Acta Psychologica 212 (2021) 103222

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy

Crossmodal correspondences as common ground for joint action

Laura Schmitz[a,g], Günther Knoblich[a], Ophelia Deroy[b,c,d], Cordula Vesper[a,e,f,*]

[a] Department of Cognitive Science, Central European University, Budapest, Hungary
[b] Faculty of Philosophy, Ludwig-Maximilians-Universität, Munich, Germany
[c] Munich Centre for Neuroscience, Ludwig-Maximilians-Universität, Munich, Germany
[d] Institute of Philosophy, School of Advanced Study, University of London, London, UK
[e] Department of Linguistics, Cognitive Science and Semiotics, Aarhus University, Aarhus, Denmark
[f] Interacting Minds Centre, Aarhus University, Aarhus, Denmark
[g] Institute for Sports Science, Leibniz Universität Hannover, Hannover, Germany

ARTICLE INFO

Keywords:
Joint action
Communication
Common ground
Crossmodal correspondences
Social cognition
Multisensory perception

ABSTRACT

When performing joint actions, people rely on *common ground* – shared information that provides the required basis for mutual understanding. Common ground can be based on people's interaction history or on knowledge and expectations people share, e.g., because they belong to the same culture or social class. Here, we suggest that people rely on yet another form of common ground, one that originates in their similarities in multisensory processing. Specifically, we focus on 'crossmodal correspondences' – nonarbitrary associations that people make between stimulus features in different sensory modalities, e.g., between stimuli in the auditory and the visual modality such as *high-pitched* sounds and *small* objects. Going beyond previous research that focused on investigating crossmodal correspondences in individuals, we propose that people can use these correspondences for communicating and coordinating with others. Initial support for our proposal comes from a communication game played in a public space (an art gallery) by pairs of visitors. We observed that pairs created nonverbal communication systems by spontaneously relying on 'crossmodal common ground'. Based on these results, we conclude that crossmodal correspondences not only occur *within* individuals but that they can also be actively used in joint action to facilitate the coordination *between* individuals.

1. Introduction

Joint actions come in countless different forms and complexities. At first glance, a simple handshake does not have much in common with an expertly synchronized swimming performance nor with an improvised dinner party. However, what co-actors in all types of joint actions rely on is the fact that they share *something*: Individuals "cannot even begin to coordinate […] without assuming a vast amount of shared information or common ground" (Clark & Brennan, 1991, p. 222; cf. also Brennan & Hanna, 2009; Clark, 1996, p. 199; Lewis, 1969; Schelling, 1960; Stalnaker, 2002). Common ground is crucial to get a joint action started; at the same time, new common ground builds up between co-actors with every joint action they perform.

1.1. Common ground

According to Clark (1996, 2006), one can broadly distinguish two different types of common ground: *Communal* common ground is based on cultural communities such as nationality, ethnic group, occupation or gender, whereas *personal* common ground is based on people's joint experience. For example, when going for a bike ride with a friend, we ride on the right side of the street because of a culturally shared norm; my friend takes the lead because our joint experience has taught us that she knows her way around the city better than I do.

When performing a joint action, the extent of common ground between co-actors may be minimal (e.g., when two strangers share nothing but a joint goal; cf. Vesper, Butterfill, Knoblich, & Sebanz, 2010) or it may be extensive (e.g., think of two synchronized swimmers who have trained together for years and also share their private lives as a married couple). Co-actors may rely on common ground implicitly or they may decide to make (specific parts of) their common ground mutually manifest (cf. Sperber & Wilson, 1995). Importantly, what is (or is not) common ground between two people is to a certain extent subjective as "we are in fact acting on our individual beliefs or assumptions about

[*] Corresponding author at: Department of Linguistics, Cognitive Science and Semiotics, Aarhus University, Jens Chr. Skous Vej 4, 8000 Aarhus C, Denmark.
E-mail address: cvesper@cc.au.dk (C. Vesper).

https://doi.org/10.1016/j.actpsy.2020.103222
Received 3 February 2020; Received in revised form 25 September 2020; Accepted 5 November 2020
Available online 7 December 2020

what is in our common ground" (Clark, 1996, p.96; also see Wilkes-Gibbs & Clark, 1992). Thus, individuals engaged in joint action will act upon what they *think* is common ground in the specific interaction and they will therefore *expect* that their behavior is comprehensible for their interaction partners. This is most obviously the case in conversations where speakers adjust to the expected shared background with their addressees, for example by simplifying word choices when interacting with children (Brennan & Hanna, 2009; Clark & Krych, 2004; Horton, 2007; Horton & Keysar, 1996; Keysar, Barr, Balin, & Brauner, 2000; Lockridge & Brennan, 2002). It is, however, possible that an individual's belief about what is common ground in an interaction is actually false and that they will thus not be understood (or misunderstood) by their interaction partner – it is only when both partners' beliefs about their common ground coincide that the interaction will be successful. Taken together, common ground supports social interaction in various forms and contexts and is, therefore, a decisive factor for the success of any joint action.

1.2. Communication

When common ground between co-actors is limited or completely absent, the interaction is prone to go awry. One way to quickly establish common ground, e.g., when agents do not share a visual context, is by using (verbal) communication in its function as a coordination device (Brennan, 2005; Clark & Kruch, 2004; Duff, Hengst, Tranel, & Cohen, 2006; Scott-Phillips, 2015; Vesper et al., 2010). It is even possible that language has evolved precisely for the purpose of facilitating joint action, as it allows, for example, to solve coordination problems efficiently over time and distances. At the same time, (verbal) communication is itself a joint action as it requires coordination of meaning between sender(s) and receiver(s) (Clark, 1996; Duff et al., 2006; Galantucci, 2009; Shintel & Keysar, 2009).

If, however, there is no shared language and hence no (linguistic) convention to rely on, how do co-actors achieve common ground and, thus, coordinate their actions? Studies in the field of 'experimental semiotics' have shown that when people need to coordinate their actions but cannot rely on conventional language or other forms of pre-established common ground, they might spontaneously invent a set of symbols and thereby bootstrap a novel communication system (e.g. de Ruiter, Noordzij, Newman-Norlund, Hagoort, & Toni, 2007; Galantucci, 2005; Scott-Phillips, Kirby, & Ritchie, 2009). Relatedly, research on 'sensorimotor communication' (Pezzulo et al., 2019; Pezzulo, Donnarumma, & Dindo, 2013; Vesper & Sevdalis, 2020) has demonstrated that when conventional communication is not feasible, people might systematically adjust particular movement parameters of their instrumental actions such that they violate an observer's motor prediction (Becchio, Sartori, Bulgheroni, & Castiello, 2008; Hommel, Müsseler, Aschersleben, & Prinz, 2001; Knoblich & Jordan, 2003; Prinz, 1997; Sebanz, Bekkering, & Knoblich, 2006; Wolpert, Doya, & Kawato, 2003). In doing so, they send nonverbal, communicative signals which in turn facilitate interpersonal coordination (Candidi, Curioni, Donnarumma, Sacheli, & Pezzulo, 2015; Sacheli, Tidoni, Pavone, Aglioti, & Candidi, 2013; Schmitz, Vesper, Sebanz, & Knoblich, 2018a; Vesper & Richardson, 2014; Vesper, Schmitz, & Knoblich, 2017). Together, these findings suggest that communication naturally emerges out of the need to interact and coordinate actions with others (Clark, 1996; Galantucci, 2005).

In contrast to the research on experimental semiotics and sensorimotor communication, which explores how communication systems emerge *from scratch* or piggyback on instrumental movements, we focus here on communication that builds on *pre-existing common ground*. Our proposal is twofold:

(1) Firstly, we suggest that so-called "crossmodal correspondences" – a perceptual phenomenon that has been identified and extensively explored by researchers in the field of multisensory perception (Spence, 2011; Spence & Deroy, 2013) – provide a form of pre-existing common ground between people.
(2) Secondly, we suggest that people can actively employ this 'crossmodal common ground' for communication. Communication, in turn, will allow for successful joint action.

In the present study, two co-actors coordinate their actions towards a joint goal or outcome (Knoblich, Butterfill, & Sebanz, 2011; Vesper et al., 2010). Critically, they need to communicate *in order* to achieve their joint goal. In turn, they achieve their joint goal *through* the act of communicating, i.e., by successfully coordinating meaning. Thus, coordination and communication are tightly intertwined in the present context: Co-actors' interaction is a joint action because of the presence of a joint goal and an act of communication because it involves an explicit communicative intent (cf. Scott-Phillips et al., 2009).

In the following, we will provide a brief background on crossmodal correspondences and then spell out our hypotheses about how these correspondences can serve to support joint action.

1.3. Crossmodal correspondences

Crossmodal correspondences are stable associations that people make between stimulus features in different sensory modalities, most commonly between stimuli in the auditory and the visual modality. For example, people associate *high-pitched* sounds with *small, light* objects (Evans & Treisman, 2010; Gallace & Spence, 2006; Spence & Parise, 2012) and *low-pitched* sounds with *big, dark* objects (Klapetek, Ngo, & Spence, 2012; Marks, 1987). Research has shown that crossmodal correspondences are already experienced early in development (Dolscheid, Shayan, Majid, & Casasanto, 2013; Walker et al., 2010) and that people make such associations a) automatically (Spence & Deroy, 2013) and b) both implicitly and explicitly (cf. Spence, 2011).

Four different kinds of crossmodal correspondences have been distinguished, namely those based on *statistical* regularities found in nature (e.g., between the size of an object and its resonance frequency; Bee, Perrill, & Owen, 2000; Coward & Stevens, 2004), those based on *structural* associations (e.g., between magnitude-related stimuli features such as size and loudness; Smith & Sera, 1992; Walsh, 2003), those that are *semantically mediated* (e.g., between auditory pitch and visual elevation; Martino & Marks, 1999; also see Walker, Francis, & Walker, 2010), and those that are *hedonically mediated* (e.g., between the taste and shape of food; Parise, Spence, & Deroy, 2016; Velasco, Woods, Deroy, & Spence, 2015). Research has shown that these four kinds of correspondences differ in terms of their qualitative experience and developmental trajectory, as well as in how they affect human information processing (cf. Spence, 2011).

As regards the function of crossmodal correspondences, it has been suggested that they facilitate the crossmodal binding problem, i.e., they help us to decide whether incoming unisensory signals originate from the same or different sources – and thus whether they should be bound together, or integrated, in the brain (Ernst, 2007; Spence, 2011; Spence, Ngo, Lee, & Tan, 2010). In other words, when no other contextual cues are available, one may "by default" rely on crossmodal associations. For example, I may rely on the above-mentioned association between pitch and size to determine whether the *low-pitched* sounds I'm hearing are emitted by the *small* or the *big* frog I'm observing at the pond.

1.4. Crossmodal common ground

Based on our proposal that multisensory processing can offer a basis for communication and joint action, the aim of the present proof-of-concept study was to test the following concrete hypothesis: Crossmodal correspondences provide a form of common ground ('crossmodal common ground') that can be actively used to facilitate joint action. More specifically, when joint action partners need to communicate to facilitate coordination yet cannot use conventional language, they will rely on crossmodal correspondences. In doing so, they will implicitly act on the assumption that these associations are part of the sum of information they in fact share with each other,[1] i.e., they will assume these associations to be part of the *common ground* between them (Clark, 1996, 2006). This assumption, in turn, will make them expect communication to be successful.

To test our hypothesis, we created a communication game to be played jointly by two participants – a 'sender' and a 'receiver' – in a public space at an art gallery (see Fig. 1). In the game, the sender's task was to inform the receiver about a particular feature of a visual stimulus (i.e., about its size, brightness or color). The sender, however, was not allowed to use speech or gesture to communicate, but was asked to pick particular stimuli in the auditory (or olfactory) modality as a communication medium. In particular, the sender could play piano tones of different auditory pitch for the receiver to hear or she could present different scents for the receiver to smell. Thus, for example, the sender could communicate to the receiver about the particular size of a visual stimulus by playing a tone of a particular height on the piano. The receiver, in turn, needed to match the height of the tone she heard to one of the differently sized visual stimuli she saw on a computer screen (see Fig. 1A). That is, the receiver needed to "translate" the auditory stimulus feature back into a visual stimulus feature.

We designed three different versions of this game; each version was based on one specific crossmodal correspondence that has been established in the literature. In particular, we relied on the correspondences between (1) audition [pitch] and vision [size] (Evans & Treisman, 2010; Gallace & Spence, 2006; Spence & Parise, 2012); (2) audition [pitch] and vision [brightness] (Klapetek et al., 2012; Marks, 1987; Martino & Marks, 1999; Melara, 1989); and (3) olfaction and vision [color] (Demattè, Sanabria, & Spence, 2006, 2009; Deroy, Crisinel, & Spence, 2013; Gilbert, Martin, & Kemp, 1996; Kemp & Gilbert, 1997; Österbauer et al., 2005). We tested each version in a separate experiment. In all three experiments, the visual modality served as the "referent" for communication (i.e., what to communicate *about*) whereas the "communication medium" (i.e., what to communicate *with*) was either audition (Experiments 1 and 2) or olfaction (Experiment 3).

The central aim of these experiments was to test whether co-actors would spontaneously rely on crossmodal correspondences to create communicative mappings to facilitate joint action. In particular, we predicted that senders would map stimuli features in the two different sensory modalities (e.g., auditory tones of different pitch and visual stimuli of different size) in line with the nonarbitrary crossmodal correspondences documented in the literature. For example, we expected senders to match *high-pitched* sounds with *small* circles and *low-pitched* sounds with *large* circles – rather than the opposite (i.e., *high-pitched* to *large* and *low-pitched* to *small*) or any other possible combination (Hypothesis 1a). Thus, we predicted that senders would rely on (their individual) crossmodal correspondences when creating communicative signals, i.e., that they would use these correspondences for a novel purpose, namely the purpose of communication. If receivers, for their part, also (implicitly) assumed that crossmodal correspondences would form the basis for the mappings created by senders, then they should be able to interpret the communicative signals accurately. This mutual understanding between senders and receivers would consequently lead to successful joint task performance (Hypothesis 1b).

A further research question we addressed in our study was concerned with how co-actors, when faced with the challenge of establishing a communication system, would deal with ambiguity with regard to the means of communication. Specifically, we asked whether, when faced with two alternative crossmodal mapping options, co-actors would manage to align their expectations and choose the same mapping – and which mapping option they would choose.

We predicted that co-actors, once they had established a communication system based on particular crossmodal mappings and thereby created a new communicative convention, would stick to their established system even if a second option to communicate becomes available (Hypothesis 2a). In particular, faced with the two options, senders now needed to choose whether to rely on the mapping that had been jointly established during the preceding interaction in part one of the experiment ('History') or to create a different communication system based on the newly available mapping option ('Novelty'). For communication to succeed, receivers' expectations needed to correspond to the senders' choice; close alignment of their expectations should therefore lead to successful joint task performance (Hypothesis 2b).

2. Methods

We conducted our study at the Tate Modern art gallery in London, UK, within the context of the public engagement series "Tate Exchange" on April 29–30, 2017. The aim of Tate Exchange is to provide a "space for everyone to make, play, talk, and reflect and to discover new perspectives on life, through art"[2] and it regularly offers workshops, talks and events for the general public – in cooperation with artists, professionals, and scientists. The general idea was to promote scientific research and to facilitate the exchange between scientists and the public, allowing a wide audience to explore, interact and discover how sensory experiences shape our world.

For the present research, this context had three main implications: First, it created an environment that prompted openness, curiosity, and exploratory behavior in our participants. All our participants were museum visitors who volunteered to take part in our research because of their own intrinsic motivation to experience something new. Second, most of our participants came together as families or couples, leading to a higher-than-usual proportion of participants who were familiar with each other. Third, conducting our research in a museum environment meant that our experiments had to be relatively short to remain interesting and understandable without long instructions. We accommodated our methodology to reflect these constraints and opportunities. In the following, we report the methods used for all our three experiments and highlight the ways in which the individual experiments differed from each other.

2.1. Participants

Participants were visitors at Tate Modern who gave their informed consent for their anonymized data to be used for research purposes. Our experiment had been approved by the museum. We asked visitors to participate in pairs of two. Most people arrived jointly in pairs (7 of 12 pairs in Experiment 1; 12 of 12 pairs in Experiment 2; 6 of 8 pairs in Experiment 3) and thus pair members knew each other beforehand; all other individuals were randomly matched with an unfamiliar partner. Demographic information can be found in Table 1.

2.2. Stimuli

2.2.1. Basic experimental setup

The setup for our communication game is shown in Fig. 1. It consisted of the following equipment. An ASUS laptop running Matlab (2015) was used to control the experimental procedure and record the data. The visual stimuli for senders were shown on the laptop's screen (resolution: 1280 × 720 pixels) placed on a normal-sized table. To present the visual stimuli to receivers and record their responses, an LCD

[1] This is also in line with Grice's idea that speakers are cooperative when they rely on the information they share with their addressee (Grice, 1975). Thus, conversations "are characteristically, to some degree at least, cooperative efforts" (Grice, 1975, p. 26).

[2] https://www.tate.org.uk/visit/tate-modern/tate-exchange (retrieved on January 26, 2020)

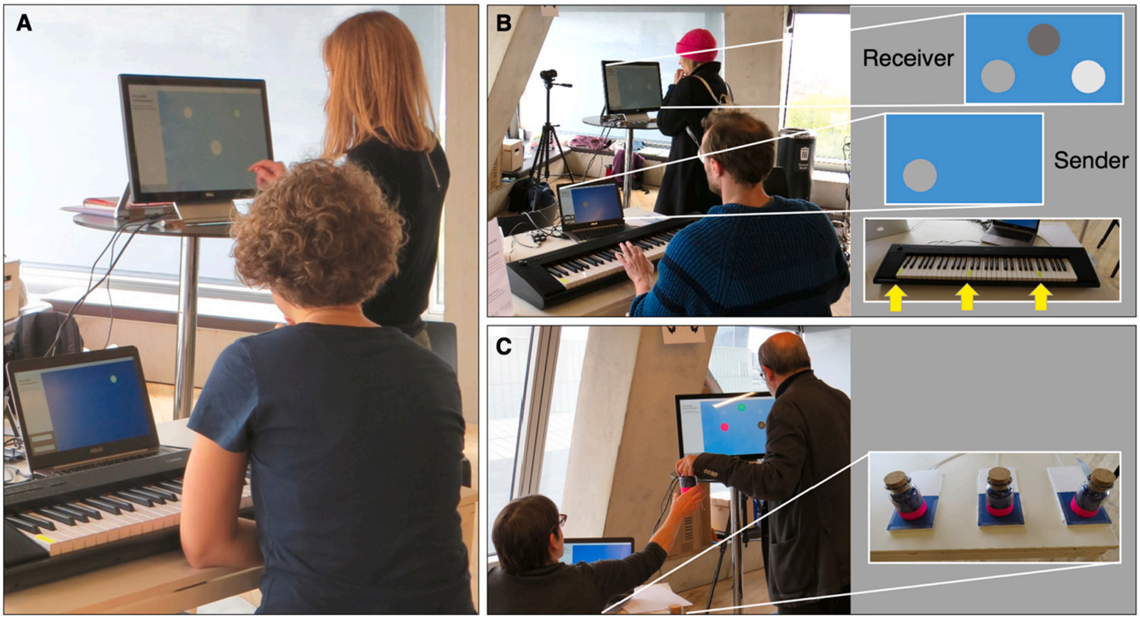

Fig. 1. Examples of our communication game where participant pairs communicated “crossmodally”. The ‘sender’ (sitting) relied on auditory pitch (by playing different piano tones) to communicate about visual stimuli of different size (**A**, Experiment 1) or of different brightness (**B**, Experiment 2); the ‘receiver’ (standing) needed to select the matching stimulus on the touch screen. (**C**) In Experiment 3, the sender relied on olfactory stimuli (by picking different scents of tea in opaque bottles) to inform the receiver (who could open the bottles to smell the scents) about visual stimuli of different colors that the receiver needed to select on the touch screen.

Table 1
Demographic information of participants across the three experiments.

Experiment	Senders		Receivers	
	Gender	Age	Gender	Age
1 Pitch & size	5 M, 7 F	32.8 (19–54)	4 M, 8 F	30.3 (20–57)
2 Pitch & brightness	5 M, 7 F	30.5 (23–42)	6 M, 6 F	29.2 (23–42)
3 Odor & color	2 M, 6 F	45.5 (26–64)	6 M, 2 F	40.4 (18–66)

touch screen (resolution: 1280 × 720 pixels) was used. It was placed on a high table at a height comfortable for adults to stand at and was located in a position that prevented receivers from seeing the senders’ screen (which was placed on the left behind receivers’ back), see Fig. 1.

2.2.2. Communication medium

Depending on the particular experiment, different technology triggered and recorded senders’ “communicative signals”. In Experiments 1 and 2, a portable digital piano (Yamaha Piaggero NP12) was used as the communication medium. Three keys on the piano (a low C2 with midi code 36, a middle C4 with midi code 60 and a high C6 with midi code 84) were marked with bright yellow stickers to facilitate participants’ choice. The other keys on the piano remained functional. The data from the piano were exported in midi format, containing information about which key was pressed, when and how long it was pressed, and what force was used to press it. If senders pressed more than one piano key in a trial (which occurred only in 11% of all trials of all three experiments combined), we analyzed the key they pressed last because this was typically the one that receivers acknowledged and based their response upon.

In Experiment 3, custom-made force plates (using an Arduino Mega 2560 microcontroller) provided a binary signal indicating whether an object was placed on each plate or not. This signal was used to record, for each trial, which instance of the communication medium – in this case odor bottles – senders picked up and handed over to receivers who then smelled the scent. The odor stimuli were three teas with strong, distinct scents: black tea (with bergamot scent), fruit tea (with cherry scent) and mint tea (with mint scent). All teas had a similar dark color and were filled into small opaque glass bottles. An easy-to-open lid allowed participants to smell the different scents (Fig. 1C).

2.2.3. Referent for communication

The referents for communication were visual stimuli that varied along one particular dimension (size, brightness or color, depending on the experiment; see Table 2). In addition to the one-dimensional stimulus features used in all experiments, participants in Experiments 1 and 2 received a further ‘two-dimensional’ set of stimuli in a second part of the experiment. These two-dimensional stimuli (Table 2) varied along two perceptual feature dimensions, i.e., in both size and brightness. Importantly, the stimuli were designed such that the mappings for the two dimensions were not congruent, e.g., the *large* circle was *white* such that its size corresponded to a low-pitched tone (large > low pitch) yet its brightness corresponded to a high-pitched tone (bright > high pitch). Thus, depending on the chosen stimulus dimension (size or brightness), the resulting crossmodal mapping would differ, allowing us to determine whether participants used the same mapping as in the first, one-dimensional part of the experiment or whether they established a different mapping.

2.3. Procedure

The experimenter introduced the task (“you will be playing a communication game together”) and informed each person about their respective role (“one of you will be the sender who sends information to the other person; one of you will be the receiver whose job it is to understand the signals they receive”) and that their communication would have to be non-verbal (“importantly, you are not allowed to speak with each other or to use gestures – so you will have to invent a new, non-verbal language!”).

Participants then performed two randomly chosen training trials to get familiarized with the task procedure (guided by the experimenter and by short written instructions on the computer screens), after which they performed 15 experimental trials (corresponding to the one-dimensional part). Each trial started with the presentation of the

Table 2
Visual stimulus properties along with the predicted crossmodal mappings (*in italics*). In the two-dimensional part of Experiments 1 and 2, two alternative predictions could be made, as participants could either stick to the same crossmodal mappings as in the one-dimensional part ('History') or switch to a different mapping ('Novelty').

Experiment	Stimulus 1	Stimulus 2	Stimulus 3
1 Pitch & size, one-dimensional			
Size (Ø in cm)	Large (15.0)	Medium (11.25)	Small (7.5)
Brightness (RGB)	Bright (255, 255, 255)	Bright (255, 255, 255)	Bright (255, 255, 255)
Pitch	*Low*	*Medium*	*High*
1 Pitch & size, two-dimensional			
Size (Ø in cm)	Large (15.0)	Medium (11.25)	Small (7.5)
Brightness (RGB)	Bright (255, 255, 255)	Dark (77, 77, 77)	Medium (166, 166, 166)
Pitch (History)	*Low*	*Medium*	*High*
Pitch (Novelty)	*High*	*Low*	*Medium*
2 Pitch & brightness, one-dimensional			
Size (Ø in cm)	Large (15.0)	Large (15.0)	Large (15.0)
Brightness (RGB)	Dark (133,133,133)	Medium (171, 171, 171)	Bright (230,230,230)
Pitch	*Low*	*Medium*	*High*
2 Pitch & brightness, two-dimensional			
Size (Ø in cm)	Medium (10.0)	Small (5.0)	Large (15.0)
Brightness (RGB)	Dark (133, 133, 133)	Medium (171, 171, 171)	Bright (230, 230, 230)
Pitch (History)	*Low*	*Medium*	*High*
Pitch (Novelty)	*Medium*	*High*	*Low*
3 Odor & color, one-dimensional			
Size (Ø in cm)	Large (15.0)	Large (15.0)	Large (15.0)
Color (RGB)	Black (0,0,0)	Green (18, 212, 133)	Pink (255, 0, 102)
Odor	*Bergamot*	*Mint*	*Cherry*

visual stimuli, i.e., three circles differing in size (Experiment 1), brightness (Experiment 2), or color (Experiment 3) appeared on the sender's and receiver's screens (identical images were shown on both screens). To avoid any implicit ordering of the stimuli on the screen (e.g., as could happen if presented in a row), stimuli were arranged in an isosceles triangle, where the respective position of the three stimuli randomly changed from trial to trial. After 2 s, two of the three circles on the sender's screen vanished and only the target stimulus remained; the receiver's screen remained unchanged. Senders were instructed to communicate the identity of this target stimulus to receivers by pressing a key on the keyboard (Experiments 1 and 2) or handing over one of the three odor bottles (Experiment 3). Based on this communicative signal, the receiver then chose a visual target on his / her screen by touching it. If the correct target was chosen, both sender and receiver got immediate positive feedback in the form of a green circle around the chosen target (or a happy smiley on the target in Experiment 3). If the incorrect target was chosen, the feedback was negative, indicated by a red circle (or an unhappy smiley in Experiment 3). The feedback stayed on the screen and, after 1 s, a "continue" button appeared on the receiver's screen. The receiver was asked to touch this button once both participants were ready for the next trial.

When participants in Experiments 1 and 2 had finished the first, one-dimensional part, the experimenter asked whether they would be willing to perform a second part of the game. All participants agreed. Participants were informed by the experimenter that the objects they now needed to communicate about looked slightly different than in the first part. Participants then continued with the two-dimensional part of the experiment, in the same sender-receiver role distribution.

Once participants had finished the full experiment, they were asked to complete a short debriefing questionnaire. Besides giving basic demographic information, they were asked to rate the confidence they had experienced during the game. Specifically, the participant in the role of the sender was asked "How confident were you that your task partner would correctly interpret your signals?" and the participant in the role of the receiver was asked "How confident were you that you correctly interpreted your task partner's signals?". Answers were given on a 5-point Likert scale ranging from "not at all" to "very much".

Each experiment took about 10 min in total.

2.4. Data analysis

For all experiments, we report 1) the sender's mapping score and 2) the dyad's joint match accuracy. The first measure indicates whether senders based their communicative mappings on crossmodal correspondences. The second measure indicates whether receivers matched the two stimuli dimensions in the same way senders did; hence, it effectively represents the pair's joint coordination performance. Specifically, joint match accuracy was computed as the percentage of trials in which the receiver chose the correct target stimulus, i.e., all trials where receiver and sender relied on the same mapping.

To calculate the sender's mapping score, we extracted the sender's responses from the piano or force sensor. The raw values of all participants and experiments are publicly available on https://doi.org/10.17605/OSF.IO/J9PTC. To compute an overall mapping score, we counted the trials in which senders created mappings in line with the crossmodal correspondences documented in the literature. We then calculated the proportion of these trials of the total number of trials. This procedure is exemplified in the following equation for the pitch & size mapping in Experiment 1:

$$Mapping\ score = \frac{\left(N_{[low\,pitch->big]} + N_{[medium\,pitch->medium]} + N_{[high\,pitch->small]}\right)}{N_{big} + N_{medium} + N_{small}} \times 100$$

In the numerator, the number of trials where the senders chose the predicted crossmodal mappings are summed up, i.e., all trials where they mapped low pitches to big targets, medium pitches to medium-sized targets, and high pitches to small targets. In the denominator, the total number of trials is summed up, i.e., all trials with big, medium-sized, and small targets (total $N = 15$). The mapping score is now computed by dividing the number of trials with predicted mappings by the total number of trials (and multiplying it by 100 to get a score in percentage). For example, the sender in the pair whose raw data is presented in Fig. 2A consistently matched *big* targets to *low-pitched* tones, *medium-sized* targets to *medium-pitched* tones and *small* targets to *high-pitched* tones in all 15 trials, leading to a mapping score precisely according to our predictions: ((5 + 5 + 5) / (5 + 5 + 5)) *100 = 100%. In contrast, if, for example, the sender had matched the *big* target to a *low-pitched* tone only 2 out of 5 times, this would result in a lower mapping score: ((2 + 5 + 5) / (5 + 5 + 5)) *100 = 80%. Generally, if senders matched all stimuli according to our predictions, the resulting mapping score would be 100%; if they never matched accordingly, the score would be 0; and if they randomly chose their mappings, the score would be around 33%. Thus, the mapping score reflects how much senders use the *predicted* mapping in comparison to *any other* mapping, irrespective of what this other mapping was, that is, whether it was a random choice or a mapping that is consistent but systematically different from our predicted crossmodal mapping.[3] A sender could, for example, consistently choose to match small stimuli to medium-high pitches, medium-sized stimuli to low pitches, and large stimuli to high pitches and would still only get a mapping score of 0% because the chosen mapping is not what we expected based on our hypothesis about crossmodal correspondences as common ground.

To test our first research question, we used the data from the one-

[3] In Figures 3 and 4, we present the proportion of trials where participants used crossmodal mappings ('Crossmodal') and the proportion of trials where they used any other, not predicted type of mappings ('Other'). Together, these proportions add up to 100%.

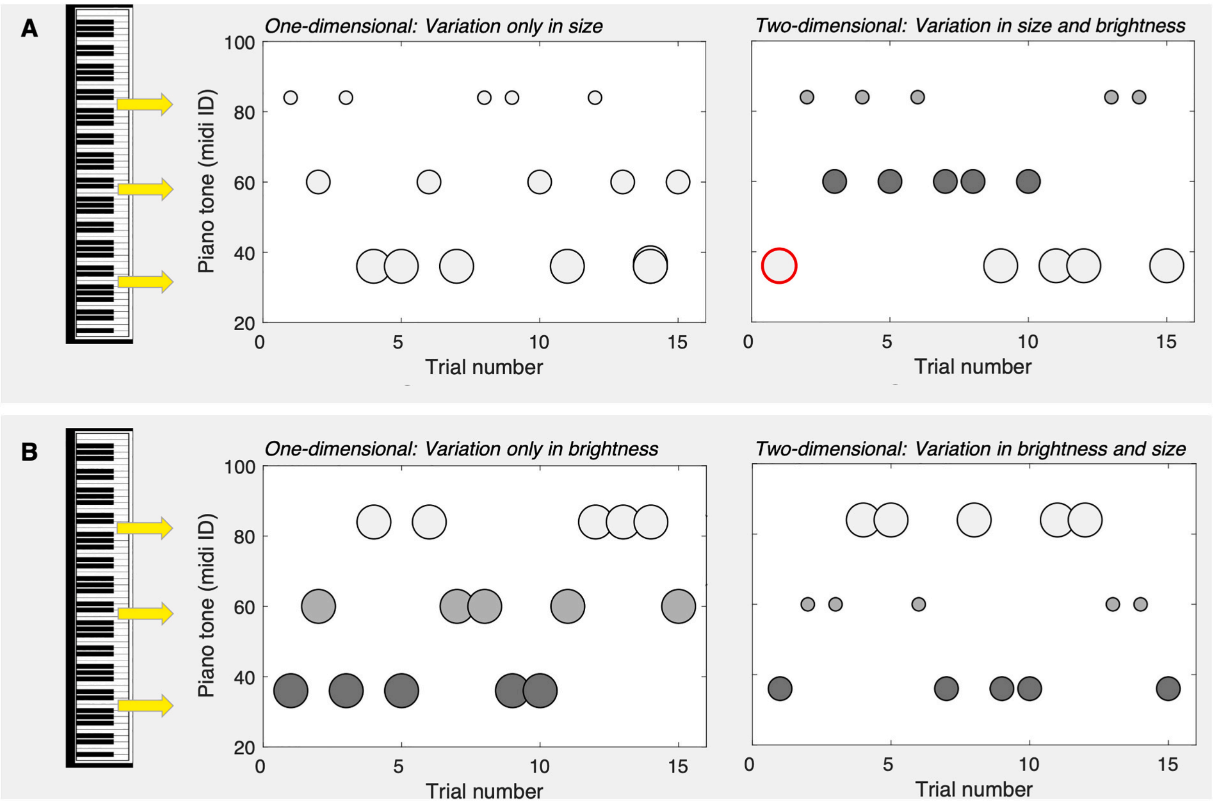

Fig. 2. Exemplary communicative signals from two senders in (A) Experiment 1 (pitch & size) and (B) Experiment 2 (pitch & brightness). The piano tones played by these senders are shown as a function of the visual target stimuli they were communicating about. The yellow arrows indicate the marked piano keys and the corresponding midi IDs. The trial highlighted with a red circle shows a joint match error, where, apparently, the sender continued to use the established crossmodal mapping (pitch & size) in the two-dimensional part, whereas the receiver did not. However, the pair immediately recovered from this initial misalignment of expectations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

dimensional part of all experiments and statistically compared senders' mapping scores to a chance level of 33% using one-sample *t*-tests. If senders' mapping scores were significantly above chance, this would indicate that they consistently mapped the sensory stimuli according to the predicted crossmodal correspondences (Hypothesis 1a). We also compared the dyads' joint match accuracies to a chance level of 33% using one-sample *t*-tests. If the joint match accuracies were significantly above chance, this would indicate that senders and receivers relied on the same crossmodal mapping (Hypothesis 1b).

To test our second research question, we used the data from the second, two-dimensional part of Experiments 1 and 2 and statistically compared two alternative mapping scores using paired-samples *t*-tests. The two scores reflected the two different crossmodal mapping options, namely the option that had already been available in the previous, one-dimensional part of the experiment (e.g., mapping *high-pitched* to *small* in Experiment 1; 'History') and the option that had become newly available in the two-dimensional part thanks to the added second stimulus dimension (e.g., mapping *high-pitched* to *bright* in Experiment 1; 'Novelty'). Thus, we tested which of the two alternative crossmodal mappings senders used more frequently (Hypothesis 2a). In addition, as in the one-dimensional part, we compared the dyads' joint match accuracies to a chance level of 33% to test whether senders and receivers relied on the same crossmodal mapping (Hypothesis 2b).

Data analysis and statistical testing were performed using customized R scripts (R Core Team, 2020).

3. Results

3.1. Crossmodal common ground for communication

To investigate Hypothesis 1a, we analyzed senders' mapping scores in the one-dimensional part of all experiments (Fig. 3A) and compared them to a chance level of 33%. In Experiment 1 (Pitch & size), we found that the score of 88.3% was significantly higher than chance, $t(11) = 9.165$, $p < .001$, Cohen's d = 2.65.[4] Similarly, the mapping score of 96.7% in Experiment 2 (Pitch & brightness) was significantly higher than chance, $t(11) = 22.878$, $p < .001$, Cohen's d = 6.6. This was also the case in Experiment 3 (Odor & color), where we computed a mapping score of 93.3%, $t(7) = 19.55$, $p < .001$, Cohen's d = 6.91.

The analysis of the dyads' joint match accuracy (Hypothesis 1b) indicated that participants' joint performance was significantly better than chance in all experiments. In Experiment 1, the accuracy was 87%, $t(11) = 14.264$, $p < .001$, Cohen's d = 4.12.; in Experiment 2, it was 89%, $t(11) = 16.462$, $p < .001$, Cohen's d = 4.75; in Experiment 3, it was 90%, $t(7) = 9.048$, $p < .001$, Cohen's d = 3.2.

In the debriefing questionnaire, we asked participants to rate on a scale from 1 ("not at all") to 5 ("very much") how confident they had felt

[4] We also performed a post-hoc analysis including only the five pairs in Experiment 1 where partners did *not* know each other. The mapping score was significantly different from chance, $t(4) = 39.396$, $p < .001$, Cohen's d = 17.62, replicating our finding also for this small sub-group of unfamiliar partners. This suggests that the successful usage of crossmodal correspondences for communication does not depend on prior familiarity between interaction partners. Further experiments are needed to test the potential role of familiarity more systematically.

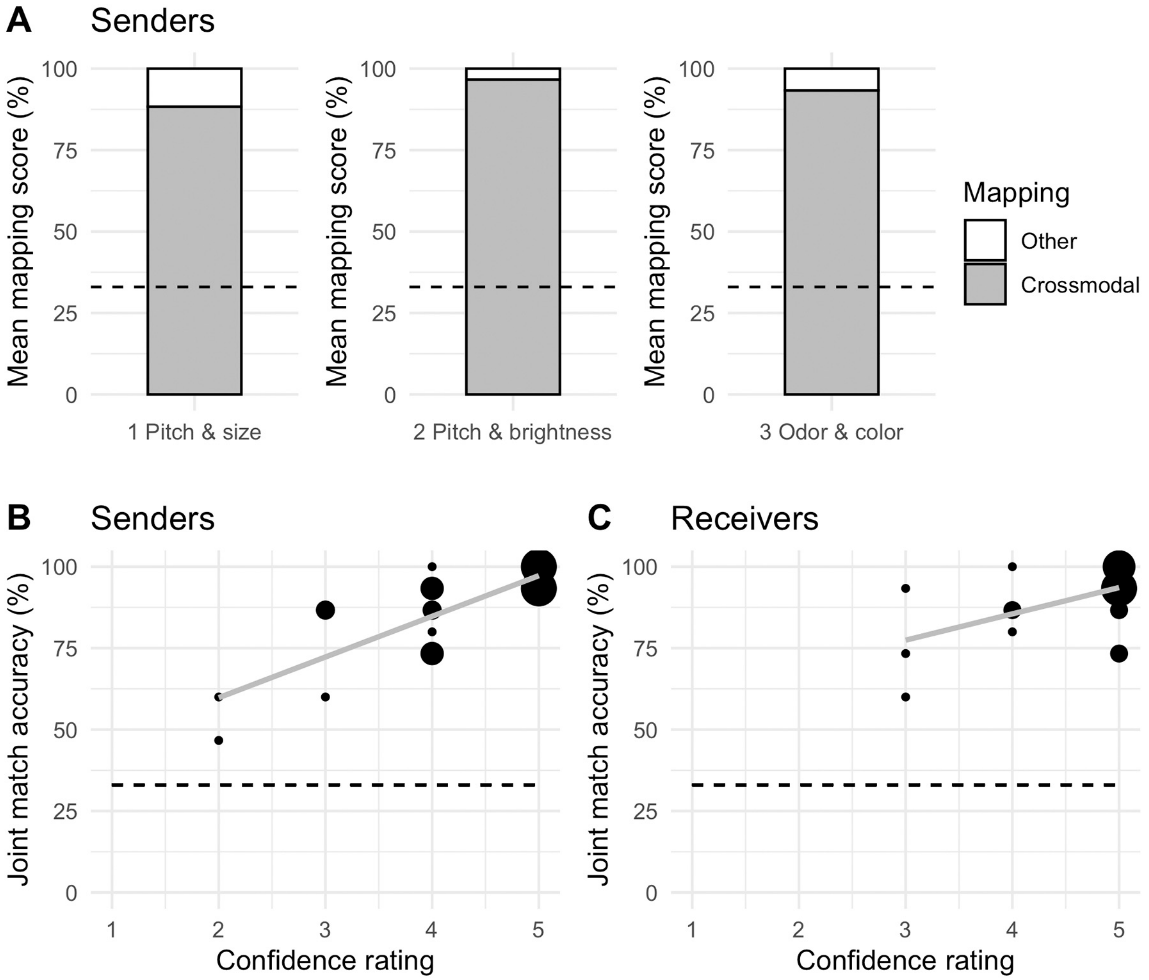

Fig. 3. (A) Senders' mapping scores for the first one-dimensional part of all three experiments. The dashed lines represents a chance mapping score of 33%. Dark grey (Crossmodal) shows the percentage of trials in which senders used the predicted crossmodal mapping; white (Other) shows the percentage of trials where senders used any other type of mapping. (B) Senders' and (C) receivers' confidence ratings as a function of joint performance represented by joint match accuracy (aggregated data from all experiments). Dot size represents number of participants; the grey lines show linear regression lines, and the dashed lines represent accuracy at chance level.

about their way of communicating. The reported confidence ratings were overall high. Aggregated over all three experiments, senders were highly confident (4.29 on average) that receivers would understand their signals. Receivers, in turn, were also highly confident (4.5 on average) that they correctly interpreted the senders' signals. To check whether senders' and receivers' confidence was related to the dyads' actual joint performance, we performed correlation analyses of confidence and joint match accuracy. We found high correlation coefficients for both senders, $r = 0.824$, $p < .001$ (Fig. 3B), and receivers, $r = 0.751$, $p < .001$ (Fig. 3C).

3.2. *Common ground and ambiguity*

To investigate Hypothesis 2a, we analyzed senders' mapping scores in the two-dimensional part of Experiments 1 and 2. As in the one-dimensional part, we first tested whether senders relied on crossmodal correspondences by comparing their mapping scores to a chance level of 33%. In particular, we initially wanted to verify whether senders relied on *any* type of crossmodal mapping at all, and thus we used a mapping score that combined the two alternative crossmodal mapping options that were available in the two-dimensional part ('History' + 'Novelty'). In Experiment 1, this combined mapping score of 92.2% was significantly higher than chance, $t(11) = 19.404$, $p < .001$, Cohen's d = 5.6. Also in Experiment 2, the combined mapping score of 97.2% was significantly above chance, $t(11) = 23.12$, $p < .001$, Cohen's d = 6.67.

To determine which of the two alternative crossmodal mappings senders prioritized, we then directly compared the two mapping scores as this comparison would indicate which mapping was used more frequently (Fig. 4). Senders' mapping scores in Experiment 1 (Pitch & size) were significantly higher for the pitch-to-size mapping that they had also used in the previous, one-dimensional part ('History'), compared to the pitch-to-brightness mapping that became available in the two-dimensional part ('Novelty'), $t(11) = 3.943$, $p < .01$, Cohen's d = 2.26. This finding was replicated in Experiment 2 (Pitch & brightness), where the interaction history (i.e., using a pitch-to-brightness mapping) dominated over the newly available pitch-to-size mapping, $t(11) = 2.592$, $p < .05$, Cohen's d = 1.49.

An analysis of the dyad's joint match accuracy in the two-dimensional parts (Hypothesis 2b) showed that, for the most part, receivers matched the two stimuli dimensions in the same way senders did, leading to high accuracy. Joint match accuracy was significantly higher than chance in both Experiment 1 with an accuracy of 71%, $t(11) = 5.214$, $p < .001$, Cohen's d = 1.51, and in Experiment 2 with an accuracy of 83%, $t(11) = 7.386$, $p < .001$, Cohen's d = 2.13.

4. Discussion

The proposal we put forward here is to consider 'crossmodal common

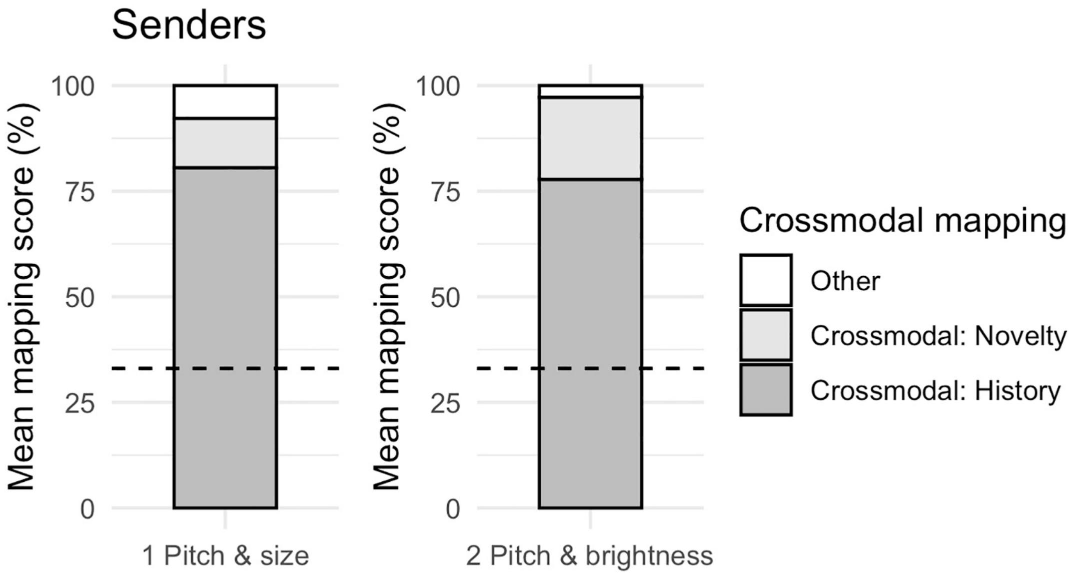

Fig. 4. Two different mapping scores are shown for the two-dimensional part of Experiments 1 and 2: The mapping score for 'History' (dark grey) reflects how often senders used the same mapping as in the previous, one-dimensional part of the experiment; the mapping score for 'Novelty' (light grey) reflects how often senders created a mapping based on the newly introduced stimulus dimension. For completeness, 'Other' (white) is shown, representing all those alternative mapping combinations that do not reflect any predicted crossmodal correspondence. The dashed lines represent a chance mapping score of 33%.

ground' as a basis for communication and joint action. We regard this form of common ground as complementary to other previously discussed forms such as *communal* common ground, based on societal norms and conventions, or *personal* common ground, based on joint experience (Clark, 1996; Lewis, 1969; Stalnaker, 2002). What sets crossmodal common ground apart from these and other forms of common ground is that it is available to groups of people who have entirely different backgrounds (e.g., do not speak the same language) and who have not previously interacted. Instead, crossmodal common ground can facilitate interaction between multiple people simply because they share the same perceptual apparatus and therefore process multisensory information in comparable ways. In fact, more generally, there seems to be a basic, overarching common ground provided to all of us by virtue of our shared perceptual capacities: "All of us take as common ground, I assume, that people normally have the same senses, sense organs, and types of sensations." (Clark, 1996, p. 106). Crossmodal common ground may be described as a special subset of our perceptual common ground because it is not based on shared "types of sensations" but rather on shared *associations between* those sensations. Crossmodal common ground is not present because we perceive a specific stimulus feature *in the same way* but because we make *the same association* between *this* stimulus feature and specific *other* features in different sensory modalities – and because we are mutually aware that we do so. In other words, if we expect that all (or at least, most) people associate *small* objects with *light* colors and *high-pitched* sounds, we can assume this association as common ground and make corresponding predictions about others' behavior which, in turn, facilitates joint action.

4.1. Empirical evidence for crossmodal common ground

As a proof of concept, we tested this general proposal in a set of three short experiments within the context of a public engagement event in an art museum. Pairs of museum visitors participated in a communication game, in which one person (the 'sender') informed their partner (the 'receiver') about different visual stimuli. Instead of using conventional forms of communication such as speech or gesture, however, our task required participants to spontaneously create a communication system using one sensory modality as the referent (e.g., *visual* stimuli of different size to communicate *about*) and another as the communication medium (e.g., piano tones of different *auditory* pitch to communicate *with*). Thus, for example, the sender could communicate to the receiver about the particular size of a visual stimulus by playing a tone of a particular height on the piano. The receiver needed to match the height of the tone she heard to one of the differently sized visual stimuli she saw on a computer screen.

The aim of a first 'one-dimensional' part of our communication game was to test whether co-actors would spontaneously rely on crossmodal correspondences to create communicative mappings. Crucially, in order for communication to succeed, this required, first, that senders created communicative signals based on crossmodal mappings, assuming that receivers would understand these, and second, that receivers interpreted the communicative signals accordingly. Confirming our hypotheses, we found overall high 'mapping scores' for senders (Hypothesis 1a) and high 'joint match accuracies' for the interacting dyads (Hypothesis 1b), indicating that both senders and receivers consistently matched visual stimuli with auditory (Experiments 1 and 2) or olfactory (Experiment 3) stimuli in line with crossmodal correspondences.

The consistent use of crossmodal correspondences for communication we observed in all experiments was further confirmed by participants' self-reports: When we asked senders after the experiment how confident they had felt that their communicative signals would be understood by their task partner, and receivers how confident they had felt that they interpreted the provided communicative signals correctly, most participants gave high ratings, which also correlated with how well they actually managed to coordinate (Fig. 3**B** and **C**). This indicates that, regardless of whether participants created or interpreted the communicative signals, they confidently relied on crossmodal correspondences and assumed that these would hold in the same way for their respective task partner.

Notably, our results clearly show that participants did not first *acquire*, or learn, crossmodal mappings. If these mappings had to be newly learned during the experiment, then most pairs should have started with several unsuccessful trials, before eventually converging on the same mapping. However, we find that most pairs are successful right from the start or improve very quickly. In addition, the fact that crossmodal mappings were consistently used by the majority of participants confirms that participants did not simply learn any new mapping. If this were the case, any other mapping system (not based on crossmodal correspondences) should have been as likely to be learned as the crossmodal mapping. However, we find a prevalence for the mapping that is based on crossmodal correspondences, indicating that 'crossmodal common ground' was already there and participants did not just learn it while performing the experimental task.

Still, despite these clear findings supporting our central hypothesis about crossmodal common ground, it is interesting to note that not all dyads were equally successful in playing the communication game. For example, one dyad experienced difficulties because the sender did not provide consistent communicative signals but unsystematically matched *big* targets with *high-*, *medium-*, and *low-pitched* tones. Consequently, the receiver, who depended on the sender's communicative signals, made a couple of 'match errors', until the sender's signals became more reliable. This failure can be attributed to a general problem underlying the representation of others and their potentially differing perspective (e.g., Keysar et al., 2000; Keysar, Lin, & Barr, 2003). In particular, previous research has demonstrated the crucial need for consistency and alignment of representation in communication. For instance, in Galantucci's seminal study in the field of experimental semiotics (Galantucci, 2005), participant dyads attempted to create novel communication systems from scratch; however, whereas most succeeded, participants in one dyad failed to converge on the same system even after more than two hours of interacting because the meaning of the signs created by the sender were ambiguous and not used consistently. This highlights that even if participants can learn through feedback, as in Galantucci's study, joint convergence is not trivial.

Such examples vividly demonstrate how senders' and receivers' expectations need to be aligned for communication to succeed. Importantly, senders need to be aware of the receivers' background (that might differ from their own) and need to carefully monitor the receivers' behavior in order to adjust their communicative signals if necessary, just as in conventional linguistic communication (Brennan & Hanna, 2009; Brennan, Kuhlen, & Charoy, 2018; Clark & Kruch, 2004; Lockridge & Brennan, 2002). In the present study, senders designed their communicative signals based on the (implicit) assumption that crossmodal correspondences would provide a background shared by receivers. This awareness about an interaction partner's task, perspective, constraints etc. is not restricted to communicative interactions but a common feature of many joint actions (Curioni, Vesper, Knoblich, & Sebanz, 2019; Schmitz, Vesper, Sebanz, & Knoblich, 2017, 2018b; Sebanz et al., 2006; Surtees, Apperly, & Samson, 2016; Vesper et al., 2010).

4.2. Turning crossmodal into personal common ground

In a second 'two-dimensional' part of our communication game, we followed up on the idea of partner-specific communication, or audience design. Our aim was to probe co-actors' communicative preferences by exposing them to an ambiguous situation: In contrast to the first part, where the stimuli serving as referents for communication varied only in one stimulus dimension (e.g., only in size) and were thus 'one-dimensional', the stimuli in the second part varied in two dimensions (e.g., in size and in brightness) – they were 'two-dimensional'. This created a source of ambiguity for co-actors because senders now needed to decide whether to create crossmodal mappings based on the same stimulus dimension as in the first part, using the already established communication system ('History'), or based on the newly introduced stimulus dimension ('Novelty'). Our findings show that co-actors had a clear preference for relying on the established communication system, as indicated by senders' overall high 'mapping scores' for the previously used crossmodal mapping (Hypothesis 2a) and by dyads' high 'joint match accuracies' (Hypothesis 2b). This finding is interesting for at least two reasons:

First, it demonstrates how crossmodal common ground – a form of common ground that does, importantly, not require personal acquaintance – can turn into personal common ground through repeated use within a joint action, or an interactive grounding process (cf. Clark & Brennan, 1991). In other words, something that began with a general expectation about the similarities in multisensory processing in other people quickly became a part of the particular interaction history of a particular sender-receiver-dyad. Future research needs to explore this finding further by systematically manipulating the type and amount of shared knowledge between different interaction partners to distinguish senders' own preferences from their assumptions about their particular partners' expectations. Moreover, a careful control of how familiar participants are with each other is desirable as in the present study, we could not entirely exclude the possibility that participants might rely on personal common ground that they had established before taking part in our research. However, it is rather unlikely that prior familiarity affected the present results, in particular, as a post-hoc analysis of our data demonstrated that also the pairs consisting of unfamiliar individuals consistently relied on crossmodal correspondences.

Second, the finding points to a system in human communication that helps resolve ambiguous situations. It thereby complements our earlier work that showed a similarly high consistency in co-actors' implicit preference for a particular way of communicating. Participant dyads in that study, when faced with two different options for communication – a sensorimotor communication strategy, in which their instrumental goal was combined with the communicative goal, and a communicative strategy, in which they could separate their instrumental and communicative action goals – quickly showed agreement as to which option to use (Vesper et al., 2017). Thus, in many cases of joint action, co-actors not only seem to have similar expectations about how to approach their coordination problem, but also which approach to give precedence in cases of ambiguity.

Moreover, our present finding that co-actors prefer to rely on their joint interaction history is in line with Clark's "Principle of joint salience". This principle states that when faced with a coordination problem, people should ideally strive for the solution that is most salient with respect to their current common ground, i.e., for what is *jointly* most salient for them (Clark, 1996). By doing so, they reduce the chances of miscoordination. Further theoretical support for our finding comes from another of Clark's principles, namely the "Principle of least collaborative effort" (Clark & Brennan, 1991; also cf. Grice, 1975). It says that people in a conversation try to minimize their collaborative effort, i.e., the combined effort of both speaker and addressee. When applying this principle to the ambiguous situation in our study, one should predict that dyads will continue to use the communication system they had already established rather than creating a different system. The former clearly creates the least collaborative effort. Participants in our study acted accordingly by choosing to continue with their established way of communicating.

4.3. Theoretical backdrop and future directions

With regard to the theoretical foundations of the present study, it is worth returning to the general concept of communication upon which our proposal about crossmodal common ground is based. We rely on the understanding that communicative behavior can be characterized as involving communicative intentions (cf. Scott-Phillips et al., 2009). Following Sperber and Wilson's (1995) influential account of (ostensive) communication, which is built around the expression and recognition of intentions, we can distinguish *informative intentions* (i.e., the intention to convey information to one's addressee) and *communicative intentions* (i.e., the intention to make one's informative intention known to one's addressee). In our experiments, the roles of 'sender' and 'receiver' were pre-defined and the two participants were explicitly instructed to communicate. Thus, they were mutually aware that the sender had an informative intention and that this intention was known by the receiver, i.e., the sender's intent to communicate was already explicit. Hence, the sender did not have to demonstrate ostensively to the receiver that she was trying to communicate; the receiver, in turn, did not have to recognize the sender's behavior as communicative (cf.

Scott-Phillips et al., 2009).

Our experiments show that by using crossmodal correspondences consistently and repeatedly for communication, participants created a communicative convention. According to many researchers in the field (e.g., Galantucci, 2009; Misyak, Noguchi, & Chater, 2016; Tamariz, Roberts, Martínez, & Santiago, 2018), the emergence of conventions is one of the characteristics of communication systems; languages might even be referred to as a set of "conventional codes" (Scott-Phillips, 2015). In our study, participants established – over time and via the repeated usage of crossmodal mappings for a communicative purpose – a new conventional code between them. Whereas crossmodal correspondences *as such* pre-existed, the *use* of crossmodal correspondences *as* communicative mappings did not and thus newly emerged into a communicative convention. This goes beyond previous research which has shown that individuals make stable associations between stimulus features in different sensory modalities by showing that individuals also *expect others* to share these associations. Notably, the results of the second part of Experiments 1 and 2 also indicate that participants stick to the convention they created in the first part of the experiment, rather than introducing a different set of mappings.

Given the apparent prevalence of crossmodal common ground in adults, an interesting direction for future research is to investigate its developmental trajectory. Although the current study was not designed to investigate this question, we actually acquired anecdotal evidence[5] from eight pairs of children aged between 7 and 15 who played the communication game in the version of Experiment 2. Their data allow us to draw two very careful conclusions at this point: On the one hand, the data from the first 'one-dimensional' part suggest that children in that age range do equally well as adults in using crossmodal correspondences for the purpose of communicating with each other. On the other hand, the data from the second 'two-dimensional' part indicate that children might differ in interesting ways from adults when facing ambiguity in communication. Specifically, it seemed that the children who participated in our study did not show the same clear preference for relying on interaction history as adults did. In fact, children chose the 'Novelty' mapping, rather than the 'History' mapping, for a good portion of trials, only matching around 60% of the trials based on 'History' – in contrast to the adults' 80% in Experiments 1 and 2. There are, of course, many aspects to consider here, such as children's natural playfulness and curiosity which might have biased them to try something new instead of going for the "safe" option; still, these data point to interesting questions to be explored further.

4.4. Conclusion

In the present work, we propose that people share a specific form of common ground because they process multisensory information in comparable ways; in particular, because they experience the same *crossmodal correspondences* between stimuli features from different sensory modalities. We further propose that people can employ this *crossmodal common ground* for communication and joint action. We thereby extend previous research, which has shown that individuals consistently exhibit crossmodal correspondences, by predicting that individuals also *expect others* to *share* and *understand* these correspondences. Initial support for our proposal comes from a series of short experiments where joint action partners – unable to use conventional forms of communication such as speech or gesture – systematically mapped *visual* stimulus features with, for example, *auditory* stimulus features in line with crossmodal correspondences to create communicative signals. With this proof-of-concept study, we thus contribute to the current issues in joint action research, specifically to the research on how communication systems emerge in the face of interpersonal coordination demands. We conclude that crossmodal correspondences not only occur *within* individuals but that they can also be actively used to facilitate the coordination *between* individuals.

CRediT authorship contribution statement

Laura Schmitz: Conceptualization, Methodology, Investigation, Resources, Writing - original draft, Writing - review & editing. **Günther Knoblich:** Conceptualization, Resources, Funding acquisition, Writing - review & editing. **Ophelia Deroy:** Conceptualization, Resources, Writing - review & editing. **Cordula Vesper:** Conceptualization, Methodology, Investigation, Software, Formal analysis, Data curation, Writing - original draft, Writing - review & editing.

Acknowledgements

This work was funded by the European Research Council under the European Union's Seventh Framework Program (FP7-2007-2013)/ Grant agreement no. 609819 (SOMICS). The authors thank the Institute of Philosophy in London and the Tate Exchange program for supporting their research. The work on this project was facilitated by a Doctoral Research Support Grant awarded to LS by the Central European University.

References

Becchio, C., Sartori, L., Bulgheroni, M., & Castiello, U. (2008). Both your intention and mine are reflected in the kinematics of my reach-to-grasp movement. *Cognition, 106*, 894–912.

Bee, M. A., Perrill, S. A., & Owen, P. C. (2000). Male green frogs lower the pitch of acoustic signals in defense of territories: A possible dishonest signal of size? *Behavioral Ecology, 11*(2), 169–177. https://doi.org/10.1093/beheco/11.2.169

Brennan, S. E. (2005). How conversation is shaped by visual and spoken evidence. In J. C. Trueswell, & M. K. Tanenhaus (Eds.), *Approaches to Studying World-situated Language Use: Bridging the Language-as-product and Language-as-action Traditio* (pp. 95–129). MIT Press.

Brennan, S. E., & Hanna, J. E. (2009). Partner-specific adaptation in dialog. *Topics in Cognitive Science, 1*, 274–291.

Brennan, S. E., Kuhlen, A. K., & Charoy, J. (2018). Discourse and dialogue. In *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (pp. 1–57). American Cancer Society. https://doi.org/10.1002/9781119170174.epcn305.

Candidi, M., Curioni, A., Donnarumma, F., Sacheli, L. M., & Pezzulo, G. (2015). Interactional leader-follower sensorimotor communication strategies during repetitive joint actions. *Journal of the Royal Society Interface, 12*(110), 20150644. https://doi.org/10.1098/rsif.2015.0644

Clark, H. (1996). *Using Language*. Cambridge University Press.

Clark, H., & Brennan, S. E. (1991). *Grounding in communication*. 222–233.

Clark, H. H. (2006). Context and common ground. In K. Brown (Ed.), *Encyclopedia of Language and Linguistics (pp. 105–108)*. Elsevier.

Clark, H. H., & Kruch, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal Memory and Language, 50*, 62–81.

Coward, S. W., & Stevens, C. J. (2004). Extracting meaning from sound: Nomic mappings, everyday listening, and perceiving object size from frequency. *The Psychological Record, 54*(3), 349–364. https://doi.org/10.1007/BF03395478

Curioni, A., Vesper, C., Knoblich, G., & Sebanz, N. (2019). Reciprocal information flow and role distribution support joint action coordination. *Cognition, 187*, 21–31. https://doi.org/10.1016/j.cognition.2019.02.006

Demattè, M. L., Sanabria, D., & Spence, C. (2006). Cross-modal associations between odors and colors. *Chemical Senses, 31*(6), 531–538. https://doi.org/10.1093/chemse/bjj057

Demattè, M. L., Sanabria, D., & Spence, C. (2009). Olfactory discrimination: When vision matters? *Chemical Senses, 34*(2), 103–109. https://doi.org/10.1093/chemse/bjn055

Deroy, O., Crisinel, A.-S., & Spence, C. (2013). Crossmodal correspondences between odors and contingent features: Odors, musical notes, and geometrical shapes. *Psychonomic Bulletin & Review, 20*(5), 878–896. https://doi.org/10.3758/s13423-013-0397-0

Dolscheid, S., Shayan, S., Majid, A., & Casasanto, D. (2013). The thickness of musical pitch: Psychophysical evidence for linguistic relativity. *Psychological Science, 24*(5), 613–621. https://doi.org/10.1177/0956797612457374

Duff, M. C., Hengst, J., Tranel, D., & Cohen, N. J. (2006). Development of shared information in communication despite hippocampal amnesia. *Nature Neuroscience, 9* (1), 140–146.

Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. *Journal of Vision, 7*(5), 1–14. https://doi.org/10.1167/7.5.7

Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision, 10*(1), 1–12. https://doi.org/10.1167/10.1.6

[5] The reason we did not perform a proper analysis on the children's data was that, with a range between 7 and 15 years, our sample was too heterogeneous in terms of age.

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science, 29*, 737–767.

Galantucci, B. (2009). Experimental semiotics: a new approach for studying communication as a form of joint action. In *, 1. Topics in Cognitive Science* (pp. 393–410).

Gallace, A., & Spence, C. (2006). Multisensory synesthetic interactions in the speeded classification of visual size. *Perception & Psychophysics, 68*(7), 1191–1203. https://doi.org/10.3758/BF03193720

Gilbert, A. N., Martin, R., & Kemp, S. E. (1996). Cross-modal correspondence between vision and olfaction: The color of smells. *The American Journal of Psychology, 109*(3), 335. https://doi.org/10.2307/1423010

Grice, H. P. (1975). Logic and conversation. In *Studies in the Way of Words*.

Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences, 24*, 849–937.

Horton, W. S. (2007). The influence of partner-specific memory associations on language production: Evidence from picture naming. *Language and Cognitive Processes, 22*(7), 1114–1139. https://doi.org/10.1080/01690960701402933

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition, 59*(1), 91–117. https://doi.org/10.1016/0010-0277(96)81418-1

Kemp, S. E., & Gilbert, A. N. (1997). Odor intensity and color lightness are correlated sensory dimensions. *The American Journal of Psychology; Urbana, 110*(1), 35–46. https://doi.org/10.2307/1423699

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science, 11*(1), 32–38. https://doi.org/10.1111/1467-9280.00211

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

Klapetek, A., Ngo, M. K., & Spence, C. (2012). Does crossmodal correspondence modulate the facilitatory effect of auditory cues on visual search? *Attention, Perception, & Psychophysics, 74*(6), 1154–1167. https://doi.org/10.3758/s13414-012-0317-9

Knoblich, G., & Jordan, J. S. (2003). Action coordination in groups and individuals: Learning anticipatory control. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(5), 1006–1016. https://doi.org/10.1037/0278-7393.29.5.1006

Knoblich, G., Butterfill, S., & Sebanz, N. (2011). Psychological research on joint action: Theory and data. In Brian Ross (Ed.)*, Vol. 54. The Psychology of Learning and Motivation* (pp. 59–101). Burlington: Academic Press.

Lewis, D. (1969). *Convention: A philosophical study*. Harvard University Press.

Lockridge, C. B., & Brennan, S. E. (2002). Addressees' needs influence speakers' early syntactic choices. *Psychonomic Bulletin & Review, 9*, 550–557. https://doi.org/10.3758/BF03196312

Marks, L. E. (1987). On cross-modal similarity: Auditory–visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance, 13*(3), 384–394. https://doi.org/10.1037/0096-1523.13.3.384

Martino, G., & Marks, L. E. (1999). Perceptual and linguistic interactions in speeded classification: Tests of the semantic coding hypothesis. *Perception, 28*(7), 903–923. https://doi.org/10.1068/p2866

Matlab. (2015). *version 8.5 (R2015a)*. Natick, Massachusetts: The MathWorks Inc.

Melara, R. D. (1989). Dimensional interaction between color and pitch. *Journal of Experimental Psychology: Human Perception and Performance, 15*(1), 69–79. https://doi.org/10.1037/0096-1523.15.1.69

Misyak, J., Noguchi, T., & Chater, N. (2016). Instantaneous conventions: The emergence of flexible communicative signals. *Psychological Science, 27*(12), 1550–1561. https://doi.org/10.1177/0956797616661199

Österbauer, R. A., Matthews, P. M., Jenkinson, M., Beckmann, C. F., Hansen, P. C., & Calvert, G. A. (2005). Color of scents: Chromatic stimuli modulate odor responses in the human brain. *Journal of Neurophysiology, 93*(6), 3434–3441. https://doi.org/10.1152/jn.00555.2004

Parise, C. V., Spence, C., & Deroy, O. (2016). Understanding the correspondences: Introduction to the special issue on Crossmodal correspondences. *Multisensory Research, 29*(1–3), 1–6. https://doi.org/10.1163/22134808-00002517

Pezzulo, G., Donnarumma, F., & Dindo, H. (2013). Human sensorimotor communication: A theory of signaling in online social interactions. *PLoS One, 8*(11), e79876. https://doi.org/10.1371/journal.pone.0079876

Pezzulo, G., Donnarumma, F., Dindo, H., D'Ausilio, A., Konvalinka, I., & Castelfranchi, C. (2019). The body talks: Sensorimotor communication and its brain and kinematic signatures. *Physics of Life Reviews, 28*, 1–21. https://doi.org/10.1016/j.plrev.2018.06.014

Prinz, W. (1997). Perception and action planning. *European Journal of Cognitive Psychology, 9*(2), 129–154.

de Ruiter, J. P., Noordzij, M. L., Newman-Norlund, S., Hagoort, P., & Toni, I. (2007). *On the origin of intentions*. 593–610.

Core Team, R. (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Sacheli, L., Tidoni, E., Pavone, E., Aglioti, S., & Candidi, M. (2013). Kinematics fingerprints of leader and follower role-taking during cooperative joint actions. *Experimental Brain Research, 226*(4), 473–486. https://doi.org/10.1007/s00221-013-3459-7

Schelling, T. C. (1960). *The strategy of conflict*. Harvard University Press.

Schmitz, L., Vesper, C., Sebanz, N., & Knoblich, G. (2017). Co-representation of others' task constraints in joint action. *Journal of Experimental Psychology: Human Perception and Performance, 43*(8), 1480–1493. https://doi.org/10.1037/xhp0000403

Schmitz, L., Vesper, C., Sebanz, N., & Knoblich, G. (2018a). When height carries weight: Communicating hidden object properties for joint action. *Cognitive Science, 42*(6), 2021–2059. https://doi.org/10.1111/cogs.12638

Schmitz, L., Vesper, C., Sebanz, N., & Knoblich, G. (2018b). Co-actors represent the order of each other's actions. *Cognition, 181*, 65–79. https://doi.org/10.1016/j.cognition.2018.08.008

Scott-Phillips, T. C. (2015). *Speaking our minds*. London: Palgrave Macmillan.

Scott-Phillips, T. C., Kirby, S., & Ritchie, G. R. S. (2009). Signalling signalhood and the emergence of communication. *Cognition, 113*(2), 226–233. https://doi.org/10.1016/j.cognition.2009.08.009

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences, 10*(2), 70–75.

Shintel, H., & Keysar, B. (2009). Less is more: A minimalist account of joint action in communication. *Topics in Cognitive Science, 1*, 260–273.

Smith, L. B., & Sera, M. D. (1992). A developmental analysis of the polar structure of dimensions. *Cognitive Psychology, 24*(1), 99–142. https://doi.org/10.1016/0010-0285(92)90004-L

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics, 73*(4), 971–995. https://doi.org/10.3758/s13414-010-0073-7

Spence, C., & Deroy, O. (2013). How automatic are crossmodal correspondences? *Consciousness and Cognition, 22*(1), 245–260. https://doi.org/10.1016/j.concog.2012.12.006

Spence, C., Ngo, M. K., Lee, J.-H., & Tan, H. (2010). Solving the correspondence problem in haptic/multisensory interface design. In M. Hosseini Zadeh (Ed.), *Advances in haptics*. BoD – Books on Demand.

Spence, C., & Parise, C. V. (2012). The cognitive neuroscience of crossmodal correspondences. *I-Perception, 3*(7), 410–412. https://doi.org/10.1068/i0540ic

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and Cognition*. Blackwell Publishers.

Stalnaker, R. (2002). Common ground. *Linguistics and Philosophy, 25*, 701–721.

Surtees, A., Apperly, I., & Samson, D. (2016). I've got your number: Spontaneous perspective-taking in an interactive task. *Cognition, 150*, 43–52. https://doi.org/10.1016/j.cognition.2016.01.014

Tamariz, M., Roberts, S. G., Martínez, J. I., & Santiago, J. (2018). The interactive origin of iconicity. *Cognitive Science, 42*(1), 334–349. https://doi.org/10.1111/cogs.12497

Velasco, C., Woods, A. T., Deroy, O., & Spence, C. (2015). Hedonic mediation of the crossmodal correspondence between taste and shape. *Food Quality and Preference, 41*, 151–158. https://doi.org/10.1016/j.foodqual.2014.11.010

Vesper, C., Butterfill, S. A., Knoblich, G., & Sebanz, N. (2010). A minimal architecture for joint action. *Neural Networks, 23*(8–9), 998–1003.

Vesper, C., & Richardson, M. J. (2014). Strategic communication and behavioral coupling in asymmetric joint action. *Experimental Brain Research, 232*, 2945–2956. https://doi.org/10.1007/s00221-014-3982-1

Vesper, C., Schmitz, L., & Knoblich, G. (2017). Modulating action duration to establish non-conventional communication. *Journal of Experimental Psychology: General, 164*(12), 1722–1737. https://doi.org/10.1037/xge0000379

Vesper, C., & Sevdalis, V. (2020). Informing, coordinating, and performing: A perspective on functions of sensorimotor communication. *Frontiers in Human Neuroscience, 14*. https://doi.org/10.3389/fnhum.2020.00168

Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., & Johnson, S. P. (2010). Preverbal infants' sensitivity to synaesthetic cross-modality correspondences. *Psychological Science, 21*(1), 21–25. https://doi.org/10.1177/0956797609354734

Walker, P., Francis, B. J., & Walker, L. (2010). The brightness-weight illusion. *Experimental Psychology, 57*(6), 462–469. https://doi.org/10.1027/1618-3169/a000057

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences, 7*(11), 483–488. https://doi.org/10.1016/j.tics.2003.09.002

Wilkes-Gibbs, D., & Clark, H. H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language, 31*(2), 183–194. https://doi.org/10.1016/0749-596X(92)90010-U

Wolpert, D. M., Doya, K., & Kawato, M. (2003). A unifying computational framework for motor control and interaction. *Philosophical Transactions of the Royal Society of London B, 358*, 593–602.