THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY
2012, i*First*, 1–15

Gaze transfer in remote cooperation: Is it always helpful to see what your partner is attending to?

Romy Müller[1], Jens R. Helmert[1], Sebastian Pannasch[1,2], and Boris M. Velichkovsky[1,3]

[1]Applied Cognitive Research/Psychology III, Technische Universität Dresden, Dresden, Germany
[2]Brain Research Unit, Aalto University School, Espoo, Finland
[3]Department of Cognitive Studies, Kurchatov Institute, Moscow, Russian Federation

Establishing common ground in remote cooperation is challenging because nonverbal means of ambiguity resolution are limited. In such settings, information about a partner's gaze can support cooperative performance, but it is not yet clear whether and to what extent the abundance of information reflected in gaze comes at a cost. Specifically, in tasks that mainly rely on spatial referencing, gaze transfer might be distracting and leave the partner uncertain about the meaning of the gaze cursor. To examine this question, we let pairs of participants perform a joint puzzle task. One partner knew the solution and instructed the other partner's actions by (1) gaze, (2) speech, (3) gaze and speech, or (4) mouse and speech. Based on these instructions, the acting partner moved the pieces under conditions of high or low autonomy. Performance was better when using either gaze or mouse transfer compared to speech alone. However, in contrast to the mouse, gaze transfer induced uncertainty, evidenced in delayed responses to the cursor. Also, participants tried to resolve ambiguities by engaging in more verbal effort, formulating more explicit object descriptions and fewer deictic references. Thus, gaze transfer seems to increase uncertainty and ambiguity, thereby complicating grounding in this spatial referencing task. The results highlight the importance of closely examining task characteristics when considering gaze transfer as a means of support.

Keywords: Gaze transfer; Cooperation; Pointing; Ambiguity; Resolution; Grounding.

For joint actions to be performed fluently and accurately, it is essential to know how a partner is interacting with task-relevant objects. In real life cooperation, this can be achieved by verbal communication and by observing a partner's actions in the environment. Consequently, conversational

Correspondence should be addressed to Romy Müller, Applied Cognitive Research/Psychology III, Technische Universität Dresden, Helmholtzstraße 10, 01069 Dresden, Germany. (E-mail: romy.mueller@psychologie.tu-dresden.de)

Parts of this research were presented at the 10th Annual Conference of the Society for Cognitive Science (KogWis 2010, Potsdam, Germany, October 2010) and at the 4th International Conference on Cognitive Science (Tomsk, Russia, June 2010). We thank Daw-An Wu, Evan F. Risko, Cathy Reed, and Cathy Nangini for very helpful comments on an earlier version of the manuscript. Furthermore, we thank Anett Reiche, Josephine Hartwig, Caroline Gottschalk, and Antje Grabowski for their support in data acquisition and Gernot Pascher for his assistance in the technical implementation of the experimental setup. This research was supported by a grant of the TU Dresden Centre for Continuing Education to RM, by the European Commission (FP7-PEOPLE-2009-IEF, EyeLevel 254638) to SP, and by the Russian Foundation for Basic Research (Interdisciplinary oriented research 09-06-12003) to BMV.

http://www.tandfonline.com http://dx.doi.org/10.1080/17470218.2012.737813

grounding can take place in a direct manner, using visual information to provide and collect evidence for a shared understanding (Clark & Brennan, 1991). In remote cooperation, special challenges for communication arise. The lack of visual cues complicates the establishment of common ground and thereby increases the need for detailed verbal descriptions and feedback (Marshall & Novick, 1995). In consequence, ambiguities can arise and hamper coordinated planning and acting. Such difficulties in mutual understanding are highly persistent and usually cannot be eliminated with the help of video-conferencing technology (Whittaker, 1995).

Therefore, when supporting remote communicative processes with technical devices, it is crucial to consider the underlying mechanisms of cooperation in a given context. In direct communication, the role of eye movements cannot be valued highly enough. A partner's gaze provides a central cue for establishing joint reference frames: By following each other's gaze, interlocutors establish a shared focus on relevant objects, a state called *joint attention* (Bruner, 1981). The degree to which partners look at the same objects at approximately the same time can indicate how well they understand each other (Richardson, Dale, & Kirkham, 2007). Furthermore, eye movements are highly predictive of verbal object descriptions, preceding them by approximately one second (Griffin & Bock, 2000). During face-to-face conversations, an interlocutor's gaze is used to disambiguate his object references even before the point of verbal disambiguation (Hanna & Brennan, 2007). Furthermore, since information processing is to some extent reflected in eye movements (Velichkovsky, 2002), gaze can provide a window to a partner's visual attention and awareness (Vertegaal, Velichkovsky, & Van der Veer, 1997). Following this argumentation, the question arises whether gaze provides valid support to reach mutual understanding when the means of communication are highly restricted in more artificial settings. Or, in other words: Do these gaze benefits hold true for remote cooperation?

Previous studies suggest that remote partners can make use of gaze information. When measuring the eye movements of one or both partners and superimposing them on the other partner's stimulus material as a gaze cursor (e.g., Carletta et al., 2010), people can continuously track the other's focus of visual attention. This form of *gaze transfer* helps to improve performance. For instance, it increases the efficiency of joint visual search, presumably because people become aware of where their partner has already been looking, thereby avoiding redundant search (Brennan, Chen, Dickinson, Neider, & Zelinsky, 2008). However, in contrast to cooperation in real world tasks, a correct interpretation of the partner's eye movements was less crucial in Brennan et al. where two participants were merely searching in parallel. Since each of the participants was able to solve the task alone, a peripheral monitoring of the partners' gaze cursor was sufficient. The results, therefore, do not permit conclusions about the potential of gaze transfer to improve conversational grounding in more interdependent settings.

Gaze benefits have also been found when interpreting the partner's gaze was crucial to solve the task successfully. Neider, Chen, Dickinson, Brennan, & Zelinsky (2010) compared gaze and speech transfer while participants jointly searched for sniper targets in a street scene. Critically, agreement on the target location was required in order to complete a trial. Therefore, the partner who first spotted the target had to indicate its location to the other one, and the trial was finished when both partners looked at the target at the same time and then pressed a key. In this study, gaze transfer did not improve performance in terms of solution rates and search times *per se*, but speeded up the communication of target locations. Participants using gaze transfer reduced the number of verbal location descriptions and speaking turns, relying on short deictic references instead. Thus, they were able to use their partner's gaze to understand what he wanted to point out to them.

Similar conclusions were reached in a joint problem solving task (Velichkovsky, 1995). Pairs of participants were asked to solve puzzles together, while one knew the solution due to prior training (expert), and the other was able to move the

pieces but had no further information (novice). Transmitting the expert's eye movements to the novice allowed for faster solutions than a purely verbal communication. Gaze transfer also diminished verbal object descriptions and evoked a more frequent use of deictic references, again indicating that participants were able to understand their partner's gaze.

Taken together, all three studies suggest that gaze transfer facilitates the mutual understanding of partners and thereby improves remote cooperation. But can we conclude from these results that gaze transfer is easily understandable and that people are generally good at using information about a partner's visual attention? Does gaze transfer provide the best possible means of support for interpersonal coordination? We believe that a note of caution is required because two things are missing in previous studies: a critical control condition and a thorough consideration of task characteristics.

The studies discussed above contrasted gaze transfer with purely verbal interaction. Since the tasks comprised the indication of particular locations, it is not very surprising that providing a gaze cursor as a spatial indicator facilitates target detection, especially when compared with a condition where no visual indicator is available at all. The actual degree to which people can interpret the gaze cursor should be better understood by comparing it with *another* spatial indicator, such as the computer mouse. Although Velichkovsky (1995) used the mouse as a control condition and found similar performance for the two types of cursor transfer, he did not compare them with regard to their effects on the interaction between partners. As performance effects are often hard to detect in remote cooperation studies, even when the underlying mechanisms are quite distinct (Doherty-Sneddon et al., 1997; Monk & Gale, 2002), a more in-depth investigation of the cooperative process is needed. To better understand how partners use gaze transfer as a conversational cue, it should be determined whether gaze and mouse transfer have differential effects on the way people establish common ground.

To address this issue properly, one needs to be more specific when asking about the effects of gaze and mouse. There is general agreement that no conversational medium is good or bad in itself, but each has its own specific profile of grounding costs (Clark & Brennan, 1991). Whether these costs outweigh the benefits largely depends on the match between the features a medium offers and the user's needs in a given context. Apparently, the most distinctive feature of gaze transfer is the abundance of information contained in the cursor movement. Gaze cursors provide the observer with a high-resolution visualization of the partner's visual attention. Although it is often claimed or at least implicitly assumed that gaze transfer benefits stem from just that sort of information, it has not yet been demonstrated. Previous studies left open the possibility that people using gaze transfer merely profit from gaze pointing as means of spatial referencing in tasks where locations need to be indicated. In this case, all relevant information in these tasks could also be transmitted by simpler forms of pointing, such as the mouse. Even more so, the additional information contained in gaze might be not only useless but even harmful if it does not match the task requirements. Interpreting a partner's pointing gestures could become difficult due to the presence of communicatively irrelevant (but not easily distinguishable) cursor movements. Whether gaze transfer can indeed impose a significant cost should become obvious in a task that makes it necessary for a partner to interpret spatial referencing cues quickly and unambiguously.

Clarifying this issue is particularly pressing as two of the previous studies (Neider et al., 2010; Velichkovsky, 1995) used tasks where gaze benefits were most likely due to its referencing function, in terms of either indicating target locations or pointing out the correct puzzle pieces. Still, the choice between gaze and mouse was treated like a merely practical one, based for example on the availability of a flat surface or free hands (Neider et al., 2010). Since no differentiation between gaze and mouse was made with respect to form and content of the cursor feedback, the impression might arise that gaze generally is an advantageous

communication device, perfectly suited to clarify what the partner is intending to show. Indeed, Neider et al. suggested that the value of mouse and gaze cursors for interpersonal coordination was different due to the latter's fast and instrumental nature, arguing that "Gaze cursors can therefore mediate coordination at a finer time scale" (Neider et al., 2010, p. 724).

To test whether gaze transfer really is as unequivocally positive as previous studies seem to indicate, particularly when compared to other means of providing spatial information, the present study contrasted gaze transfer with speech and mouse transfer in a spatial referencing task. Similar to the Velichkovsky (1995) study, pairs of participants had to solve computerized puzzles. The expert knew the solution and instructed the novice, who had no further information but was able to move the pieces. In this setup, the knowledge and action capabilities were distributed between expert and novice, so that solving the task required interpreting and using gaze for a step-by-step coordination of their individual actions.

Building on previous work, we compared gaze and speech transfer from the expert to the novice with purely verbal interaction to replicate gaze benefits over speech, in terms of both performance and the efficiency of reaching mutual understanding about the objects referred to. Furthermore, we included a control condition where only gaze was transmitted. This was done because previous studies using joint visual search had reported that gaze alone could be at least as effective as in combination with speech (Brennan et al., 2008; Neider et al., 2010). We wanted to investigate whether this sufficiency of gaze still held true when a fine-grained coordination was required throughout the solution process.

Most importantly, to examine whether gaze transfer comes with a cost when used for spatial referencing, we contrasted it with the purely intentional pointing information derived from a mouse cursor. In this situation the information about attention and search processes contained in gaze might complicate the understanding of the expert's intentions but should not necessarily impair overall performance (see Velichkovsky, 1995). However, difficulties in understanding the cursor should clearly be reflected in measures of the cooperative process: Novices should become more careful when reacting to the gaze cursor, and more verbal effort and clarification should be needed to establish common ground.

Furthermore, to test whether the effects depend on the level of cooperation a task affords, we manipulated whether the novice had to follow each of the expert's instructions (low autonomy) or was allowed to move pieces of his own choosing as well (high autonomy).

METHODS

Subjects

Overall, 96 students of the Technische Universität Dresden (74 females) with a mean age of 22.9 years (range 17–38 years, $SD = 4.28$) took part in this study. Half of them participated in the high-autonomy version of the experiment and the other half participated in the low-autonomy version. Subjects were invited in pairs and randomly assigned to their roles in the experiment (expert or novice). All participants had normal or corrected-to-normal vision, were native German speakers, and received either course credit or a payment of €5 per hour. Informed consent was obtained according to local ethical guidelines and the experiment was conducted in conformity with the Declaration of Helsinki.

Apparatus

The two participants were seated in the same room, close enough to hear each other speaking but visually separated by a portable wall. For the experiment, five computers were connected via Ethernet. Two computers were used for stimulus presentation, two computers recorded the speech, and one computer served as host PC for the eye movement recordings. Eye movements of the expert were sampled monocularly at 500 Hz using the SR EyeLink 1000 infrared eye tracking system (SR Research, Ontario, Canada) in the

remote recording mode with an online detection of saccades and fixations and a spatial accuracy of better than 0.5°. Saccades were identified by deflections in eye position in excess of 0.1°, with a minimum velocity of 30°/s and a minimum acceleration of 8000°/s^2, maintained for at least 4 ms.

Stimuli

Fourteen puzzles were tested in a pre-study. Four of them were selected based on their similar difficulty and served as stimuli in the experiment. The puzzles consisted of photographs of natural scenes. Each puzzle was surrounded by a coordinate frame, consisting of digits (1–4) along the vertical axis and letters (A–E) along the horizontal axis. Both participants saw the puzzle and the coordinate frame throughout the whole experiment. Each puzzle piece on the expert's screen contained an additional coordinate marker (e.g., A,3), indicating the target location of the piece.

The stimuli were presented on CRT displays (19 inch Samtron 98 PDF) with a resolution of 1024 by 768 pixels at a refresh rate of 100 Hz. Puzzles were shown with a size of 935 × 704 pixels within the coordinate frame (see Figure 1). All puzzles consisted of 20 (5 × 4) pieces, each with a size of 187 × 176 pixels, corresponding to 5.8°.horizontally and 5.4°.vertically of visual angle, respectively. Pieces were positioned without overlap or free space in between. During the gaze transfer conditions, the expert's eye movements were transmitted to the novice's screen, and in the mouse transfer condition the expert's mouse movements were transmitted. Gaze and mouse cursors were visualized as a tricolor eye icon (see Figure 1), which was chosen to keep visual attributes of the cursor constant across transfer types and to assist the novice in differentiating between his own and the expert's mouse cursor in the mouse condition.

Procedure

Each pair of participants completed four blocks of different communication conditions (see below); the order of blocks was counterbalanced across participants. Each block was composed of four puzzles. We presented the same puzzles in each block to achieve a close replication of the Velichkovsky (1995) study. One (and in all blocks the same) puzzle always served as a practice trial and was not part of the analyses. The order of the three experimental puzzles within the blocks was randomized. Each block began with the practice trial followed by the experimental puzzles. Before each block, participants received a written task instruction. Subsequent to each block, participants rated the task difficulty, the efficiency of cooperation, and the ease of use of the respective communication medium.

Eye movements of the expert were recorded throughout the whole experiment, but only transferred to the novice in the respective communication conditions (see below). A nine-point calibration and validation was performed before each block and repeated between two puzzles, if necessary. To start a trial, both participants had to press the space key. A trial started with a 4 s preview of the solved puzzle, followed by the randomized arrangement of all pieces.

While the expert could not move the pieces, his task was to guide the novice's actions by the communication means of the respective block. In the *gaze* condition, the raw eye movement signal of the expert was transmitted to the novice's screen. The expert was not allowed to speak in the gaze condition, whereas the novice could speak during all conditions. In the *gaze & speech* condition, the expert's gaze was transferred and both participants could speak freely except for naming the puzzle coordinates. The *mouse & speech* condition was similar to gaze & speech but the expert's mouse movements were transferred instead of gaze. In the *speech* condition, no cursor was transferred and free verbal communication was allowed, as in gaze & speech and mouse & speech.

The novices' task was to complete the puzzle as fast and accurately as possible. Half of the novices had to follow the guidance of the expert strictly (low autonomy), while the other half were free to decide when to follow the expert's guidance (high autonomy). To move a piece, the novice clicked on it with the left mouse button and kept the

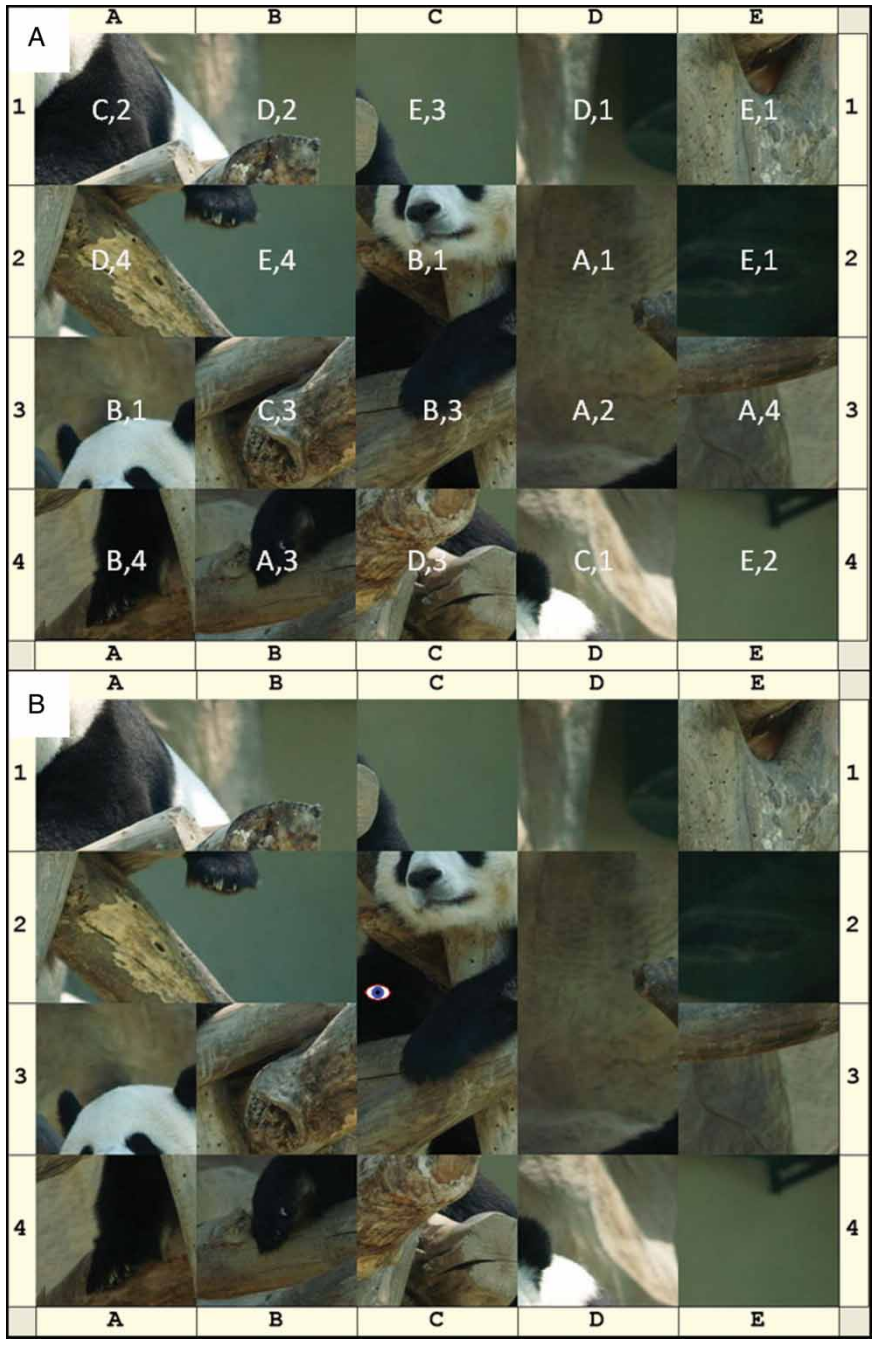

Figure 1. *Randomly arranged pieces for one puzzle as shown to the expert (A) and to the novice (B). The gaze or mouse position of the expert is indicated by an eye-icon on the novice's screen. To view a colour version of this figure, please see the online issue of the Journal.*

button pressed while dragging the piece to the new location. Once the mouse button was released, the clicked puzzle piece moved to the new position and the piece originally located at this position was swapped to the starting position of the moved piece. While the mouse button was pressed, the selected piece was superimposed by a transparent greyish layer and turned back to its original colour only when the button was released. Thus, both partners could see when a piece was selected. Also, all changes the novice made to the puzzle were visible to the expert immediately, but he did not see the novice's mouse cursor or the paths the novice took when moving a piece. A puzzle was terminated once the final piece arrived at its correct location. In total, the experiment took about one hour.

Data analysis

For the three puzzles of each block, we analysed performance measures as well as the number and quality of verbal interactions. All analyses were carried out performing 4×2 factorial repeated measures analyses of variance (ANOVAs) with communication condition (gaze, gaze & speech, mouse & speech, speech) serving as within-subjects factor and autonomy (low, high) as between-subjects factor. A time factor was not included in the final model because testing revealed that although there were significant practice effects, they did not differ between communication conditions and were therefore not relevant for the purposes of this study.[1] When additional factors were included, this information is given at the respective position in the text. Post hoc comparisons were computed using Tukey honest significant differences tests.

Due to technical problems during the recording, one performance dataset of a low-autonomy gaze & speech puzzle, one performance dataset of a high-autonomy gaze puzzle, and one verbal interaction dataset of a low-autonomy expert for gaze & speech are missing. The missing values were replaced by the mean of all data in the corresponding conditions.

RESULTS

The solution time was defined as the time between the puzzle onset and the final mouse button release. Significant main effects were found for communication condition, $F(3, 138) = 42.14$, $MSE = 2453.48$, $p < .001$, and autonomy, $F(1, 46) = 10.39$, $MSE = 4974.46$, $p = .002$. The interaction between communication condition and autonomy was marginally significant, $F(3, 138) = 2.55$, $MSE = 2453.48$, $p = .058$ (see Figure 2A). Post hoc tests revealed that the main effect for communication condition was exclusively based on the longest solution times for speech (119.1 s), in contrast to the other communication conditions ($\leq$76 s), all p values $< .001$, no further differences were found, all p values $> .1$. In the low-autonomy condition, solution times were longer compared to high autonomy (89.4 vs. 70.4 s). However, these performance benefits for high autonomy were only present in speech and in mouse & speech, both $p = .015$, but failed to reach the significance level in gaze & speech, $p = .065$ and were not present in gaze, $p = .152$.

Moving a piece to a wrong target location was defined as an error. Statistical testing revealed significant main effects for communication condition, $F(3, 138) = 11.85$, $MSE = 79.26$, $p < .001$, and autonomy, $F(1, 46) = 23.57$, $MSE = 384.48$, $p < .001$, together with an interaction for the two factors, $F(3, 138) = 4.11$, $MSE = 79.26$, $p = .008$. Post hoc analyses revealed the highest error rates for speech in contrast to the other communication conditions, all p values $< .05$. Moreover, more errors were made in gaze compared to mouse & speech, $p = .020$, while error rates were similar for gaze & speech compared to gaze as

[1] To test for practice effects, we analysed solution times separately for each blocks with regard to the order in the experiment. We found the same pattern of speech leading to slower performance than all other communication conditions, although the difference between speech and gaze & speech missed the significance level in block 4, $p = .077$. Critically, the cursor transfer conditions did not differ from each other in any block, all p values $> .3$.

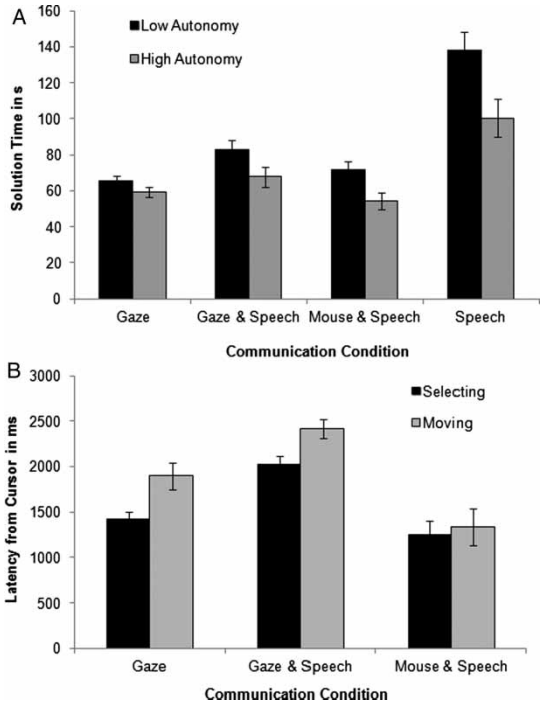

Figure 2. *Solution times for all experimental conditions (A) and latencies from first cursor reference on a piece to selecting and moving, depending on the type of cursor transfer (B). Bars represent standard errors.*

well as mouse & speech, both p values $> .2$. More errors occurred under high autonomy; this difference between autonomy levels was present in all communication conditions, all p values $< .05$, but descriptively largest when only speech was used (see Table 1).

According to the overall performance measures, the conditions gaze, gaze & speech, and mouse & speech seem to be similarly effective (cf. Figure 2A). Does this mean that novices can understand and use the expert's gaze cursor in the same way as the mouse cursor? To examine this question, situations need to be identified that require a reaction to the cursor. Second, such situations must be differentiated with regard to the necessity of a *correct* interpretation of the cursor. This is important because it is conceivable that possible interpretation difficulties of gaze transfer might be less detrimental to performance when they have no severe consequences than in situations where a misinterpretation of the partner's spatial reference is highly disrupting.

There are two such instances that require a manual reaction to the cursor but differ in the risk associated with an incorrect interpretation of the cursor signal: *selecting* the piece the expert is pointing to with his eye or mouse cursor, and *moving* a selected piece to a target location the expert is pointing to. In both cases, the novice must detect the cursor, interpret its meaning (i.e., differentiate between search and communicative purposes), and then perform the respective action. Thus, while selecting refers to the latency between the expert's cursor landing on the original position of a piece and the novice's click, moving encompasses the time between the expert's cursor landing on the destination and the novice's mouse button release. Moving a piece to the wrong place results in an error, whereas an inaccurate selection can be undone quite easily by simply releasing the mouse button. Therefore, these types of reaction differ in how strongly they presuppose an accurate interpretation of the expert's cursor.

At the same time, they require similar motor actions by the novice. He has to move the mouse to the location cued by the cursor (original position in the case of selecting and destination for moving, respectively) and then either click (selecting) or release (moving) the mouse button. Therefore, differences in the latencies from the expert's first

Table 1. *Mean values and standard deviations of error rates (%) for all experimental conditions*

	Gaze	*Gaze and speech*	*Mouse and speech*	*Speech*
Low autonomy	6.35 (5.73)	5.26 (4.92)	5.24 (3.71)	7.66 (4.74)
High autonomy	14.34 (10.42)	13.24 (10.17)	9.44 (9.62)	19.23 (8.03)

cursor landing on a piece to the novice's action (selecting vs. moving) cannot be explained by low-level processes. Instead, they should reveal differences between the two situations in the novices' strategies of using the two cursor types.

We compared the selecting and moving latencies for gaze, gaze & speech, and mouse & speech, for all occasions where a novice action was preceded by an expert cursor reference on the same piece. Thus, in the gaze transfer conditions, latencies were collected from all trials where the expert looked at the piece prior to the novice's action. Mouse latencies stem from all trials where the novice compliantly did what the mouse cursor was indicating, regardless of where the expert was looking. Since no cursors were available in speech, this condition was omitted from the analysis. Latencies were calculated as the temporal difference between the beginning of the first fixation or mouse landing on the piece to be used later within the same trial, and the action of the novice.[2] If the expert rested his cursor on a piece while waiting for the novice to finish autonomously performed actions, the latency calculation started at the beginning of the new trial (directly after the novice's previous move).

Latencies were applied to a 3 (communication condition: gaze, gaze & speech, mouse & speech) $\times$ 2 (autonomy: high, low) $\times$ 2 (reaction: selecting, moving) repeated measures ANOVA. We found significant main effects for reaction, $F(1, 46) = 13.80$, $MSE = 530468.81$, $p < .001$, and communication condition, $F(2, 92) = 26.79$, $MSE = 773864.62$, $p < .001$, together with a significant interaction for communication condition and reaction, $F(2, 92) = 6.28$, $MSE = 164005.91$, $p = .003$. No main effects or interaction including autonomy were found, all F values < 3, all p values $> .1$.

Shortest reference-to-action latencies were found for mouse & speech, followed by gaze and then gaze & speech, all p values $< .02$. Overall, latencies were shorter for reacting to a cursor reference by selecting than by moving. According to the interaction, latencies for selecting were shorter than for moving only for gaze and gaze & speech, both $p < .001$, but not for mouse & speech, $p > .9$ (see Figure 2B). Accordingly, only when the cursor indicated the expert's gaze position were the novices faster to react to it by selecting than by moving.

To determine how subject pairs used speech to establish common ground in the different communication conditions, we analysed word numbers and the specificity of verbal object references used by the experts. All non-task-related utterances were removed from the data. The subject's role in the task was introduced as a factor for the analysis of word numbers in order to consider verbal utterances of both experts and novices. Gaze-only transfer was omitted from the following analyses since the expert could not speak. Because some subjects did not speak at all (mostly novices in the low-autonomy condition), the sum of the averaged percentages can be less than 100.

First, we analysed the number of words spoken throughout the solution of a puzzle. Word numbers as a measure of verbal effort should increase when there is no visual information to enable direct grounding. Therefore, we expected to find most words in the speech condition, especially when every move needed to be instructed under low autonomy. If gaze transfer causes an increased need for disambiguation as suggested by the previous analyses, pairs might also speak more when using gaze transfer instead of the mouse.

Word numbers differed across the communication conditions, $F(2, 184) = 53.03$, $MSE = 2101.56$, $p < .001$, and role, $F(1, 92) = 135.82$, $MSE = 4254.58$, $p < .001$, but were similar for the two levels of autonomy, $F(1, 92) = .14$, $MSE = 4254.58$. Significant interactions were obtained for communication condition and role,

[2] The first cursor landing does not necessarily indicate pointing, especially during gaze transfer where visual scanning is transmitted as well. However, the expert immediately knew the correct location of a piece when looking at it due to the coordinate marks. Therefore, the need for scanning was minimal after the initial fixation. Moreover, even if scanning differed between gaze and mouse, it should be similar for selecting and moving, so that the relative effect for each communication condition should not be affected.

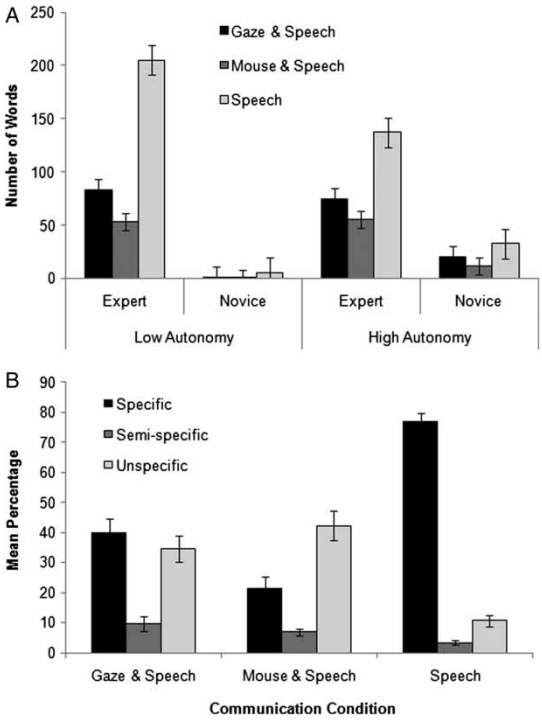

Figure 3. *Numbers of words for experts and novices (A) and specificity of the experts' referring expressions (B) in both levels of autonomy. Bars represent standard errors.*

$F(2, 184) = 34.64$, $MSE = 2101.56$, $p < .001$, and for communication condition, role, and autonomy, $F(2, 184) = 5.88$, $MSE = 2101.56$, $p = .003$ (see Figure 3A). Post hoc testing revealed more than twice as many words for speech (95.0) as for gaze & speech (44.7) and for mouse & speech (30.1), both $p < .001$. No reliable difference was found between gaze & speech and mouse & speech, $p = .068$. Experts uttered about eight times more words than novices (101.4 vs. 11.8); this difference was highest in speech. In gaze & speech and mouse & speech, the experts produced fewer words, resulting in more balanced dialogues between both partners. Under high autonomy, the number of words spoken by the experts in speech was smaller, $p < .01$; this was not the case in gaze & speech and mouse & speech, both $p > .5$.

The analysis of the reactions to the cursor suggests that novices were reluctant to take action in response to the gaze cursor, presumably due to its higher ambiguity. To test whether experts counteracted this difficulty by putting more effort into preventing ambiguity through their utterances, we analysed the precision of their verbal references to pieces and locations. If experts engage in ambiguity prevention more strongly, their object references should be more precise. Therefore, we expected them to contain the most descriptive content in speech and least in mouse & speech, with gaze & speech somewhere in between.

In the following analysis, *specific references* are instructions or comments with exact spatial descriptions (e.g., "move the third piece in the first row one position to the left and two rows down"), while *semi-specific references* give a direction but no exact position (e.g. "move the left piece upwards") and *unspecific references* do not provide any positional information (e.g. "move this one there", also known as verbal deixis). In addition, there were referential utterances describing a piece by its content or by relating it to a previous action (e.g. "the piece with the head", "the one you had before"). These utterances (4.4 %) were equally distributed across the experimental conditions and strongly varied in their level of precision, so they were not considered in this analysis. We performed a 3 (communication condition: gaze & speech, mouse & speech, speech) × 2 (autonomy: low, high) × 3 (specificity: specific, semi-specific, unspecific) repeated measures ANOVA. Only effects related to specificity are reported.

There was a main effect of specificity, $F(2, 92) = 56.15$, $MSE = 1000.31$, $p < .001$, and a significant interaction for communication condition and specificity, $F(4, 184) = 59.11$, $MSE = 425.92$, $p < .001$. No further interactions with respect to specificity were obtained, all F values < 3, all p values $> .1$, which reveals that there were no differences regarding the level of autonomy. There were more specific verbal references (46.0%) than unspecific (29.1%) and semi-specific references (6.6%). This distribution strongly varied across the communication conditions: In gaze & speech and mouse & speech we found fewer specific but more unspecific references than in speech, all p values $< .001$. The number of semi-specific

references was similar across the communication conditions, all p values $> .8$ (see Figure 3B).

To examine further whether the two types of cursor transfer had a differential effect on the expert's choice of referential form, we conducted planned comparisons between communication conditions with cursor transfer (gaze & speech, mouse & speech) and specificity (specific, unspecific). The significant interaction, $F(1, 46) = 11.32$, $MSE = 743.57$, $p = .002$, demonstrated that more unspecific than specific references were used in mouse & speech, $p < .001$, as opposed to gaze & speech where there was no difference, $p > .9$.

DISCUSSION

To better understand how a partner's gaze is used to achieve mutual understanding during remote cooperation, we systematically compared the effects of gaze versus mouse transfer in a puzzle task. Specifically, we asked whether there was a cost attached to receiving a high-resolution display of the partner's visual attention in a task that requires spatial referencing. Pairs of participants—an expert and a novice—jointly solved puzzles under two levels of autonomy. The communication conditions for transmitting the expert's knowledge to the novice were gaze, gaze & speech, mouse & speech, or speech.

As expected, gaze transfer as compared to purely verbal interaction revealed faster performance, fewer errors and a reduction of verbal effort, as evidenced in shorter dialogues and less specific object descriptions. So far, our results confirm earlier findings by Velichkovsky (1995) and demonstrate the beneficial function of gaze transfer in remote cooperation. Furthermore, the present findings go beyond previous studies in demonstrating that the usability of gaze transfer can be affected by the need for close cooperation. For increased novice autonomy, enabling him to perform easy moves on his own accord, we obtained improved performance speed except in the gaze transfer condition. One possible explanation for this result could be that the constantly and unpredictably moving gaze cursor led to an automatic attention capture (see for instance Mulckhuyse, van Zoest, & Theeuwes, 2008), thereby binding partners together regardless of task affordances. This assumption is supported by additional analyses which indicate that during gaze transfer the degree of cooperation was barely affected by the autonomy manipulation.

Our main interest concerned the comparison between gaze and mouse transfer with regard to the cooperative process. As expected, the overall performance was quite similar for both forms of cursor transfer: Solution times were within the same range and error rates were only higher for gaze than for mouse transfer when the former was used without concurrent speech. The absence of a solution time difference in favour of gaze is not trivial when considering the timing of information transmission: Gaze cursors necessarily arrive earlier at the target location than the mouse since targets must be seen by the expert before he can manually point to them. The lack of faster performance despite this temporal advantage indicates difficulties on the novice's side when interpreting the gaze cursor.

In order to understand these difficulties, one must consider the information contained in gaze transfer and its relation to the requirements of the task at hand. While a gaze cursor provides a high-resolution display of the partner's visual attention and search process, the puzzle task merely requires the explicit indication of particular positions which the novice has to interpret appropriately. Eye movements, and fixations in particular, can reflect multiple cognitive functions, varying with respect to their mode of processing and the underlying mechanisms (Velichkovsky, 2002). Thus, in the puzzle task the transmitted gaze behaviour of the expert only sometimes serves the required indicating function, and often it does not. Therefore, an extra effort from the novice is necessary to decide whether a fixation is meant for pointing or merely reflects the ongoing search for a target.[3] In contrast, the mouse serves only a

[3] An analysis of the experts' eye movements revealed that they tried to prevent such ambiguities by adapting temporal and spatial parameters of their gaze when using it for communication. They fixated longer and on fewer pieces, and especially tried to keep their

single function and can therefore be used instantly. A similar phenomenon has been observed when applying gaze as an input device in human computer interaction (Majaranta & Räihä, 2002) and is referred to as the *Midas touch problem* (see Jacob, 1991; Velichkovsky, Sprenger, & Unema, 1997).

The assumption of gaze-related uncertainty is supported by the comparison of the novices' selecting and moving reactions. First, novices generally reacted faster to the expert's mouse cursor than to his gaze. More importantly, they reacted faster to gaze cursors by selecting a piece than by moving a piece to the cursor location, whereas no difference between reaction types was found for the mouse cursor. The difference for gaze cursors may be related to the costs of an erroneous interpretation. Selecting the wrong piece causes almost no costs, as the mouse button can be released and nothing has changed. In contrast, moving a piece to the wrong position produces an error that requires further actions for correction, and is therefore more costly. This higher impact of misinterpretations may have urged novices to collect more evidence for the expert's communicative intention. Such cost-related procrastination in our novices' reactions suggests that they had trouble when dealing with the ambiguity inherent in gaze transfer. Thus, gaze transfer appears to be more usable in situations where an accurate interpretation of the cursor signal is less crucial.

In turn, these difficulties in the joint process of communication require partners to adapt their communicative strategies to avoid ambiguity and ensure mutual understanding. Accordingly, an increase in conversational effort in the gaze conditions was observed. The trend towards more words in gaze as compared to mouse transfer did not quite reach significance, but evidence for more careful interactions stems from the precision experts chose when verbally referring to objects. Visual information about the partner's attentional engagement can enable a more direct grounding, reducing the need for cumbersome verbal descriptions (Clark & Brennan, 1991). In our study, this was reflected in a decreased precision of verbal references for both types of cursor transfer. However, there also was a difference between them, with specific references being used more frequently during gaze transfer.

This finding corresponds with results from studies of pragmatic tools in cognitive linguistics, showing that the choice of a referential form—the decision about referring to an object with a specific description versus using a shorter pronoun—depends on the clarity of the linguistic context (Chafe, 1976) and can be predicted from the features of this context (Kibrik, Dobrov, Loukachevitch, & Zalmanov, 2010). Referential choice is also sensitive to the ambiguity and salience of the objects within the visual scene (Ferreira, Slevc, & Rogers, 2005; Fukumura, van Gompel, & Pickering, 2010). When less salient referents are present, specific descriptions are preferred, whereas for highly salient referents, unspecific reference is used to avoid additional working memory load (Almor, 1999) and to minimize joint effort when establishing common ground (Clark, 1996). In turn, if the specificity of referential choice is related to the perceived need for clarification, our results imply that pairs using gaze transfer found it rather necessary to prevent ambiguity.

Additional analyses revealed a higher amount of verbal feedback during gaze than mouse transfer, indicating that partners provided more explicit evidence when the interpretability of the implicit evidence from the cursor was not warranted. Taken together, our analyses of the cooperative process indicate that establishing common ground was harder when using gaze as compared to mouse transfer. Novices seemed uncertain about the cursor intention and, consequently, pairs in general and experts in particular adapted to the situation by communicating more precisely and explicitly.

Some of the confusion related to gaze transfer might be reduced through technical modifications

gaze still during periods when the novice needed to interpret it. However, controlling one's eye movements can severely impair performance (Ballard, Hayhoe, & Pelz, 1995), which implies that although such strategies may ameliorate some problems of gaze transfer, they can give rise to others.

of the cursor signal. In particular, a temporal filter averaging over several gaze samples (Helmert, Pannasch, & Velichkovsky, 2008) or a threshold-based transmission depending on certain gaze parameters such as fixation durations is conceivable. However, with regard to our initial question about the effect of eye movement information on the grounding process, the utility of gaze transfer for spatial referencing *per se* needs to be reconsidered. In our paradigm, all differences between gaze and mouse transfer manifested in gaze costs, indicating that all useful information was available through pointing alone. Additional cues to the partner's visual attention did not make communication any easier, but markedly impaired the interpretation of the cursor. Thus, our findings qualify the generally positive evaluation of gaze transfer as a cooperative tool that can be found in the literature (e.g., Neider et al., 2010).

However, it is important to note that the lack of gaze over mouse benefits in our particular task does not mean that these benefits cannot be present in other contexts. Jointly solving puzzles may simply not require a detailed and continuous attention monitoring, whereas in other situations this information might well be useful. Here we can learn a lesson from human computer interaction, where gaze input also turned out to be problematic when used as an explicit command for discrete actions (Jacob, 1991). Just like in our study, interpreting the purpose of a given fixation was challenging and made performance unstable. Therefore, gaze input has been applied for more implicit forms of control, using eye movement parameters as additional cues to interpret gestural or verbal reference (Kaur et al., 2003; Koons, Sparrell, & Thorisson, 1993). Similarly, gaze behaviour as a reliable indicator of a person's interest has been used to determine the successive content of interactive films or storylines (Hansen, Andersen, & Roed, 1995; Starker & Bolt, 1990; Vesterby et al., 2005).

One might conceive of similar applications for gaze in cooperative contexts. Ideally, this should be tasks where the precise focus of a person's attention is of direct relevance to an observer, for instance when trying to understand a partner's search strategies, implicit knowledge, or interests. Indeed, novices can benefit from having watched expert gaze recordings in different training scenarios (Litchfield, Ball, Donovan, Manning, & Crawford, 2008; Mehta, Sadasivan, Greenstein, Gramopadhye, & Duchowski, 2005; Stein & Brennan, 2004). In co-present situations, pioneering investigations used gaze transfer as a window to a listener's interest and thereby supported speakers in appropriately choosing and abandoning topics in a conversation (Qvarfordt, Beymer, & Zhai, 2005).

Another note of caution is required when interpreting the present findings: They do *not* show that a transmission of information about a partner's attention is irrelevant altogether in spatial referencing tasks. Instead, it can be questioned whether this information should be available on a microscopic level such as single fixations. In a broader sense, attention information is present throughout each interaction, because people usually attend to the objects they are talking about (Griffin, 2004) and acting on (Land, Mennie, & Rusted, 1999). Therefore, even speech, gestures, and actions are informative of a partner's attention. Moreover, people can use their mouse to indicate reliably where they are looking when this is necessary during cooperation (Müller, Helmert, Pannasch, & Velichkovsky, 2011).

In conclusion, we demonstrated that gaze transfer can help establish common ground in remote cooperation. However, when used to convey spatial information it also has costs in terms of a higher ambiguity than conventional referencing devices. In order to utilize the full potential of gaze as a tool for communication, task characteristics need to be considered more thoroughly, especially with regard to the necessity of attention information. Our findings suggest that only the information directly relevant to the task should be transmitted, unless practical constraints make it necessary to do otherwise. Whether gaze transfer is a part of this should be determined with respect to the way grounding takes place in a particular context.

Manuscript received 23 May 2012
Revised manuscript received 14 September 2012
First published 13 November 2012

REFERENCES

Almor, A. (1999). Noun-phrase anaphora and focus: The informational load hypothesis. *Psychological Review*, *106,* 748–765.

Ballard, D. H., Hayhoe, M. M., & Pelz, J. B. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, *7,* 66–80.

Brennan, S. E., Chen, X., Dickinson, C. A., Neider, M. B., & Zelinsky, G. J. (2008). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition*, *106,* 1465–1477.

Bruner, J. (1981). The pragmatics of aquisition. In W. Deutsch (Ed.), *The child's construction of language* (pp. 42–64). New York, NY: Academic Press.

Carletta, J., Hill, R. L., Nicol, C., Taylor, T., de Ruiter, J. P., & Bard, E. G. (2010). Eyetracking for two-person tasks with manipulation of a virtual world. *Behavior Research Methods*, *42,* 254–265.

Chafe, W. L. (1976). Givenness, contrastiveness, definiteness, subjects, topics and point of view. In C. N. Li (Ed.), *Subject and topic* (pp. 25–55). New York, NY: Academic Press.

Clark, H. H. (1996). *Using language*. Cambridge, UK: Cambridge University Press.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). Washington, DC: APA Books.

Doherty-Sneddon, G., Anderson, A., O'Malley, C., Langton, S., Garrod, S., & Bruce, V. (1997). Face-to-face and video mediated communication: A comparison of dialogue structure and task performance. *Journal of Experimental Psychology: Applied*, *3,* 105–125.

Ferreira, V. S., Slevc, L. R., & Rogers, E. S. (2005). How do speakers avoid ambiguous linguistic expressions? *Cognition*, *96,* 263–284.

Fukumura, K., van Gompel, R. P. G., & Pickering, M. J. (2010). The use of visual context during the production of referring expressions. *Quarterly Journal of Experimental Psychology*, *63,* 1700–1715.

Griffin, Z. M. (2004). Why look? Reasons for eye movements related to language production. In J. M. Henderson & F. Ferreira (Eds.), *The integration of language, vision, and action: Eye movements and the visual world* (pp. 213–247). New York, NY: Taylor and Francis.

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, *11,* 274–279.

Hanna, J. E., & Brennan, S. E. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, *57,* 596–615.

Hansen, J. P., Andersen, A. W., & Roed, P. (1995). Eye-gaze control of multimedia systems. In Y. Anzai, K. Ogawa, & H. Mori (Eds.), *Advances in human factors/ergonomics: Symbiosis of human and artifact – Future computing and design for human–computer interaction, Proceedings of the Sixth International Conference on Human–Computer Interaction* (Vol. 20, pp. 37–42). Amsterdam: Elsevier Science B.V.

Helmert, J. R., Pannasch, S., & Velichkovsky, B. M. (2008). Influences of dwell time and cursor control on the performance in gaze driven typing. *Journal of Eye Movement Research*, *2,* 1–8.

Jacob, R. J. K. (1991). The use of eye movements in human–computer interaction techniques: What you look at is what you get. *ACM Transactions on Information Systems (TOIS)*, *9,* 152–169.

Kaur, M., Tremaine, M., Huang, Wilder, J., Gacovski, Z., Flippo, F., et al. (2003). *Where is "it"? Event synchronization in gaze–speech input systems*, Paper presented at the 5th International Conference on Multimodal Interfaces, New York, NY.

Kibrik, A. A., Dobrov, G. B., Loukachevitch, N. V., & Zalmanov, D. A. (2010). *Referential choice as a probabilistic multi-factorial process*, Paper presented at the Fourth International Conference on Cognitive Science, Tomsk, Russia.

Koons, D. B., Sparrell, C. J., & Thorisson, K. R. (1993). Integrating simultaneous input from speech, gaze, and hand gestures. In M. Maybury (Ed.), *Intelligent multimedia interfaces* (pp. 257–276). Menlo Park, CA: MIT Press.

Land, M. F., Mennie, N., & Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception*, *28,* 1311–1328.

Litchfield, D., Ball, L. J., Donovan, T., Manning, D. J., & Crawford, T. (2008). Learning from others: Effects of viewing another person's eye movements while searching for chest nodules. In B. Sahiner & D. J. Manning (Eds.), *Proceedings of SPIE Medical Imaging 2008: Image Perception, Observer Performance, and Technology Assessment*. San Diego, CA.

Majaranta, P., & Räihä, K.-J. (2002). *Twenty years of eye typing: Systems and design issues*, Paper presented at the Eye Tracking Research & Applications Symposium, ETRA '02, New York, NY.

Marshall, C. R., & Novick, D. G. (1995). Conversational effectiveness in multimedia communications. *Information Technology & People*, *8,* 54–79.

Mehta, P., Sadasivan, S., Greenstein, J., Gramopadhye, A. K., & Duchowski, A. T. (2005). *Evaluating different display techniques for communicating search strategy training in a collaborative virtual aircraft inspection environment*. Paper presented at the Human Factors and Ergonomics Society Annual Meeting, Orlando, FL.

Monk, A. F., & Gale, C. (2002). A look is worth a thousand words: Full gaze awareness in video-mediated conversation. *Discourse Processes*, *33,* 257–278.

Mulckhuyse, M., van Zoest, W., & Theeuwes, J. (2008). Capture of the eyes by relevant and irrelevant onsets. *Experimental Brain Research*, *186,* 225–235.

Müller, R., Helmert, J. R., Pannasch, S., & Velichkovsky, B. M. (2011). *Following closely? The effects of viewing conditions on gaze versus mouse transfer in remote cooperation*. Paper presented at the European Conference on Computer-Supported Cooperative Work, ECSCW '11, Aarhus, Denmark.

Neider, M. B., Chen, X., Dickinson, A., Brennan, S. E., & Zelinsky, G. J. (2010). Coordinating spatial referencing using shared gaze. *Psychonomic Bulletin & Review*, *17,* 718–724.

Qvarfordt, P., Beymer, D., & Zhai, S. (2005). *RealTourist – A study of augmenting human–human and human–computer dialogue with eye–gaze overlay.* Paper presented at the Human–Computer Interaction Conference, INTERACT '05, Rome, Italy.

Richardson, D. C., Dale, R., & Kirkham, N. Z. (2007). The art of conversation is coordination: Common ground and the coupling of eye movements during dialogue. *Psychological Science*, *18,* 407–413.

Starker, I., & Bolt, R. A. (1990). *A gaze-responsive self-disclosing display.* Paper presented at the Conference on Human Factors in Computing Systems, SIGCHI '90, New York, NY.

Stein, R., & Brennan, S. E. (2004). *Another person's eye gaze as a cue in solving programming problems.* Paper presented at the 6th International Conference on Multimodal Interfaces, State College, PA.

Velichkovsky, B. M. (1995). Communicating attention: Gaze position transfer in cooperative problem solving. *Pragmatics and Cognition*, *3,* 199–222.

Velichkovsky, B. M. (2002). Heterarchy of cognition: The depths and the highs of a framework for memory research. *Memory*, *10*(5/6), 405–419.

Velichkovsky, B. M., Sprenger, A., & Unema, P. J. A. (1997). *Towards gaze-mediated interaction: Collecting solutions of the "Midas touch problem*". Paper presented at the Human–Computer Interaction Conference, INTERACT '97, London, UK.

Vertegaal, R., Velichkovsky, B. M., & Van der Veer, G. C. (1997). Catching the eye: Management of joint attention in cooperative work. *ACM SIGCHI Bulletin*, *29,* 87–92.

Vesterby, T., Voss, J. C., Hansen, J. P., Glenstrup, A. J., Hansen, D. W., & Rudolph, M. (2005). Gaze-guided viewing of interactive movies. *Digital Creativity*, *16,* 193–204.

Whittaker, S. (1995). Rethinking video as a technology for interpersonal communications: Theory and design implications. *International Journal of Human–Computer Studies*, *42,* 501–529.