Plausible (746 ms; $z = 7.24$, $p < 0.001$) or the Context-dependent condition (789 ms; $z = 5.69$, $p < 0.001$). The latter two conditions did not differ significantly ($z = 1.55$, $p = 0.32$). During the Joint session, RTs were significantly longer in the Implausible (903 ms) and Context-dependent conditions (855 ms) than the Plausible condition (740 ms; Implausible vs. Plausible: $z = 7.51$, $p < 0.001$; Context-dependent vs. Plausible: $z = 5.72$, $p < 0.001$). The two former conditions did not differ significantly ($z = 1.78$, $p = 0.22$). Thus, there was a processing cost for sentences in the Context-dependent condition during the Joint session. The data are available from https://osf.io/fnt6v/.

2.2.2. 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 Figures 2 (for the sensibility judgment task) and 3 (for the passive reading task). Mean ERP amplitudes in the N400 time-window are provided in Table 3.

Figure 2. *ERP waveforms evoked by the target words in the Plausible (black solid line), Implausible (red dashed line), and Context-dependent (blue dotted line) conditions in the sensibility judgment task in Experiment 1 (Does the sentence make sense to you / the other person?) during the Alone (top) vs. Joint (bottom) sessions. Here and in Figures 3-5, the x-axis shows time in ms, and the y-axis – ERP amplitudes in μV.*

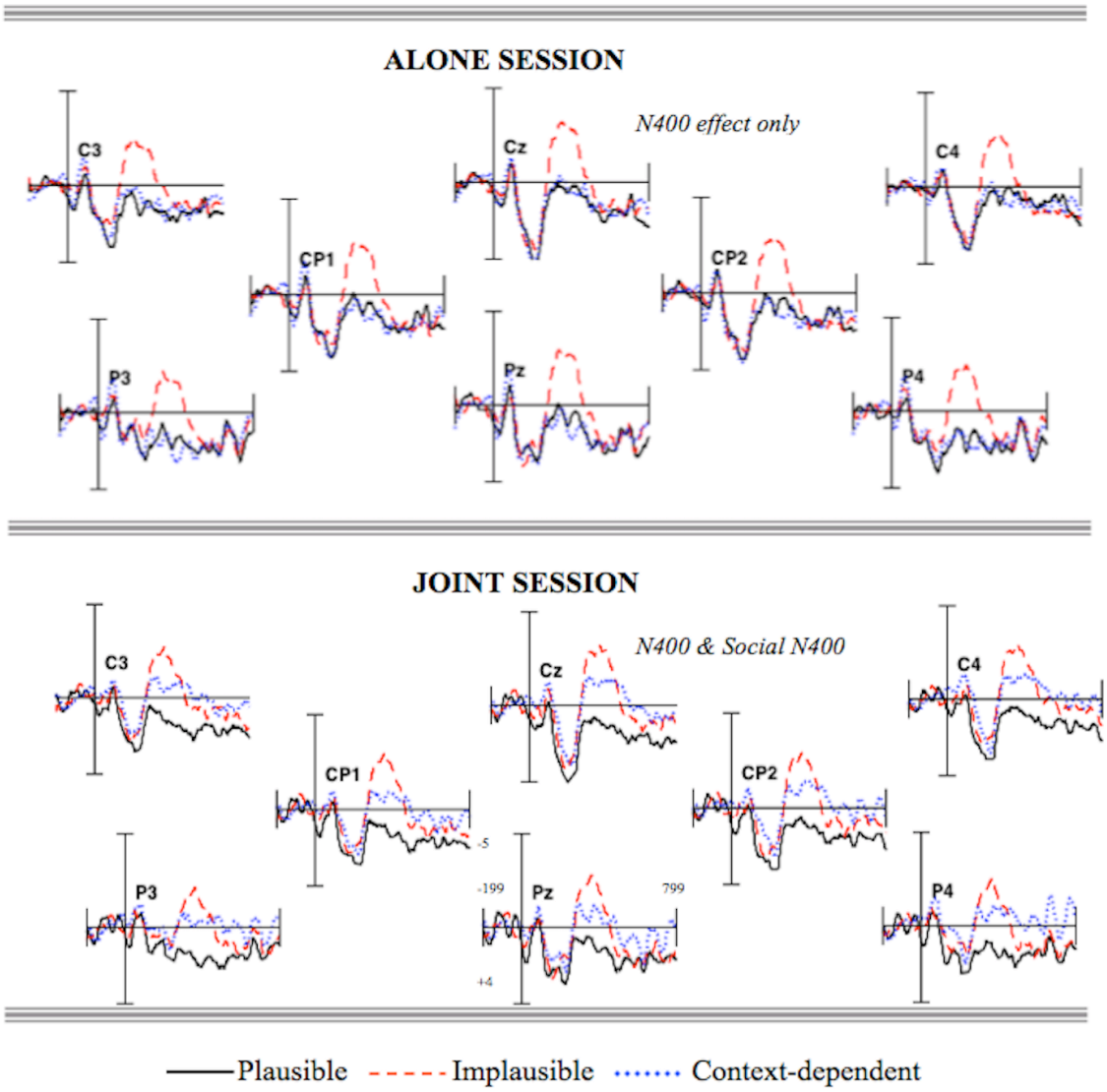


Figure 3. *ERP waveforms evoked by the target words in the Plausible (black solid line), Implausible (red dashed line), and Context-dependent (blue dotted line) conditions in the passive reading task in Experiment 1 in the Alone (top) vs. Joint (bottom) sessions.*

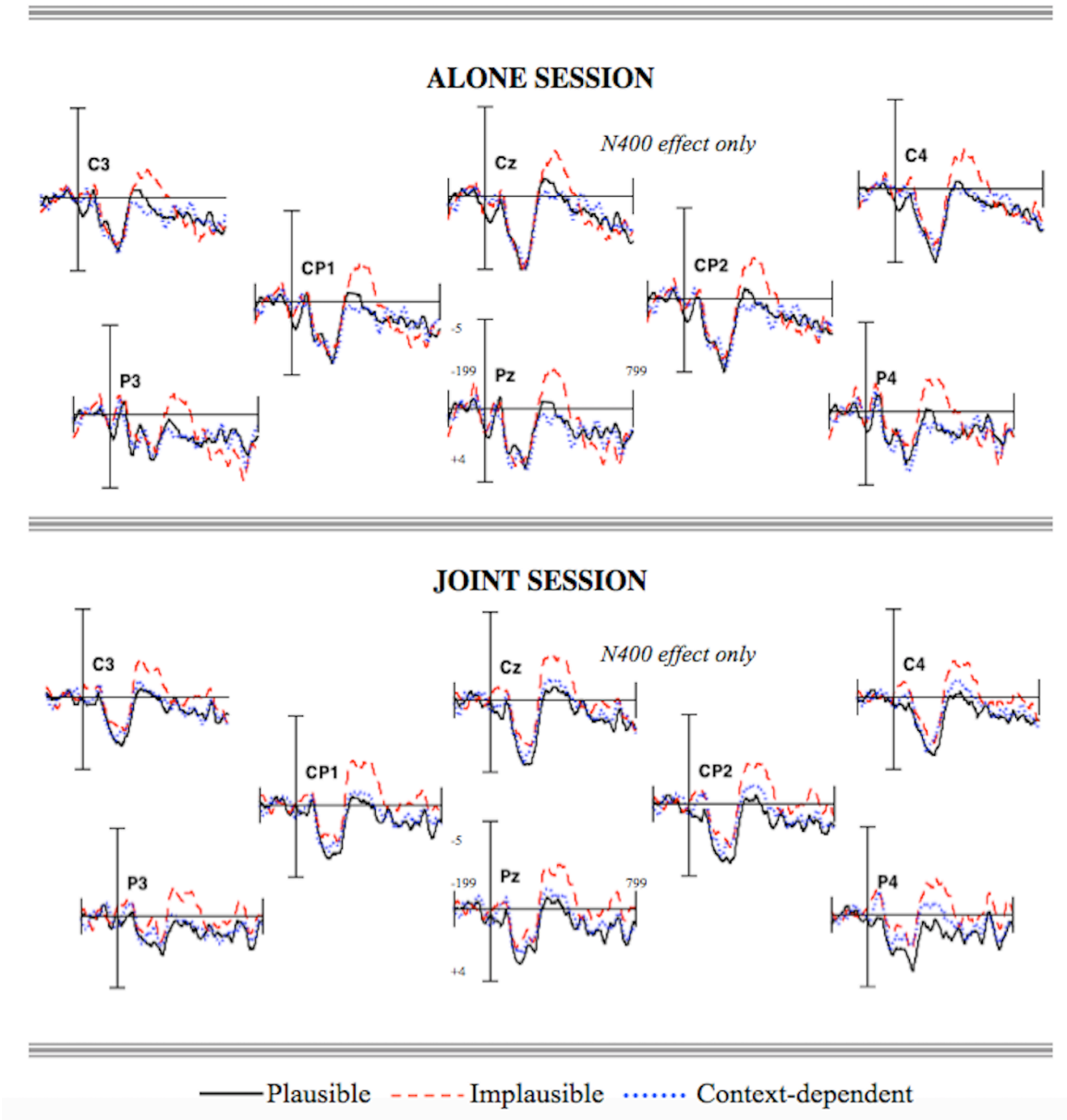


Table 3. Average ERP magnitudes (in microvolts) in the N400 time-window evoked by the target words in the sensibility judgment task and in the passive reading task of Experiment 1. Standard errors of the mean by participants are provided in parentheses.

	Task			
	Sensibility Judgment		Passive Reading	
Condition	Alone	Joint	Alone	Joint
Plausible	0.87 (.41)	1.22 (.37)	0.38 (.52)	0.27 (.49)
Implausible	-1.60 (.40)	-1.47 (.40)	-1.07 (.36)	-1.43 (.29)
Context-dependent	1.00 (.08)	-0.77 (.36)	0.72 (.39)	- 0.21 (.51)

For the sensibility judgment task, we observed a main effect of condition ($F(2,42) = 26.93, p < 0.001, \eta^2 = 0.56$), with ERPs being significantly more negative in the Implausible than the Plausible (-1.54 vs. 1.05, $t(21) = 5.29, p < .001$) or Context-dependent condition (-1.54 vs. 0.12, $t(21) = 3.39, p = .004$). The latter two conditions did not differ significantly ($t(21) = 1.91, p = .18$). Critically, we observed a significant interaction between condition and session ($F(2,42) = 7.57, p = 0.002, \eta^2 = 0.27$). Planned comparisons revealed that during the Alone session, the magnitude of the N400 was reduced in the Plausible (0.87) and Context-dependent condition (1.00) compared to the Implausible condition (-1.6; Plausible vs. Implausible: $t(21) = 4.24, p < 0.001$; Context-dependent vs. Implausible: $t(21) = 4.01, p < 0.001$). The Plausible and Context-dependent conditions did not differ significantly ($t(21) = 0.21, p = 0.99$). The fact that the Context-dependent condition patterned with the Plausible condition is in line with prior work that had established that contextual information can alleviate 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 magnitude of the N400 in the Plausible (1.22), but not the Context-dependent condition (-0.77), was significantly reduced compared to the Implausible condition (-1.47; Plausible vs. Implausible: $t(21) = 5.03, p < 0.001$; Context-dependent vs. Implausible: $t(21) = 1.30, p = 0.58$). ERPs in the Context-dependent condition were significantly more negative than in the Plausible condition ($t(21) = 3.72, p = 0.001$). 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 ($F(2,42) = 14.44, p < 0.001, \eta^2 = 0.41$), with ERPs being significantly more negative in the Implausible than the Plausible (-1.25 vs. 0.32, $t(21) = 3.15, p = 0.008$) or the Context-dependent condition (-1.25 vs. 0.26, $t(21) = 3.02, p = 0.01$). The latter two conditions did not differ significantly ($t(21) = 0.12, p = 0.99$). We found no evidence of a significant interaction between condition and session ($F(2,42) = 0.72, p = 0.49, \eta^2 = 0.03$). The data for this experiment and for Experiment 2 are available from https://osf.io/fnt6v/.

3. Experiment 2

In the sensibility judgment task of Experiment 1, where 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 obtains. 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.

3.1. Methods

3.1.1. Participants: Twenty-three participants (10 males; $M(\text{age}) = 26.1, SD = 5.4$, range 20-40 years) from MIT and the surrounding Boston community participated for payment. All were right-handed (by self report) native speakers 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 one participant were excluded due to an excessive number of artifacts in the EEG signal, with more than 25% of trials affected, leaving 22 participants for the analysis.

3.1.2. 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 trial "*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 trial "*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 from https://osf.io/fnt6v/.)

3.1.3. Procedure: 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 hours.

3.1.4. EEG recording: The 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) were excluded due to the presence of artifacts.

3.1.5. Behavioral and EEG/ERP analyses: 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.

3.2. Results

3.2.1. Behavioral Results: Average proportions of yes responses and RTs in the sensibility judgment task are reported in Table 4, and average accuracies and RTs in the comprehension question task are reported in Table 5.

Table 4. *Average proportions of yes-responses (% Yes) and response times (RTs, in ms) in the sensibility judgment task in Experiment 2. Standard errors of the mean by participants are provided in parentheses.*

	Session			
	Alone		Joint	
Condition	% Yes	RTs	% Yes	RTs
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)

Table 5. *Average accuracies (proportion correct) and reaction times (RTs, in ms) in the comprehension question task in Experiment 2. Standard errors of the mean by participants are provided in parentheses.*

	Session

	Alone		Joint	
Condition	Accuracy	RTs	Accuracy	RTs
Plausible	.96 (.04)	1780 (109)	.95 (.05)	1791 (115)
Implausible	.91 (.06)	1745 (114)	.92 (.07)	1773 (109)
Context-dependent	.91 (.06)	1774 (112)	.92 (.07)	1820 (112)

In the sensibility judgment task, similar to 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, it was only present in the response data (χ^2 (2) = 14.54, $p < .001$), but not RTs (χ^2 (2) = 1.73, p = 0.42). Planned comparisons revealed that during the Alone session, proportions of yes responses were higher in the Plausible (.95) and Context-dependent conditions (.94) than the Implausible condition (.22; Plausible vs. Implausible: $z = 17.3, p < 0.001$; Context-dependent vs. Implausible: $z = 17.05, p < .001$). The Plausible and the Context-dependent conditions did not differ significantly ($z = 1.04, p = 0.64$). 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 Joint session, proportions of yes responses differed across all three condition pairs (Plausible vs. Context-dependent: $z = 5.41, p < 0.001$; Plausible vs. Implausible: $z = 17.21, p < 0.001$; Context-dependent vs. Implausible: $z = 16.29, p < 0.001$), with the largest 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 at least some of the time when deciding whether the sentence makes 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) either in the accuracies

(χ^2 (2) = 0.69, p = .71) or the RTs (χ^2 (2) = 0.63, p = .73).

3.2.2. 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 Figures 4 (for the sensibility judgment task) and 5 (for the comprehension question task). Mean ERP amplitudes in the N400 time-window are provided in Table 6.

Figure 4. *ERP waveforms evoked by the target words in the Plausible (black solid line), Implausible (red dashed line), and Context-dependent (blue dotted line) conditions in the sensibility judgment task in Experiment 2 (Does the sentence make sense?) during the Alone (top) vs. Joint (bottom) sessions.*

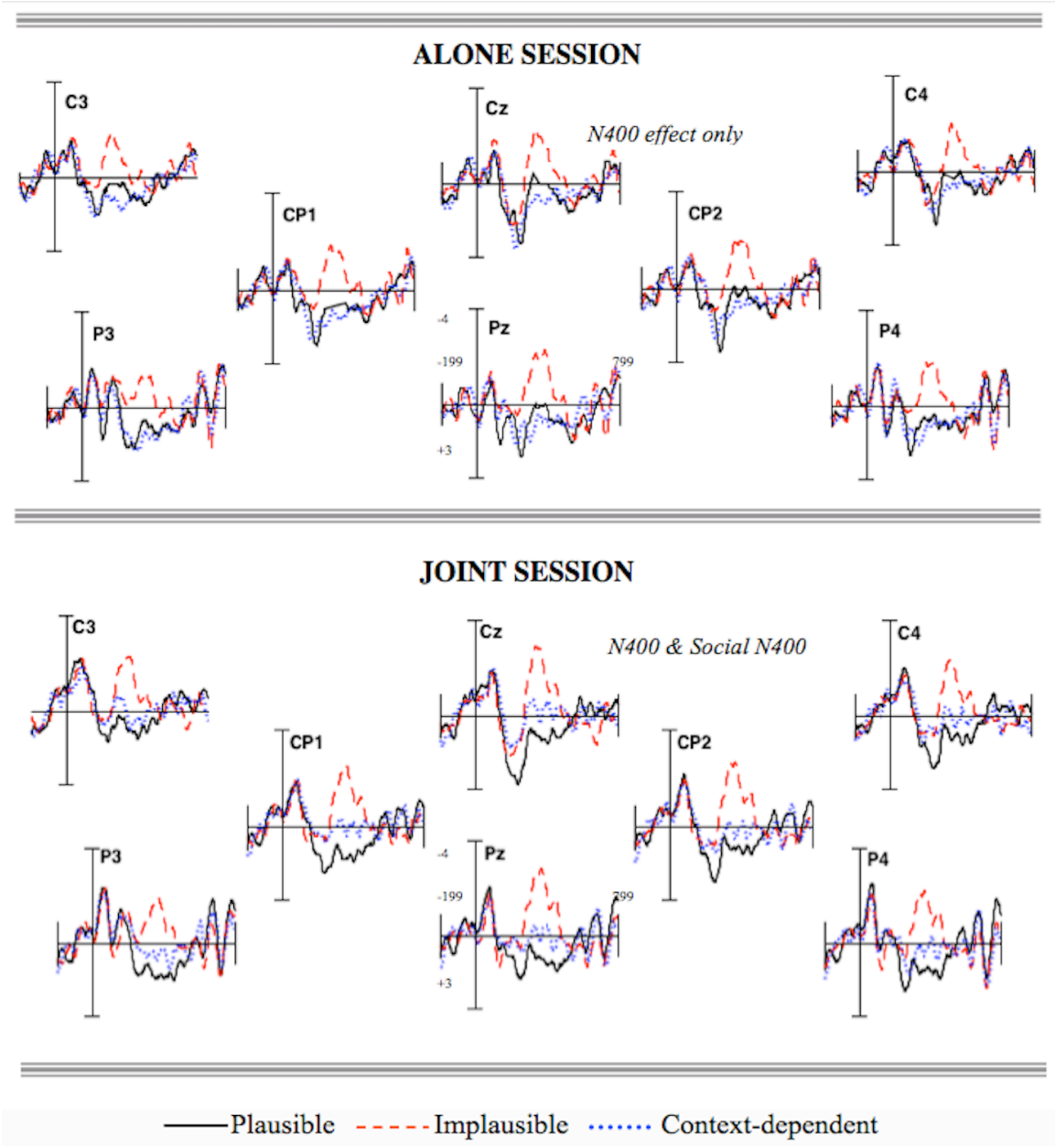


Figure 5. *ERP waveforms evoked by the target words in the Plausible (black solid line), Implausible (red dashed line), and Context-dependent (blue dotted line) conditions in the comprehension question task in Experiment 2 in the Alone (top) vs. Joint (bottom) sessions.*

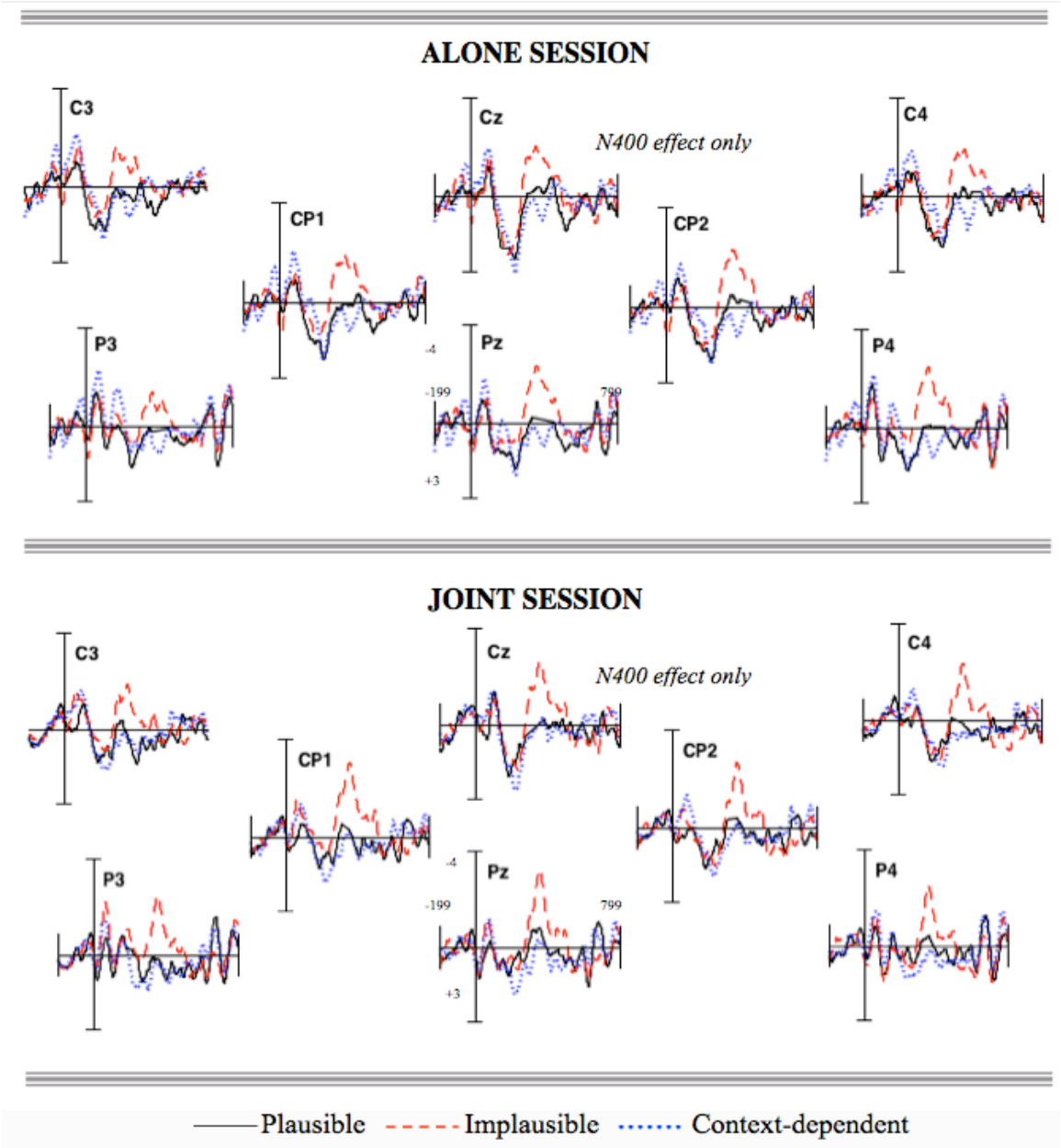


Table 6. Average ERP magnitudes (in microvolts) in the N400 time-window evoked by the target words in the sensibility judgment task and in the comprehension question task of Experiment 2. Standard errors of the mean by participants are provided in parenthesis.

	Task

Condition	Sensibility Judgment		Comprehension Question	
	Alone	Joint	Alone	Joint
Plausible	0.46 (.34)	0.93 (.38)	0.03 (.30)	0.13 (.28)
Implausible	-0.92 (.29)	-1.32 (.40)	-1.15 (.32)	-1.14 (.34)
Context-dependent	0.78 (.38)	0.07 (.40)	0.40 (.28)	0.44 (.34)

For the sensibility judgment task, we observed a main effect of condition ($F(2,42) = 20.02$, $p < 0.001$, $\eta^2 = 0.49$), with ERPs being significantly more negative in the Implausible than the Plausible (-1.12 vs. 0.70, $t(21) = 4.04$, $p < 0.001$) or Context-dependent condition (-1.12 vs. 0.43, $t(21) = 3.45$, $p = 0.003$). The latter two conditions did not differ significantly ($t(21) = 0.59$, $p = 0.99$). Critically, we observed a marginally significant interaction between condition and session ($F(2,42) = 3.33$, $p = 0.05$, $\eta^2 = 0.14$). Planned comparisons revealed that during the Alone session, the magnitude of the N400 was reduced in the Plausible (0.46) and Context-dependent conditions (0.78) compared to the Implausible condition (-0.92; Plausible vs. Implausible: $t(21) = 3.36$, $p = 0.01$; Context-dependent vs. Implausible: $t(21) = 3.95$, $p = 0.003$). The Plausible and Context-dependent conditions did not differ significantly ($t(21) = 0.81$, $p = 0.81$). Thus, similar to 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 magnitude of the N400 in the Plausible (0.93) and Context-dependent conditions (0.07) was significantly reduced compared to the Implausible condition (-1.32; Plausible vs. Implausible: $t(21) = 4.86$, $p < 0.001$; Context-dependent vs. Implausible: $t(21) = 3.85$, $p = 0.003$). Further, ERPs in the Context-dependent condition were significantly more negative than in the Plausible condition ($t(21) = 2.98$, $p = 0.02$). 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

($F(2,42) = 12.48, p < 0.001, \eta^2 = 0.37$), with ERPs being significantly more negative in the Implausible than the Plausible (-1.14 vs. 0.08, $t(21) = 3.35, p = 0.004$) or the Context-dependent condition (-1.14 vs. 0.42, $t(21) = 4.29, p < 0.001$). The latter two conditions did not differ significantly ($t(21) = 0.93, p = 0.74$). The interaction between condition and session was not significant ($F(2,42) = 0.02, p = 0.98, \eta^2 = 0.01$).

4. An exploratory analysis: The effect of social competence on perspective taking.

4.1. Methods and Analyses

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, for a total of 44 participants. For each participant, we computed the average magnitude of the Social N400 effect (Plausible minus Context-dependent), and the average magnitude of the classic N400 effect (Plausible minus Implausible). We next performed three regressions predicting the size of the Social N400 from each of the behavioral measures (the ASQ, Baron-Cohen et al., 2001; the RMET, Baron-Cohen 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).

4.2. Results

Descriptive statistics for the three tests of social competence (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 Social N400 effect. In the critical correlation analyses, the size of the Social N400 effect was correlated with the ASQ scores, although this effect did not survive the Bonferroni correction ($r(42) = -0.31, p = 0.04$ *uncorrected*; Figure 6), but not with the RMET or the EQ scores (rs$(42) < 0.03$, ps > 0.84). The relationship between the ASQ 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

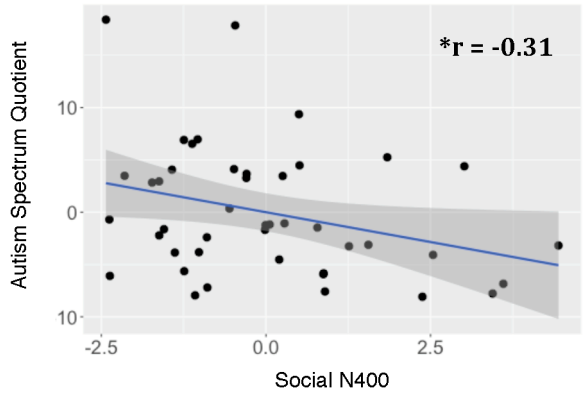
in adopting the perspective of their co-listeners.

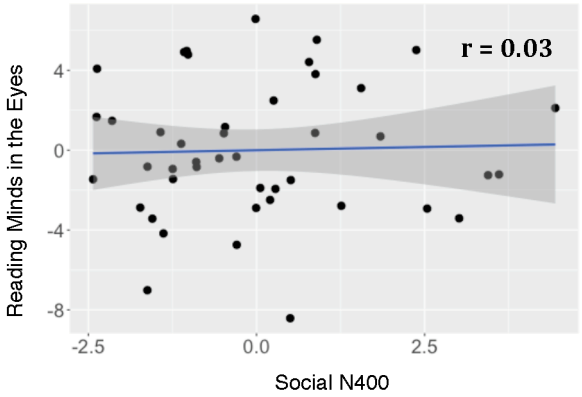
Table 7. Descriptive statistics for the three tests of social competence.

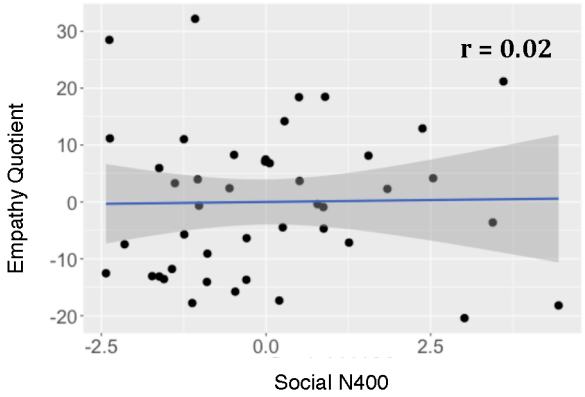
Tests of Social Competence	Mean (SD)	Range	N of participants with clinically significant ASD traits
Autism Spectrum Quotient (ASQ)	17.39 (6.29)	9 - 36	2
Reading the Mind in the Eyes (RMET)	27.66 (3.54)	19 - 33	
Empathy Quotient (EQ)	43.30 (12.93)	22 - 75	10

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 ASQ questionnaire (cf. 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 et al., 2001). Similarly, a score of 30 or lower on the EQ questionnaire is considered to indicate clinically significant levels of lack of empathy (Baron-Cohen & Wheelwright, 2004).

Figure 6. Correlations between the magnitude of the Social N400 effect and behavioral scores on the ASQ, RMET, and EQ tests of social competence. The Social N400 magnitude values and the behavioral scores are unstandardized residuals, controlling for the magnitude of the classic N400 effect. An asterisk before the *r* value indicates statistical significance at the $p < 0.05$, uncorrected, level (the corrected level is $p < 0.017$).







5. General Discussion

Endowed with powerful social skills, humans can extract rich information about others' mental states. 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 girl dressed up as a canary for Halloween*). In line with prior work (Nieuwland & Van Berkum, 2006b; Van Berkum et al., 2007), no semantic difficulty ensued when participants were reading the critical sentences alone. 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 (N400). Given the evidence for the automaticity of speech processing (e.g., Hugdahl et al., 2003; Scott et al., 2017), it is unlikely that participants strategically ignored the context sentences when accompanied by co-listeners. Thus, we argue that processing difficulty resulted because participants experienced empathetic confusion for their co-listener 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 et al., 2017). We conceptually replicated this effect and established its robustness to changes in design (within- 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 (Lin et al., 2010; Epley et al., 2004).

A number of questions remain about the nature and scope of the Social N400 effect. Is this effect limited to situations where a co-listener is *physically* present, or would it emerge if a co-listener is present via a video-conference or phone call? How does the *nature of our relationship* with the co-listener affect the likelihood of us adopting their perspective? Does it matter if the co-listener is someone whose opinion we care about? And how do these differences in our relationships with the co-listeners affect the nature and dynamics of our mentalizing in situations with multiple co-listeners? Results from our demanding comprehension task suggest that we have limited resources for perspective taking, so how do we distribute these resources across multiple co-listeners? Do we select and track one co-listener at a time, or do we track multiple co-listeners but in a less detailed manner?

Finally, in an exploratory component of the study, we found that social competence, measured by the ASQ (Baron-Cohen 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 et al., 2013). To the extent that the Social N400 effect proves to be stable and reliable within individuals over time, it might be a candidate neural marker of autism and communicative difficulties more generally.

References

1. 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*(6251), 253–267.
2. 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(2), 163-175.
3. 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. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *42*(2), 241–251.
4. Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism Spectrum Quotient□: Evidence from Asperger syndrome/high functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, *31*(1), 5–17.
5. Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7, http://CRAN.R-project.org/package=lme4. *R Package Version*.
6. Brennan, S. E., Galati, A., & Kuhlen, A. K. (2010). *Two Minds, One Dialog: Coordinating Speaking and Understanding*. *Psychology of Learning and Motivation* (Vol. 53).
7. Brown-Schmidt, S. (2009). The role of executive function in perspective taking

during online language comprehension. *Psychonomic Bulletin & Review*, *16*(5), 893–900.

8. Brown-Schmidt, S., Gunlogson, C., & Tanenhaus, M. K. (2008). Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition*, *107*(3), 1122–1134.
9. Clark, H. H. (1992). Arenas of Language Use. *Arenas of Language Use*, *18*(6), xviii, 419.
10. Clark, H. H., & Carlson, T. B. (1982). Hearers and speech acts. *Language*, *58*(2), 332–373.
11. 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/ Evoked Potentials*, *88*(3), 188–209.
12. 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*(4), 487–498.
13. 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*(6), 760–768.
14. 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*(6), 1842–1853.

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

16. 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*(3), 378–391.

17. Hagoort, P. (2003). Interplay between syntax and semantics during sentence comprehension: ERP effects of combining syntactic and semantic violations. *Journal of Cognitive Neuroscience*, *15*(6), 883–899.

18. 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*(1), 210–218.

19. 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*(1), 43–61.

20. Heller, D., Grodner, D., & Tanenhaus, M. K. (2008). The role of perspective in identifying domains of reference. *Cognition*, *108*(3), 831–836.

21. 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(1), 37-48.

22. 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*(1), 32–38.

23. Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*(1), 25–41.

24. Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science, 207*(4427), 203–205.

25. 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*(6), 1466–1481.

26. Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. MIT press.

27. 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*(3), 551–556.

28. Nieuwland, M. S., & Van Berkum, J. J. A. (2006a). Individual differences and contextual bias in pronoun resolution: Evidence from ERPs. *Brain Research*, *1118*(1), 155–167.

29. 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*(7), 1098–1111.

30. 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(1), 128-134.

31. 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*(5), 898–915.

32. Sebanz, N., Knoblich, G., & Prinz, W. (2003). Representing others' actions: Just like one's own? *Cognition*, *88*(3), B11-B21.
33. 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*(5), 859–870.
34. 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(3), 167-176.
35. Tager-Flusberg, H., Paul, R., & Lord, C. (2013). Language and Communication in Autism. In *Handbook of Autism and Pervasive Developmental Disorders,* pp. 335–364.
36. Tanner, D., & Van Hell, J. G. (2014). ERPs reveal individual differences in morphosyntactic processing. *Neuropsychologia*, *56*(1), 289–301.
37. 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*(1), 158–171.
38. 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*(2), 173–183.

39. 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.

40. Wilkes-Gibbs, D., & Clark, H. H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language*, *31*(2), 183–194.

SUPPLEMENTARY MATERIALS

Peak-to-peak analysis

To ensure that the difference in the N400 time-window is not driven by differences emerging due to earlier ERP effects (e.g., P200), we performed a peak-to-peak analysis. In particular, for each participant and each condition, we identified the most positive amplitude in the P300 time-window and the most negative amplitude in the N400 time-window. Next, we subtracted the peak N400 amplitude from the peak P300 amplitude for each condition and used these as the dependent variable in the ANOVAs.

In Experiment 1, the results of the peak-to-peak analysis fully replicated the results observed in the mean amplitude analysis. In the sensibility judgment task, there was a significant interaction between condition (Plausible vs. Implausible vs. Context-dependent) and session (Alone vs. Joint): $F(2,42) = 3.23$, $p = .05$. In the passive reading task, the interaction was not significant: $F(2,42) = 0.05$, $p = .94$. In Experiment 2, in the sensibility judgment task the interaction between condition and session was marginally significant: $F(2,42) = 2.93$, $p = .06$. However, there was a significant condition by session by electrode interaction ($F(14,294) = 2.39$, $p = .03$), such that the condition by session interaction tended to be strongest over the central electrode sites. In the comprehension question task, the condition by session interaction was not significant: $F(2,42) = 0.33$, $p = .64$. Overall, the results of the peak-to-peak analysis replicated the results of the mean amplitude analysis, suggesting that the observed patterns of results are unlikely to be caused by ERP components that emerged before the N400 time-window.