Listeners' eyes reveal spontaneous sensitivity to others' perspectives

Heather J Ferguson [1]

Richard Breheny [2]

[1] University of Kent, England, UK

[2] University College London, England, UK

Correspondence to:

Heather Ferguson
School of Psychology
University of Kent
Keynes College
Canterbury, Kent
CT2 7NP, UK

Email: H.Ferguson@kent.ac.uk
Phone: +44 (0)1227 827120
Fax: +44 (0)1227 827030

Word count: 4996 (not including Abstract and References)

Author Notes.

This work was carried out with the support of a grant from the Arts and Humanities Research Council (Ref: AH/E002358/1), and the Centre for the Study of Mind in Nature (Oslo), 'Linguistic Agency' project..

Abstract

During everyday social interactions, we typically anticipate (or explain) others' behaviour according to their current mental states (e.g. their knowledge, beliefs and intentions). To date, very little is known about the time-course with which such perspective information influences communication. We report a novel interactive 'visual world' study examining these processes. Here, two communicators watched videos depicting transfer events and subsequently described these events to each other. Critically, on half the trials a screen blocked the speakers' (but not the listeners') view part-way through the video, establishing a discrepancy in the knowledge held by the two communicators. Eye-tracking analyses showed that listeners were rapidly sensitive to their partner's perspective, as evidenced by a significantly reduced reality-bias when speakers held out-of-date knowledge about a privileged transfer event. However, we also found that under these conditions, listeners suffered ongoing interference from their own knowledge of reality, which inhibited successful anticipation of the speaker's intended referents.

Keywords: Theory of Mind; Social communication; Eye tracking

Introduction

Conversation is a form of social interaction whose success relies heavily on speakers and hearers being able to anticipate or infer the mental states of others. Given the frequency with which conversational interactions occur in everyday life, one might assume that we can perform the relevant tasks effortlessly. Contrary to this intuition, however, psychological research has shown that predicting other peoples' behaviour is more difficult than simply considering our own perspective and can be made even more difficult when we hold conflicting, privileged information about the world (Apperly et al., 2006; German & Hehman, 2006). This problem has been studied extensively in children (e.g. de Villiers & Pyers, 2002), though similar research on healthy adults is much more limited. Nevertheless, recently emerging work suggests that even among healthy adults, inferring someone else's perspective is cognitively costly and subject to interference from our own point of view (Apperly et al., 2009; 2010; Birch & Bloom, 2007). Clearly then, being able to override one's own perspective emerges as a necessary but complex process when taking someone else's perspective. An important question is whether or not the costly nature of 'Theory of Mind' (ToM) processes observed in past studies is an integral part the inferences themselves, making them different or delayed with respect to inferences about behaviour not drawing on mental state information. This issue has recently been addressed in terms of whether ToM inferences are made automatically- or whether they only come into play under certain circumstances (as in Apperly et al., 2006; Back & Apperly, 2010; Cohen & German, 2009).

Over the last decade, the topic of perspective-taking has generated increasing interest in relation to language comprehension. In particular, much recent focus has been on the viability of Keysar and colleagues' Strategic Egocentrism model (Keysar et al., 2000). According to this account, listeners automatically interpret referring expressions according to their own perspective, then later adjust to fit with the speaker's perspective. As such, taking someone

else's perspective is not a default process during communication, but a more deliberate secondary process that can be activated according to need (i.e. to resolve confusion in conversation). Although many early studies favoured such an egocentric account (e.g. Barr & Keysar, 2002; Keysar, & Barr, 2005), much subsequent research suggests that listeners employ an interactive language comprehension process that simultaneously makes use of multiple probabilistic constraints, including the speaker's perspective (Brown-Schmidt et al., 2008; Hanna et al., 2003; Heller et al, 2008).

Typically, this previous research has manipulated referentially ambiguous expressions as part of an interactive reference assignment task in which participants follow a speaker's instructions to move objects around a static grid. Importantly, some of the objects in the display are occluded from the speaker's but not the listener's view, thus setting up different perspectives for the two communicators. In a critical trial, the speaker might ask the participant to "move the cup right", when there are two cups in the grid but only one of them is visible to the speaker. In such a trial, the referential ambiguity can only be solved by adopting the speakers' perspective, which limits the set of potential targets. The listeners' eye movements around the scene are recorded for accurate assessment of perspective-taking as the description unfolds.

In this way it has been established that the linguistic input has a very strong bottom-up effect on decisions about the referent of an expression (i.e. early anticipatory looks reveal that all linguistically possible candidates are considered initially and that information about what referents the speaker knows about can have little initial effect on this process). Whether this is a result of an absolute modular divide (Barr, 2008) or just a very strong constraint of language input (Brown-Schmidt et al., 2008; Hanna et al., 2003; Heller et al, 2008) remains an open question. In this paper one of our aims is to explore whether there are other factors that can interfere with the rapid integration of mental state information during utterance

interpretation. The current study does not involve any referential ambiguities but involves the speaker holding an out-of-date perspective on an object's location. In this way we will explore the effect of so-called 'pull of reality' (Birch & Bloom, 2004) as an independent factor in discourse comprehension.

The Current Study & Predictions

The current experiment used a novel interactive video task to examine how perspective information influences adults' online expectations about reference. Pairs of communicators watched short videos depicting an actor transferring objects to one of two locations, then subsequently answered on-screen questions, as in (1), which prompted them to describe the events to each other, as in (2). Importantly, on half the trials, a screen covered the speaker's but not the listener's view half way through the video. On those occasions, the unseen video footage may or may not have contained subsequent transfer of the target object, meaning that on these trials, the speaker held out-of-date information about the object's location.

(1) *Where is the umbrella?*

(2) The umbrella is in box A.

This task is similar to Wimmer and Perner's (1983) classic "transfer" ToM task, where children were prompted to answer questions about an object's location, based on a character's false belief. Critical to both of these tasks is the fact that the object's transfer event was either witnessed or explicitly missed by a relevant other party (i.e. Maxi in Wimmer and Perner's study, and the speaker in the current study). However, there are also important differences when compared to Wimmer and Perner, in that the speaker in our task was aware that they held out-of-date information about the location of the objects and was

aware that on some trials the object might move location whilst the screen was covered. Though these factors would not change to the speaker's answer to the question (since they can only answer according to what they *know*, and not guess), it is possible that the speaker's ignorance may impact on the strength of listener's predictions. These points will be addressed in detail in light of our results, in the discussion below.

The fully crossed knowledge (shared/ privileged) by movement (moved/ not moved) design employed here provides useful baselines to examine genuine effects of perspective-taking in comparison with other factors that may be involved in directing eye movements around visual scenes. We predicted that under conditions where both communicators shared knowledge about the object's location, listeners would be able to successfully predict the mutually appropriate continuation long before the speaker has finished their description, regardless of the type of transfer event (*moved* or *not moved*). As such, any differences emerging in the strength of the reality-bias between the two shared conditions can only be attributed to low-level visual factors (e.g. influences from the 'last seen' location). In contrast, if the speaker was ignorant of a critical part of a 'moved' transfer sequence and they answer based on out-of-date information, their description will be at odds with the knowledge held by the listener. The effects of this conflict can be contrasted with a condition where the privileged transfer event did not involve a change in the object's location- meaning that the listener's knowledge of reality and the speaker's out-of-date information about the object's location were fully consistent. Therefore, if listeners do consider the speaker's perspective on such trials, we would expect to see this reflected in their visual biases, with a significantly reduced reality-bias for 'privileged' trials when the object was moved compared to when the object was not moved. The extent to which the difference in reality-bias between moved and not moved trials increases under privileged conditions compared to shared conditions, can only be attributed to successful perspective-taking.

An additional question is whether listeners are able to fully adopt the speaker's perspective on 'false belief' trials and show marked anticipation for the speaker-appropriate location, regardless of their own knowledge. If listeners can access and utilize their partner's perspective online, visual biases should show a significant preference for the alternative box prior to the auditory location onset. Alternatively, if accommodating a perspective that conflicts with salient reality is subject to a 'pull of reality', visual bias towards the appropriate location may be delayed or even subject to a bias to the reality perspective.

In order to test these predictions, we compared the time-course and location of listeners' predictive visual biases when the listener and speaker shared knowledge about an event's outcome with when the two communicators had conflicting perspectives on the object's location. Importantly, we eliminated interference from low-level visual cues by hiding the target object inside one of two plain opaque containers.

Method

Participants

Forty participants from the University of Cambridge were paired with one of two confederates. All were paid to participate.

Stimuli and Design

Twenty sets of experimental videos and pictures were paired with an auditory description in one of four conditions. Video clips were recorded in a single session involving one male and one female 'actor' and edited using Adobe Premier. All visual images were presented on a 17 inch colour monitor in 1024 x 768 pixels resolution.

Two different video scenarios set up relevant contexts. Both video scenarios began with two objects in the centre (e.g. an umbrella and a bunch of keys) and two possible target

locations (boxes labelled A and B). All video clips began with the actor moving one of the objects into one of the boxes (the other object simply served as a distracter). To set up the two visual states, a second part of the video depicted the actor lift the target object out of the original box then either replace it back into the same box (i.e. the *no-move* state), or move it into the other box (the *move* state). Subsequent pictures then depicted the final state from each of these scenarios (i.e. the two closed boxes), and were created by extracting the final frame from each video clip. Systematic viewing strategies were prevented by counterbalancing the spatial arrangement of the objects across items.

The two communicators were 'randomly' assigned roles for the experiment (i.e. participants were made to believe that the roles of speaker and listener were randomly assigned): a 'speaker' (always the confederate) who described events in the video, and a 'listener' (the participant) who listened to the speaker's description while viewing an image from the video. Confederates were naïve to the purpose of the study as well as whether the object was moved on a given trial or not (i.e. whether their description reflected accurate or out-of-date information). Note that participants also experienced the speaker's role during a short practice block and at two 'switch' points during the experiment, where participant and confederate swapped roles. This validated the experimental manipulation and ensured that participants fully understood the listener's task.

To set up the two perspective states, on half the trials a screen was used to obscure the speaker's (but not the listener's) view during the second part of the video, therefore setting different levels of knowledge for the two communicators (privileged *vs.* shared). The listener (participant) believed that the speaker's description was prompted by a question that appeared on their screen, e.g. *Where is the umbrella?*, and they experienced this first-hand on trials where they played the role of speaker. Speakers were instructed to answer the question only as far as their knowledge allowed, and not to guess. For experimental trials, auditory

descriptions were scripted to ensure consistency of the object names and description types for analysis. We can assume that participants were clearly aware that speakers only answered the questions 'as far as they knew' since in practice trials and at switch points, participants behaved in this way also.

Under these conditions, in the *no-move* state, the speaker's perspective always fits with the actual state of affairs, regardless of whether the second part of the video was *privileged* or *shared* between both communicators. In contrast, in the *move* state, if the speaker did not see the second part of the video, a conflicting perspective arises, whereby the speakers' description will mismatch the privileged knowledge held by the listener.

One version of each item was assigned to one of four presentation lists, with each list containing twenty unique experimental items, five in each of the four conditions. Participants were randomly assigned to one of these four lists, which ensured that across these lists (and therefore across participants) each video was seen in all four conditions. By using this fully counterbalanced design, we can be confident that any differences between conditions cannot be due to natural cues (e.g. lighting, contrast) in the video stimuli themselves. In addition, fifty-six filler items were added to each list. Of these, twenty involved conflicting perspective in the speaker (referring to quantity or type of transfer object) and thirty-six were shared by the two communicators. All depicted a transfer action, involving either one or more target objects and the same two target locations (boxes A and B). These filler items were interspersed randomly among the experimental trials to create a single random order. Comprehension questions followed half of the experimental and half of the filler trials and both the listener (participant) and speaker (confederate) responded to these questions. All questions probed the participant's knowledge of the real state of events. All participants scored at or above 80% accuracy on the comprehension questions.

Procedure

Participants sat in front of a colour monitor while eye movements were recorded using a stand-alone eye tracking system (Tobii X120) running at 120 Hz sampling rate. Viewing was binocular and eye movements were recorded from both eyes simultaneously. The confederate sat at a separate monitor and spoke into a microphone, which recorded audio responses to file. See Figure 1 for a schematic diagram of the experimental setup.

----------FIGURE 1 ABOUT HERE----------

As illustrated in Figure 2, each trial began with the presentation of a single centrally-located cross, which participants fixated for 1000msec. At this point, the cross was replaced by the video depicting a transfer scenario. On *privileged* trials, a message appeared on both monitors half-way through the video instructing the experimenter to put the "Screen up", thus covering the speaker's screen during the second part of the video. Both confederate and participant pressed a button to continue when prompted by the experimenter. Video clips lasted on average 25 seconds (range = 19s to 34s) in total and were followed by a blank screen for 500ms. Next, the corresponding picture was presented to the participant as the confederate gave a spoken description of events. During this time, participants believed that the speaker's screen showed a question, which prompted the verbal description, as described earlier. Confederates were instructed to begin their description when a 'beep' was heard, 500ms after picture onset. This ensured that the onset of the picture preceded the onset of the corresponding spoken description by at least 500ms. The picture stayed on-screen until the confederate's verbal description was finished and both confederate and participant pressed a button to move on.

----------FIGURE 2 ABOUT HERE----------

At the beginning of the experiment, and every ten trials thereafter, the eye-tracker was calibrated against nine fixation points. This procedure took about half a minute and an entire session lasted about an hour and a half. After the experiment, we tested participants' belief in the confederate and experimental design by asking them to rate how strongly they agreed or disagreed with the four statements below (on a scale of 1-7, with 7 being 'strongly agree'). No participants responded below 5 on any question; Mean responses are shown in brackets.

1. My partner in the experiment gave me accurate descriptions of the videos, as far as their knowledge allowed (6.82);
2. Apart from when the screen was up, I believe that my partner was watching the same video clips as me (6.79);
3. I believe that my partner could not view the videos when the screen was covered (7);
4. I believe that my partner was a real participant (6.93).

Results & Discussion

Data Processing

Eye-movements that were initiated during the auditory description were processed according to the relevant picture and word onsets. For analysis, we removed any sample that was deemed 'invalid' due to blinks or head movements. The spatial coordinates of the eye movement samples (in pixels) were then mapped onto the appropriate object regions. Finally, temporal onsets and offsets of the gazes were recalculated relative to the corresponding picture onset.

Probabilities of fixating the 'reality' or 'alternative' box as a function of time were analysed using the log-ratio measure (see Arai et al., 2007): log(Reality/Alternative) =

$ln(P_{(Reality)} / P_{(Alternative)})$. Here, $P_{(Reality)}$ refers to the probability of fixating the reality box (i.e. where the target object *really* is) and $P_{(Alternative)}$ to the probability of fixating the alternative box; *ln* refers to the natural logarithm. The output is therefore symmetrical around zero such that a positive score reflects higher proportions of fixations on the reality box and a negative score reflects higher proportions of fixations on the alternative box.

Log(Reality/Alternative) scores were analysed for five consecutive regions, determined according the onsets and offsets of words in the corresponding auditory input. These word-regions were identified and synchronised for each participant on a trial-by-trial basis, relative to the actual (i.e. non-adjusted) onsets and offsets of relevant words. The resulting average word-region durations are detailed in Table 1. None of these word lengths differed significantly across the four conditions (Fs < 2). Figure 3 plots the average log(Reality/Alternative) data for each condition, for every 20 ms time-slot. Note that eye movements and auditory input have been resynchronized according to individual word onsets (see Altmann & Kamide, 2009), and as such represent more accurate plots of evolving visual biases around the scene[1].

----------TABLE 1 ABOUT HERE----------

----------FIGURE 3 ABOUT HERE----------

Main Analyses

For each participant (respectively item) and condition, we calculated a weighted average log(Reality/Alternative) score over the 20ms time slots per analysis region. Statistical analyses used an ANOVA with Movement (*no-move* vs. *move*) and Knowledge (*privileged*

[1] Statistical tests were performed on time windows defined by the absolute onsets (i.e. non-adjusted) of relevant critical words. This is due to the variability in estimates of the delay between the time taken to program and execute a saccadic eye movement following the auditory onset of relevant words (See Altmann, in press, for a full discussion).

vs. *shared*) as the repeated-measures factors. Table 3 displays the statistical details of the effects for each time window of interest, allowing generalization to participants (F_1; in which participants are seen as a random factor and items as a fixed factor) and items (F_2; in which items are seen as a random factor and participants as a fixed factor). Significance on both these tests will ensure generalizability of the results across the different participants and experimental items. Strength of association is reported in terms of partial eta-squared ($_p\eta^2$).

----------TABLE 2 ABOUT HERE----------

Immediately from "the" and [Object] onset and persisting throughout the remaining word-regions, the ANOVAs showed a main effect of Movement. This effect was due to a higher proportion of fixations on the box that actually contained the object (the 'reality box') when only one box had been actively involved in the transfer event (*no-move* conditions), compared to when the object had been moved between the two boxes (*move* conditions). Thus, memory of the object's previous location has the effect of weakening the strength of bias to the reality box. A main effect of Knowledge also emerged from the [Object] through "is in", "box" and "A/ B". These effects reflected a stronger bias to the 'reality box' when both speaker and listener *shared* knowledge about the object's location, compared to when the listener held *privileged* knowledge about the object's location. As such, it seems that listeners here have taken the speakers' limited knowledge into account to direct their visual attention towards appropriate visual objects.

Additionally, during "is in" and "box" the results showed a significant Movement*Knowledge interaction (marginal by items); follow-up analyses of the simple main effects examined the nature of these interactions. Analyses revealed that for *privileged* conditions, the strength of the reality-bias was reduced for *move* versus *no-move* trials in both

"is in" ([$F_1(1,39) = 47.28$, $p < .001$, $_p\eta^2 = .55$; $F_2(1,19) = 52.11$, $p < .001$, $_p\eta^2 = .73$]) and "box" ([$F_1(1,39) = 15.81$, $p < .001$, $_p\eta^2 = .29$; $F_2(1,19) = 14.1$, $p < .001$, $_p\eta^2 = .43$]) word regions. However, when the transfer events were *shared* between communicators, this reduced reality-bias for *move* trials only emerged during "is in" ([$F_1(1,39) = 6.33$, $p < .02$, $_p\eta^2 = .13$; $F_2(1,19) = 4.21$, $p < .05$, $_p\eta^2 = .18$]), with no difference between *move* and *no-move* trials during "box" (all Fs < 2). Further, when the object was moved into a new box, participants were significantly less likely to fixate the reality box when knowledge of the second transfer event was *privileged* (compared to *shared*) during both "is in" ([$F_1(1,39) = 32.86$, $p < .001$, $_p\eta^2 = .46$; $F_2(1,19) = 30.23$, $p < .001$, $_p\eta^2 = .61$]) and "box" ([$F_1(1,39) = 24.33$, $p < .001$, $_p\eta^2 = .38$; $F_2(1,19) = 12.31$, $p < .001$, $_p\eta^2 = .39$]) word regions. In contrast, when the object was not moved to a new location, participants were equally likely to fixate the reality box for *shared* and *privileged* conditions (all Fs < 3). These results suggest that perspective has a greater impact on expectations when the *privileged* knowledge by one communicator is at odds with the other person's perspective.

Finally, it is interesting to note that while clearly showing a reduced bias to the reality box when the speaker held out-of-date information about the object's location, listeners did not appear to wholeheartedly adopt the speaker's perspective. This was evidenced by the finding that, in *privileged* knowledge trials (*move* condition), listeners did not direct their attention reliably to the speaker-appropriate box, but instead showed no significant anticipatory bias to either of the potential object containers (all word regions: ts < 1.2).

General Discussion

These results demonstrate that during interactive communication, perspective information is rapidly integrated with other contextual cues to direct expectations about forthcoming referents. As predicted, an interaction between knowledge and movement emerged in

anticipation of the disambiguating auditory input (i.e. during "is in" and "box"), reflecting a significantly decreased reality-bias when a transfer event occurred but was not observed by the speaker. Such an effect demonstrates that listeners were indeed considering the speaker's perspective when directing their gaze around the scene. However, in the earlier [Object] region, the main analyses found only a main effect of knowledge instead of the predicted interaction, with a significantly reduced reality-bias for *privileged* versus *shared* conditions. Indeed, this main effect persisted throughout the time-series from the [Object] onset. We argue that this specific pattern of results is not inconsistent with the view that perspective information has a very early effect on bias formation.

Recall that in the introduction we outlined the similarities between this task and Wimmer and Perner's classic "transfer" ToM task, but we also noted that an important difference between the two tasks was the naïve other party's confidence in their knowledge. Specifically, in Wimmer and Perner's task, Maxi had no reason to expect his chocolate to have been moved in his absence, so one can easily assume that he will look for the chocolate in the last place he saw it. In contrast, speakers in the current study were made aware (through their experience on *shared* trials) that the target object changed location half the time. Thus, in privileged conditions, listeners held two pieces of information about the speaker: (i) that they are ignorant of the actual answer to the question, and (ii) that they are answering based on knowledge of events up to half-way through the video. A strategy based on the first piece of information (speaker's ignorance) would be to not expect any particular location since an ignorant person can only guess. Such a strategy would be revealed by a uniform effect in both privileged conditions (*move* and *no-move*) toward the no-bias line. In contrast, a strategy based on the second piece of information (speakers using out-of-date knowledge) would be to represent the speaker's knowledge from half-way through the video and form expectations based on that. The early main effects of knowledge, where the reality-

bias was lower in *privileged* versus *shared* conditions, offers partial support that listeners initially considered the first strategy, possibly due to the salience of the speaker's ignorance of the final outcome. However, the clear difference in bias formation between the *shared* and *privileged moved* conditions that emerged during "is in" and "box" words regions, suggests the latter strategy is predominantly being employed. Listeners appear to have rapidly accepted the speaker's ignorance and based their expectations on knowledge that the speaker is answering on the basis of out-of-date information. Thus, in our results we see the influence of two separate pieces of perspective information: an early ignorance effect followed by a lasting effect of knowing that the speaker is answering on the basis of out-of-date knowledge.

Overall, our results clearly demonstrate that perspective influences anticipatory visual biases online. It is interesting to note that this perspective-taking occurs spontaneously, even in situations where communicators have not been given an explicit reason to track another person's mental state. Recent research in this area suggests that the type of discourse used may have an effect on the time-course of perspective use. Specifically, the most commonly employed paradigm involves the participant following instructions from a speaker, as described above, where the focus is on what the speaker's mental state (what she wants) (Barr & Keysar, 2002; Hanna et al., 2003; Heller et al., 2008; Keysar et al., 2000; Keysar, & Barr, 2005). However, as one might expect, improved perspective-taking is evident where an interactive question-answer discourse between the speaker and hearer explicitly established what the speaker did and did not know (Brown-Schmidt et al., 2008; Brown-Schmidt, 2009). This raises the question of whether the degree to which ToM is spontaneously employed in comprehension may depend on the type of discourse. For example it may be that question-answer discourse tends to trigger spontaneous employment of ToM while simple narrative descriptions do not, or at least do so to a lesser degree.

The fact that our study used a 'look-and-listen' paradigm and not 'follow-the-instruction' (Barr & Keysar, 2002; Hanna et al., 2003; Heller et al., 2008; Keysar et al., 2000; Keysar & Barr, 2005) or 'question-answer' (Brown-Schmidt et al., 2008; Brown-Schmidt, 2009) paradigms used in previous online perspective-taking research, provides evidence that hearers *spontaneously* seek to incorporate information about the speaker's mental state during communication. Moreover our results suggest that even in interpreting simple narrative discourse, participants are attempting to anticipate the speaker's intended referent rather than merely the most likely referent based on current knowledge from the visual context.

Overall, our findings fit with previous research showing the spontaneous online use of perspective in look-and-listen narrative comprehension tasks, which have shown clear and early use of perspective when predicting events in the narrative according to a character's false beliefs (Ferguson et al., 2010) or conflicting desires (Ferguson & Breheny, 2011). The current paper substantially extends our current understanding by examining realistic interactive two-person communication. Moreover, although our studies do not directly test whether ToM inferences are automatic (Apperly et al. 2006; German & Hehman, 2006), the results do suggest that the spontaneous use of ToM is very pervasive in utterance comprehension, suggesting that the deployment of ToM may be automatic in that domain.

However, despite this clear evidence of spontaneous sensitivity to mental state information, the results also showed some evidence that listeners experienced conflict from their own knowledge. The key indication of such an effect was apparent when the speaker's out-of-date knowledge was at odds with the actual state of affairs (*move-privileged* trials). Despite showing a clear *reduction* in reality-bias, listeners did not successfully predict reference according to the speaker's perspective (i.e. no significant preference to fixate the alternative box). While this pattern could be explained in terms of listeners assuming that the ignorant speaker is just guessing, this explanation is not consistent with the pattern in the *no*

move-privileged condition which does not differ from the *no move-shared* condition in later time windows. As suggested above, the fact that the speaker is ignorant of the real state of affairs has a strong effect early in the spoken items but in later windows, participants seemed to be more influenced by the fact that the speaker does not guess but answers on the basis of out-of-date information. Instead we attribute the lack of bias to the alternative location in the *move-privileged* condition more to the fact that the object's actual location acted as a 'curse of knowledge' for our listeners, eliciting a pull of reality (PoR) or reality-bias (Birch & Bloom, 2004, 2007; Mitchell et al., 1996), and suggest that our results shed some light on the nature of such an effect.

By definition, a PoR effect arises when an agent must construct and utilise a representation that differs from that of the actual state of affairs. So, the source of PoR effects could lie in the construction/ accessing the non-actual representation on one hand, or in 'competition' or 'interference' from the reality representation on the other, or in a combination of the two (a similar proposal has been put forward by Altmann & Kamide, 2009). Here, we have shown evidence that even when knowledge of the second movement event was *shared*, listeners maintained a representation of the object in its former location (as indexed by the stronger reality-bias for *no-move* compared to *move* trials). These data suggest that in our task, participants did not find the process of constructing/ accessing the non-reality representation particularly difficult. So, it seems that the lack of prediction in the *moved-privileged* condition was more due to interference from the final reality-based representation. This kind of interference is thought to result from inhibition difficulties during the formulation of online anticipatory hypotheses (Friedman & Leslie, 2004).

In sum, our results suggest that perspective-taking and other ToM processes are spontaneously recruited during all forms of discourse comprehension. However, they also demonstrate some interference from the listeners' knowledge of the object's actual location.

We argue that such an effect reflects a PoR effect arising from difficulty in inhibiting 'reality' representations, which leads to listeners holding multiple representations of depicted events that compete with each other during processing. Importantly, such interference effects are seen here, operating independently of any lower-level language-driven effect established in previous comprehension research.

References

Altmann, G.T.M. (in press). Language can mediate eye movement control within 100 milliseconds, regardless of whether there is anything to move the eyes to. *Acta Psychologica.*

Altmann, G.T.M. & Kamide, Y. (2009). Discourse-mediation of the mapping between language and the visual world: eye movements and mental representation. *Cognition*, *111*, 55-71.

Apperly, I.A., Riggs, K.J., Simpson, A., Samson, D., & Chiavarino, C. (2006). Is belief reasoning automatic? *Psychological Science, 17*, 841-844.

Apperly, I.A., Samson, D., & Humphreys, G.W. (2009). Studies of adults can inform accounts of theory of mind development. *Developmental Psychology, 45*, 190-201.

Apperly, I.A., Carroll, D.J., Samson,D., Qureshi, A., Humphreys, G.W., & Moffitt, G. (2010). Why are there limits on theory of mind use? Evidence from adults' ability to follow instructions from an ignorant speaker. *Quarterly Journal of Experimental Psychology, 63*, 1201-1217.

Arai, M., van Gompel, R.P.G., & Scheepers, C. (2007). Priming ditransitive structures in comprehension. *Cognitive Psychology*, *54,* 218- 250.

Back, E. & Apperly, I.A. (2010). Two sources of evidence on the non-automaticity of true and false belief ascription. *Cognition, 115*, 54-70.

Barr, D.J. (2008). Pragmatic expectations and linguistic evidence: Listeners anticipate but do not integrate common ground. *Cognition, 109*, 18-40.

Barr, D.J., Gann, T.M., & Pierce, R.S. (in press). Anticipatory effects and information integration in visual world studies. Acta Psychologica.

Barr, D.J., & Keysar, B. (2002). Anchoring comprehension in linguistic precedents. *Journal of Memory and Language, 46*, 391-418.

Birch, S.A.J., & Bloom, P. (2004). Understanding children's and adults' limitations in reasoning about the mind. *Trends in Cognitive Sciences, 8*, 255-260.

Birch, S.A.J., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science, 18*, 382-386.

Brown-Schmidt, S. (2009). The role of executive function in perspective-taking during on-line language comprehension. *Psychonomic Bulletin and 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.

Cohen, A.S., & German, T.C. (2009). Encoding of others' beliefs without overt instruction. *Cognition, 111,* 356-363.

de Villiers, J., & Pyers, J. (2002). Complements to cognition: a longitudinal study of the relationship between complex syntax and false-belief understanding. *Cognitive Development, 17*, 1037-1060.

Ferguson, H.J., & Breheny, R. (2011). Eye movements reveal the time-course of anticipating behaviour based on complex, conflicting desires. *Cognition, 119*, 179-196.

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

Friedman, O., & Leslie, A.M. (2004). A developmental shift in processes underlying successful belief-desire reasoning. *Cognitive Science*, *28*, 963-977.

German, T.P., & Hehman, J.A. (2006). Representational and executive selection resources in 'theory of mind': Evidence from compromised belief-desire reasoning in old age. *Cognition, 101*, 129-152.

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., & Tananhaus, M.K. (2008). The Role of Perspective in Identifying Domains of References. *Cognition 108*, 831-836.

Keysar, B., & Barr, D.J. (2005). Coordination of action and belief in communication. In J.C. Trueswell & M.K. Tanenhaus (Eds.), *Approaches to Studying World Situated Language Use: Bridging the Language-as-Product and Language-as-Action*. Cambridge, MA: MIT Press.

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.

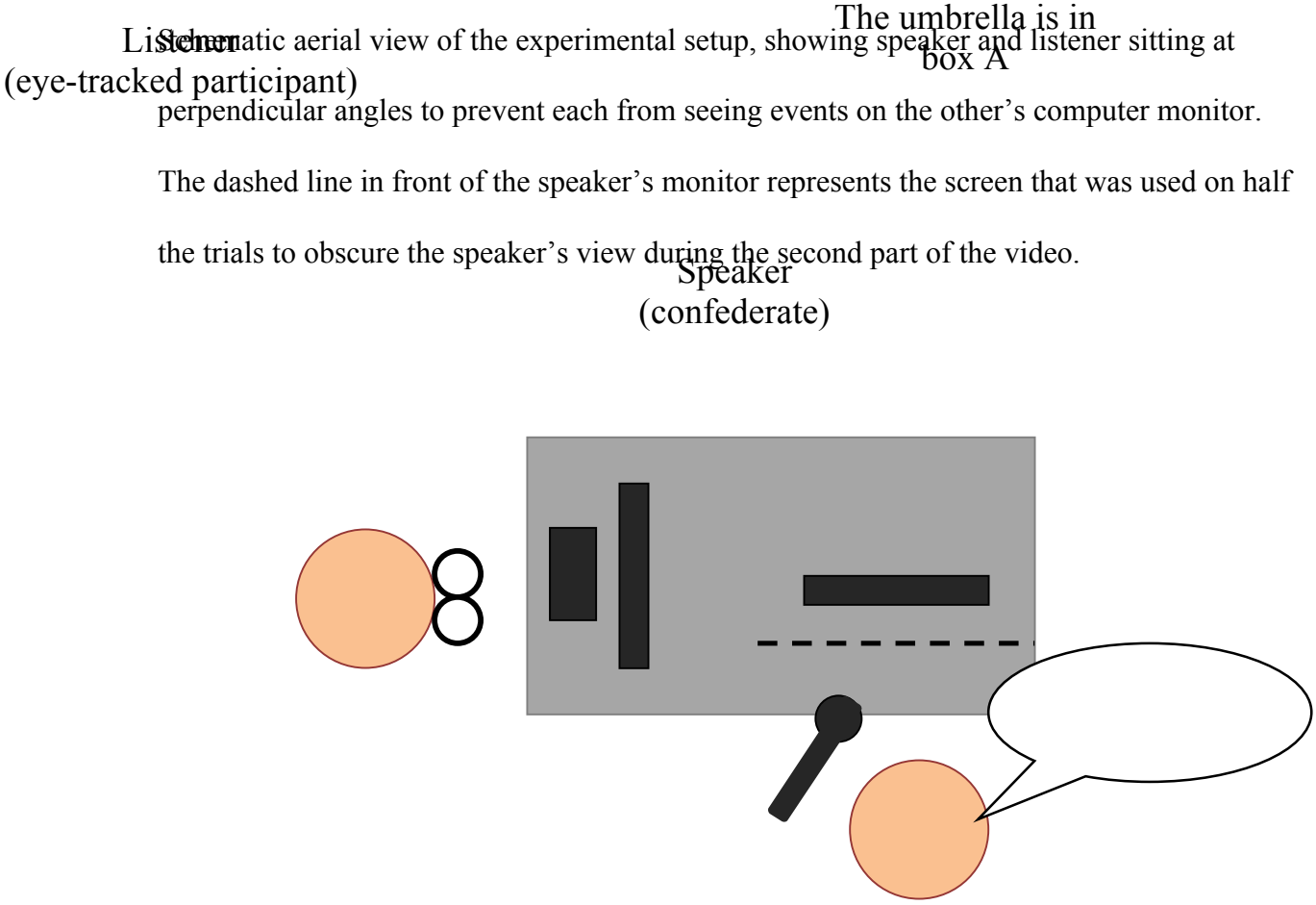
Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition, 89*, 25-41.

Mitchell, P., Robinson, E.J., Isaacs, J.E., & Nye, R.M. (1996). Contamination in reasoning about false belief: An instance of realist bias in adults but not children. *Cognition, 59*, 1-21.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*, 103-28.

Figure & Table captions

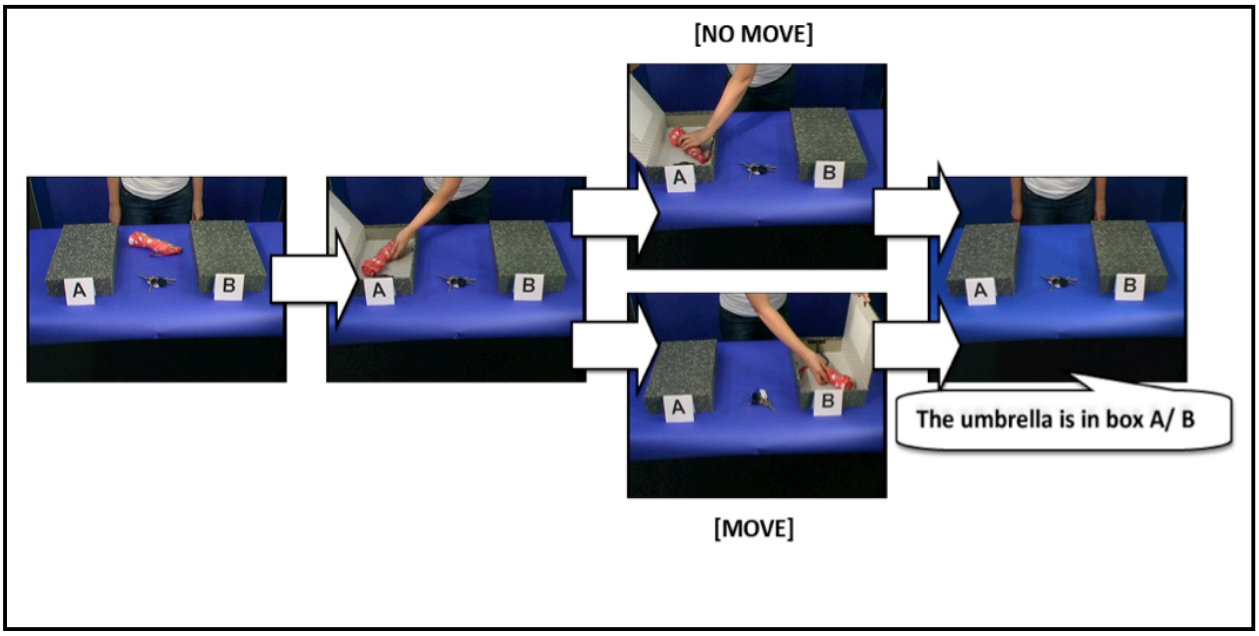
Figure 1:



Schematic aerial view of the experimental setup, showing speaker and listener sitting at perpendicular angles to prevent each from seeing events on the other's computer monitor. The dashed line in front of the speaker's monitor represents the screen that was used on half the trials to obscure the speaker's view during the second part of the video.

Figure 2

Schematic trial sequence of visual displays presented to participants. Stage 1 depicts the 'start state'. In stage 2, a video showed one object being put into one of two boxes. Stage 3 either showed the object being lifted out of the original box then replaced back into the same box (i.e. *no-move* state), or moved into the other box (*move* state). Note that on half the trials, events in Stage 3 were only seen by the participant (the listener), thus setting up a false belief for the speaker on *move* trials. Finally, Stage 4 shows the 'final state' picture that participants saw while they listened to their partner's description of events (which was prompted by a question, such as, *Where is the umbrella?*).



Stage 1 *Stage 2* *Stage 3* *Stage 4*