Intern. J. Comput.-Support. Collab. Learn (2016) 11:227–253
DOI 10.1007/s11412-016-9235-5

The symmetry of partner modelling

Pierre Dillenbourg[1] · Séverin Lemaignan[1] · Mirweis Sangin[2] · Nicolas Nova[3] · Gaëlle Molinari[4]

Received: 18 June 2015 / Accepted: 25 April 2016 / Published online: 7 May 2016

Abstract Collaborative learning has often been associated with the construction of a shared understanding of the situation at hand. The psycholinguistics mechanisms at work while establishing common grounds are the object of scientific controversy. We postulate that collaborative tasks require some level of mutual modelling, i.e. that each partner needs some model of what the other partners know/want/intend at a given time. We use the term "some model" to stress the fact that this model is not necessarily detailed or complete, but that we acquire some representations of the persons we interact with. The question we address is: Does the quality of the partner model depend upon the modeler's ability to represent his or her partner? Upon the modelee's ability to make his state clear to the modeler? Or rather, upon the quality of their interactions? We address this question by comparing the respective accuracies of the models built by different team members. We report on 5 experiments on collaborative problem solving or collaborative learning that vary in terms of tasks (how important it is to build an accurate model) and settings (how difficult it is to build an accurate model). In 4 studies, the accuracy of the model that A built about B was correlated with the accuracy of the model that B built about A, which seems to imply that the quality of interactions matters more than individual abilities when building mutual models. However,

✉ Séverin Lemaignan
severin.lemaignan@plymouth.ac.uk

1 Computer-Human Interaction in Learning and Instruction, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015, Lausanne, Switzerland

2 Use-able consulting, CH-1004, Lausanne, Switzerland

3 Haute École Arts, Design de Genéve, CH-1201, Genéve, Switzerland

4 Distance Learning University Switzerland (Unidistance), CH-3960, Sierre, Switzerland

these findings do not rule out the fact that individual abilities also contribute to the quality of modelling process.

Keywords Cognitive modelling · Grounding · Theory of mind

Introduction

From its inception, Computer Supported Collaborative Learning (CSCL) research has been following the suggestion by Roschelle and Teasley (1995) that collaborative learning has something to do with the process of constructing and maintaining a *shared understanding* of the task at hand. Building a shared/mutual understanding refers to the upper class of collaborative learning situations, those in which students should build upon each other's understanding to refine their own understanding: what is expected to produce learning is not the mere fact that two students build the same understanding but the cognitive effort they have to engage to build this shared understanding (Schwartz 1995). This effort can be observed by the frequency of rich interactions, i.e. interactions whose occurrence has been related to learning: (self-) explanations in cognitive science (Chi et al. 1989; Webb 1991), conflict resolution in socio-cognitive theories (Doise et al. 1975) and mutual regulation (Blaye and Light 1995) in a Vygostkian perspective. The construction of shared understanding has been investigated for several decades in psycholinguistics, under the notion of *grounding* (Clark and Wilkes-Gibbs 1986). However, the relevance of grounding mechanisms for explaining learning outcomes has been questioned in the learning sciences. Grounding mechanisms are appropriate to explain conversational events, such as referential failures in short dialogue episodes, but they hardly predict deeper phenomena such as *conceptual change* (i.e. the acquisition, acceptance and integration of a new belief into one's mental model) over longer sessions (Dillenbourg and Traum 2006). The cumulative effect of grounding episodes can probably be better understood from a socio-cultural perspective:

> Collaborative learning is associated with the increased cognitive-interactional effort involved in the transition from *learning to understand each other* to *learning to understand the meanings of the semiotic tools that constitute the mediators of interpersonal interaction* (Baker et al. 1999, p.31)

Along this line, several scholars suggest that CSCL research should go deeper towards understanding how partners engage in shared meaning making (Stahl 2007) or *intersubjective* meaning making (Suthers 2006).

Paradoxically, while Clark's theory is somewhat too linguistic from a learning viewpoint, it is criticized at the same time as being too cognitivist by some psycholinguists, i.e. as overestimating the amount of shared knowledge and mutual representations actually necessary to conduct a dialogue. The fundamental issue, as old as philosophy, is the degree of coupling between the different levels of dialogue, mostly between the lexical/syntactical level and the deeper semantic levels. Pickering and Garrod (2006) argue that mutual understanding starts mostly with a *superficial alignment* at the level of the linguistic representations, due to priming mechanisms, and that this local alignment may – in some cases – lead to a *global alignment* of the semantic level (*deep grounding*). For these authors, the convergence

in dialogue, and even the repair of some misunderstandings, is explained by this mimetic behavior more than by a monitoring of each other's knowledge:

> [...] interlocutors do not need to monitor and develop full common ground as a regular, constant part of routine conversation, as it would be unnecessary and far too costly. Establishment of full common ground is, we argue, a specialized and non-automatic process that is used primarily in times of difficulty (when radical misalignment becomes apparent). (Pickering and Garrod 2006, p.179)

This view is actually not incompatible with Clark's *grounding criterion* (Clark and Schaefer 1989): the degree of shared understanding that peers need to reach depends upon the task they perform. For instance, a dialogue between two surgeons might rely on superficial alignment if they talk about their friends but has to guarantee accurate common grounds when talking about which intervention will be conducted in which way on which patient. In this paper, we operationalized the grounding criterion, i.e. the necessity for accurate modelling, as *the correlation between the accuracy of partner models and measures of team performance*.

This interesting cognitive science debate occurred mostly outside the field of learning. In education, the question is to relate these mechanisms to learning outcomes: Is linguistic alignment sufficient to trigger conceptual change? Does negotiation of meaning only occur when partners monitor and diagnose each other's knowledge? If the ratio between shallow alignment and deep grounding depends upon the task, and if deep grounding is a condition for learning, then the pedagogical challenge is indeed to design tasks that require deep grounding. Most empirical studies on grounding and alignment are conducted with simple referencing tasks such as asking the way to the train station or helping the peer to choose a picture among many. In the studies we report here, we explore several richer tasks such as arguing about a sensitive issue or building a concept map.

Deep grounding or shared meaning making requires some cognitive load. For Clark and Wilkes-Gibbs (1986), what is important is not the individual effort made by the receiver of a communicative act, but the overall *least collaborative effort*. The cost of producing a perfect utterance may be higher than the cost of repairing the problems that may arise through misunderstandings, and in fact, subjects tend to make less efforts adapting their utterances to a specific partner when they know that they can later provide feedback on his/her understanding (Schober 1993). We introduced the notion of *optimal collaborative effort* (Dillenbourg et al. 1995) to stress that misunderstanding should not be viewed as something to be avoided (if this was possible), but as an opportunity to engage into verbalization, explanation, negotiation, and so forth. This issue is related to the global argument regarding cognitive load in learning activities, especially in discovery learning environments: there is no learning without some cognitive load, but overload may hinder learning (Paas et al. 2003). In the context of collaborative learning, we understand the cognitive load induced by mutual modelling as part of Schwartz (1995) notion of effort towards a shared understanding. For instance, CSCL researchers expanded the use of *collaboration scripts* (Kobbe et al. 2007). A script is a pedagogical method that frames collaborative learning activities in order to foster the emergence of productive interactions such as argumentation, explanation or conflict. Conflict-resolution scripts such as the ARGUEGRAPH (Dillenbourg and Hong 2008) form pairs of students with opposite opinions, which increases the difficulty of consensus building, requiring more justifications, more negotiation, and more load. Similarly, JIGSAW

scripts (Aronson et al. 1978) provide peers with different but complementary knowledge for augmenting (reasonably) the efforts that group members have to engage into to reach a shared solution.

In summary, the controversy around the cognitive depth of shared understanding pertains to psycholinguistics, which investigates natural conversations. The situation is different in collaborative learning; tasks are actually designed for requiring a deeper negotiation of meaning. The expertise of CSCL is to design collaborative situations that create interdependence, avoid group think, and allow students to detect any "illusion of shared understanding" (Cherubini et al. 2005). Our question is hence not anymore "do peers build a shared understanding ?", but rather "whenever peers have to build a shared understanding, how do they achieve it ?". This question addresses the mechanisms of grounding, in which the basic sequence is: make a proposition, detect misunderstanding and, if any, repair them. This paper focuses on the middle part, the detection of misunderstandings. Detecting a misunderstanding means that the emitter of a message identifies a mismatch between his communicative intention and the way his message is understood by his or her partner. Detecting one peer's misunderstanding is investigated hereafter under the general umbrella of partner modelling.

Partner modelling

We refer to *Partner modelling* as the process of inferring one's partner's mental states. Any claim that students carry out a detailed monitoring of their peers would be as incorrect as any claim that they do not maintain any representation at all. If mental modelling had to be permanently detailed and accurate, subjects would obviously face a huge cognitive load. Conversely, peers could not collaborate without some minimal amount of mutual modelling. Collaborative learning dialogue include many instances of utterances such as "I thought he would do that" (first order modelling) or even "He thought I would do that but I intended something else." (second order modelling).

The content of the partner model ranges from *dispositional* to *situational* aspects. The *dispositional* aspects refer to A's representation of B's long term knowledge, skills or traits. It is thus closely related to the notion of transactive memory (Wegner 1987; Moreland 1999). *Situational* aspects refer to A's representation of B's knowledge, behavior or intentions specifically activated in the situation in which A and B are collaborating, some of them being valid for 2 seconds, other ones for 2 hours. Examples of fragments that constitute A's model of B regarding to aspects X, i.e. $Model(A, B, X)$, abbreviated $\mathcal{M}(A, B, X)$, could be:

- $Model(A, B, knowledge)$: what does A know about B's knowledge with respect to the task at hand or, inversely, about B's knowledge gaps? When can A consider B's statements as reliable?
- $Model(A, B, skills)$: what does A know about B's skills with respect to the task at hand? May A expect B to perform well in a specific subtask? The effectiveness of division of labor depends on the quality of this mutual model.
- $Model(A, B, goals)$: what does A know about B's intentions with respect to the project, including B' motivation and commitment? Can A trust B when B promises to deliver?
- $Model(A, B, task)$: what does A know about B's representation of the situation and the task: does A knows whether B has the same understanding of the problem at stake?

- $Model(A, B, plans)$: what does A know about B's strategy. Does A understand why B did what he did? Is A able to anticipate what B will do next?
- $Model(A, B, \text{“}urgent\text{”})$: what does A know about B's understanding of A's last utterance: does "urgent" mean now, soon or "not too late"?

The list of what X stands for in $\mathcal{M}(A, B, X)$ is possibly infinite: beliefs, emotions, history, status, etc. A partner model is likely not a "box", i.e. not a monolithic representation but rather a mosaic of information fragments about the partner, with various granularity and various life cycles. This mosaic is elaborated through a variety of mechanisms, first for building an initial model of the partner, then for updating this model. As two students meet for the first time, partners models are initialized by the assumptions they make upon each other based on cues such as his/her belonging to broad categories (age, culture, profession, ...), stereotypes (sportsmen, junkie, business women, Swiss,...) as well as physical appearance. Scholars studied how initial modelling impacts communication. In their experiment, Slugoski et al. (1993) pretended to their subjects that their (confederate) partner had or had not received the same information. They observed that the subjects adapted their dialogue by focusing the explanation on the items that he/she was supposed to ignore. Brennan (1991) showed that the subjects used different initial strategies in forming queries depending on who they were told their partner was.

Initial common grounds are also initiated by co-presence: they include events to which A and B attended together (Clark and Marshall 2002) in the physical space or in their cultural space (e.g. "09-11"). While co-presence means that they can refer to shared objects and events, it does not imply that they give them the same meaning. Namely, a shared screen does not mean a shared understanding (Dillenbourg and Traum 2006).

After initialization, partners models are updated during the collaborative work through verbal and non-verbal interactions. A default inference rule is that "my partner agrees with me unless he disagrees", which rejects the critiques that partner modelling generates an unbearable cognitive load. This default rule is superseded by the several mechanisms for monitoring and repairing the partner understanding: acknowledgement, continuous attention, relevance of next turns, facial expressions including gaze signals, etc.

Finally, partner modelling does not occur in a vacuum but it is highly contextualized. Clark and Brennan (1991) review how the features of the collaborative situation, namely the media (co-temporality,...), may facilitate or hamper mutual modelling. Hutchins and Palen (1997) reported a study in which a short silence between two pilots was perfectly interpreted because it occurred in a highly constrained communication context. Some environments are more productive than others in helping peers to detect their misunderstandings. Roschelle and Teasley (1995) reformulate the design of CSCL interfaces to provide ways for peers to detect and repair their misunderstanding.

Mutual modelling

Mutual modelling is bi-directional. During dyadic problem solving, partner A builds some model of B and B build some model of A. Moreover, these two processes are not independent: A's model of what B knows, includes what B knows about A. This leads to nested levels of modelling. If A states "B thinks I am good in maths", A builds a *second level* model: $\mathcal{M}(A, B, \mathcal{M}(B, A, \text{maths-skill}))$. This leads to possibly infinite regress of nested models: A saying "B knows that I don't expect him to solve this statistics problem" corresponds to $\mathcal{M}(A, B, \mathcal{M}(B, A, AB\text{statistic-skills}))$. As we will see in one study we report

on, mutuality also applies to triads in which A will elaborate a model of B and of C, and reciprocally. This is probably also true for larger groups although one may hypothesize that there exists (yet undefined) group size from which it is not possible to model all partners individually and, therefore, members model the group as a whole instead as a collection of individuals. We do not explore this hypothesis here and limit ourselves to dyads and triads.

This mutuality allows us to address a fundamental question: *does the quality of partner modelling depend upon the cognitive skills of each partner (some people being better in perceiving other's states) or does it result from the quality of interactions among them ?* Behind his question, the reader may perceive the long lasting debate between tenets of, respectively, the individual and social views of human cognition. The simple hypothesis is that when individuals are randomly paired for an experiment, there is no reason for which their individual cognitive skills would correlate. Therefore, if it occurs that the quality of the $Model(A, B)$ is correlated with $Model(B, A)$, one may infer that this quality depends upon what A and B have built together while interacting.

To answer this question, we went back to five previous studies that addressed various other research questions, but in which the correlation between the quality of these two models could be computed. The studies we report do not hence constitute a clean sequence of experiments to investigate mutual modelling but the ad-hoc revisiting of previous experiments to explore a question that we had then neglected. Some of these studies are about collaborative learning while others are only collaborative problem solving, but the latter rely on rich tasks that are similar to those we use in CSCL.

A notation for discussing mutual modelling

Natural language becomes cumbersome when describing things such as "the model that A builds about the model that B builds about A". Therefore, to define hypotheses and report on experiments on mutual modelling, we use the notation $\mathcal{M}(A, B, X)$ to denote "A knows that B knows X". This notation is not proposed as a formal theory of mutual modelling but as useful simplification for communicating about mutual modelling. This notation does not mean A has an explicit, monolithic representation of B: it must be understood as an abstraction referring to complex socio-cognitive processes. As explained in the previous section, the model built by A can be fragmented, multi-dimensional, etc. This notation is neither presented as a computational model of mutual modelling, nor as some universal formalism; its usefulness is internal to this paper. Additionally, we refer to the *degree of accuracy* of the model as $\mathcal{M}^{\circ}(A, B, X)$. We discuss in the next section the methodological difficulty in measuring this accuracy.

We parametrize and assess the mutual modelling effort through 3 variables:

1. Tasks vary a lot with respect to how much they require mutual understanding. The *grounding criterion* – denoted $\mathcal{M}^{\circ}_{min}$ – represents the minimum level of modelling accuracy required for a task T to succeed. Qualitatively, if the performance on a given task T is correlated to $\mathcal{M}^{\circ}(A, B, X)$, then $\mathcal{M}(A, B, X)$ is significant to T success, and the grounding criterion of X for T ($\mathcal{M}^{\circ}_{min}(A, B, X, T)$) is non-zero. Under the assumption that the higher the correlation, the more critical $\mathcal{M}(A, B, X)$ is to T, we hereafter use the correlation coefficient as an estimate of $\mathcal{M}^{\circ}_{min}$.
2. Before any specific grounding action, there is generally a non-null probability that X is mutually understood by A and B (e.g. X is part of A's and B's cultures, it is manifest to co-present subjects or simply there is not much space for misunderstanding or

disagreement about X). We simply could not collaborate without a certain level of initial grounds. We denote the theoretical accuracy of *initial grounds* with $\mathcal{M}^{\circ}_{t_0}(A, B, X)$.

3. The *cost of grounding* X refers to the physical and cognitive effort required to perform a grounding act α: a verbal repair (e.g. rephrasing), a deictic gesture, a physical move to adopt one partner's viewpoint, etc. This cost varies according to media features (Clark and Brennan 1991).

Based on these 3 parameters, the probability of making an action α_t about content X at time t during task T in order to increase $\mathcal{M}^{\circ}(A, B, X)$ is the ratio between how much it is needed and how much it costs (Traum and Dillenbourg 1996):

$$p(\alpha_t(X,T)/_{\mathcal{M}^{\circ}_{t+1}(A,B,X)\nearrow}) \simeq \frac{\mathcal{M}^{\circ}_{min}(A,B,X,T) - \mathcal{M}^{\circ}_{t}(A,B,X)}{cost(\alpha_t)} \quad (1)$$

This formula is presented as a qualitative summary, not as a real equation, since several parameters are hard to quantify (e.g. the cost of a communication act depends upon the user as well).

We can further clarify the parameters in the context of the experiments we present hereafter:

- $\mathcal{M}^{\circ}(A, B, X)$: our experiments address different contents that can be represented in mutual models:
  1. $\mathcal{M}^{\circ}(A, B, actions)$ is about how well A guesses what action B has performed (study 2) or will perform next (study 1),
  2. $\mathcal{M}^{\circ}(A, B, emotion)$: how accurately A perceives B's emotional state (study 3),
  3. $\mathcal{M}^{\circ}(A, B, knowledge)$: how accurately A estimates B's knowledge with respect to the material they learn together (study 4 and 5).
- $\mathcal{M}^{\circ}_{min}(A, B, X, T)$: our studies build upon various collaborative tasks: argumentation (study 3), games (study 1 and 2) and concept mapping (study 4 and 5). By varying the tasks, we do actually vary the grounding criterion. The tasks were all designed to require a reasonably high grounding criterion, as they are meant for the participants to have to actually build a solution or a representation together.
- $\mathcal{M}^{\circ}_{t0}(A, B, X)$: along the same reasoning, the initial degree of common grounds should be rather low (and hence the difference between initial and required degrees rather high) in order to make mutual modelling effort more observable. Studies 1, 4, and 5 have been conducted with teams of students who did not know each other. They came nonetheless from the same university (and they hence had some general common grounds). For studies 2 and 3, students knew each other before for reasons explained later on. In study 5, we manipulated the initial mutual modelling by using a JIGSAW script.
- $cost(\alpha)$: in all studies but study 4, the cost of grounding is an independent variable. Study 3 uses media richness as independent variable, with the hypothesis that modelling emotions is "cheaper" with a richer medium, i.e. when peers can see each other. Studies 1, 2, and 4 use awareness tools which, in principle, reduce the cost of mutual modelling, but do not eliminate all costs: if the tool provides A with information about what B does/knows, this additional information may actually increase cognitive load. Awareness tools constitute a kind of mutual modelling prosthesis, and, like any prosthesis, they may augment mutual modelling (by facilitating it or even scaffolding it) or inhibit it (by making it useless).

While we introduce here formally the cost of grounding $cost(\alpha)$ as one relevant variable for the discussion of mutual modelling situations, we will not attempt to characterize it beyond these qualitative observations in the studies we present hereafter.

Methodological issues

Since a mental model is not directly observable, the study of mutual modelling is methodologically difficult. How can we for instance measure $\mathcal{M}^{\circ}(A, B, X)$? We proceed in two steps, first to capture $\mathcal{M}(A, B, X)$ and then to estimate $\mathcal{M}^{\circ}(A, B, X)$.

Capturing $\mathcal{M}(A, B, X)$ The simplest method is to ask A what he/she believes about what B knows, feels, intends to do, etc. This raises obvious methodological concerns since such a question triggers a modelling process beyond what would naturally occur. To avoid this bias, one can estimate mutual modelling after task completion. Then, the obvious drawback are memory losses and post-hoc reconstruction. The first method was used in study 2 and the second one in the other studies. Another option would be to rely on external behavioural metrics like eye-tracking: we hypothesise for instance that the fixation time reflects the efforts engaged by the human to understand, hence, model, the others. Such an approach has however not been investigated in the presented studies.

Estimating $\mathcal{M}^{\circ}(A, B, X)$ Once $\mathcal{M}(A, B, X)$ is captured, we need to access the reference model $\mathcal{M}(B, X)$ to estimate its accuracy. Since we can only indirectly access it via what B reports (i.e. $\mathcal{M}(B, B, X)$), accuracy can be estimated in 2 ways:

- Subjective accuracy: In study 3, for instance, we compute $\mathcal{M}^{\circ}(A, B, X)$ by measuring if A describes B's emotions in the same way B reports its emotions ($\mathcal{M}(A, B, X) = \mathcal{M}(B, B, X)$).
- Objective accuracy: In studies 4 and 5, we compute $\mathcal{M}^{\circ}(A, B, X)$ by comparing $\mathcal{M}^{\circ}(A, B, K)$ to B's actual knowledge as it has measured by a test.

Our method for investigating mutual modelling relies on the observation of the variations of accuracy that result from variations of external parameters (the variables of the formula 1 above): for instance, the accuracy should go down if we increase the cost of grounding acts, and conversly go up if we increase the grounding criterion (i.e. the necessity to build a shared understanding). One way to produce these variations relies what CSCW researchers call "awareness tools", i.e. functionalities that inform A about B's actions that A can not directly perceive due to B working in a different subset of the virtual space. Different awareness tools are used in the following studies as methodological levers to experimentally manipulate the mutual modelling activity.

Hypotheses and questions

The experiments we report here address mutual modelling across different tasks, some with dyads, others with triads. They were conducted over 6 years in two different institutions by different researchers. They used different independent, intermediate, and dependent variables. Nonetheless, we were able to retroactively address 3 research questions across these studies.

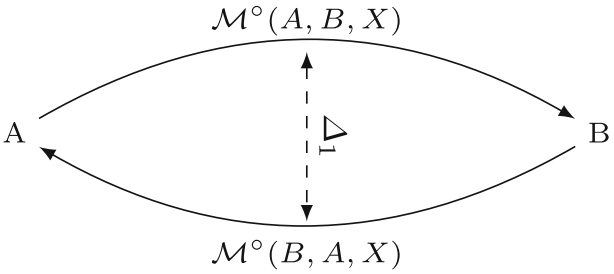


Fig. 1 Mutual modelling in a dyadic interaction, $\Delta_1 = \Delta(\mathcal{M}^\circ(A, B, X), \mathcal{M}^\circ(B, A, X))$

The symmetry question As stated earlier, the fundamental challenge is to determine if mutual modelling is an individual skill (hypothesis $\mathcal{H}_1$ below) or the emergent property of social interactions (hypothesis $\mathcal{H}_3$). This question is empirically translated into the symmetry of mutual modelling (Fig. 1): what is the relationship between $\mathcal{M}^\circ(A, B, X)$ and $\mathcal{M}^\circ(B, A, X)$? A low symmetry would mean that mutual modelling is mainly an individual attitude/aptitude ($\mathcal{H}_1$). A high correlation might support $\mathcal{H}_3$ since there is a low probability that randomly formed pairs integrate peers with the same level of mutual modelling skills. However, a high correlation could also have another explanation, stated in $\mathcal{H}_2$: it could be that A is good at modelling B and good at helping B to model herself or himself.

- $\mathcal{H}_1$: $\mathcal{M}^\circ(A, B)$ depends upon A's ability or effort to model B,
- $\mathcal{H}_2$: $\mathcal{M}^\circ(A, B)$ depends upon B's ability or effort to help A to model him/herself,
- $\mathcal{H}_3$: $\mathcal{M}^\circ(A, B)$ depends upon the quality of interactions among A and B.

$\mathcal{H}_2$ relates to second level modelling since B needs to monitor A to see if A understood him/her ($\mathcal{M}(B, A, \mathcal{M}(A, B))$). We will see that $\mathcal{H}_2$ and $\mathcal{H}_3$ are actually difficult to differentiate.

The triangle questions With triads, we may compute the accuracy of 6 models: $\mathcal{M}^\circ(A, B, X)$, $\mathcal{M}^\circ(B, A, X)$, $\mathcal{M}^\circ(A, C, X)$, $\mathcal{M}^\circ(C, A, X)$, $\mathcal{M}^\circ(C, B, X)$ and $\mathcal{M}^\circ(B, C, X)$. This leads to two *triangle questions* (Fig. 2): Do A and B have the same accuracy when modelling C ($\Delta_2 = \Delta(\mathcal{M}^\circ(A, C, X), \mathcal{M}^\circ(B, C, X))$)? A significant correlation would support , $\mathcal{H}_2$ (C has helped both A and B to model C) or support $\mathcal{H}_3$ (the quality of triad interactions enables all partners to model each other accurately)

Conversely, does model with similar accuracy A and B? (low $\Delta_3 = \Delta(\mathcal{M}^\circ(C, A, X), \mathcal{M}^\circ(C, B, X))$)? A positive answer would support $\mathcal{H}_1$, being simply good at modelling any partner. It could also support $\mathcal{H}_3$, since the quality of interactions

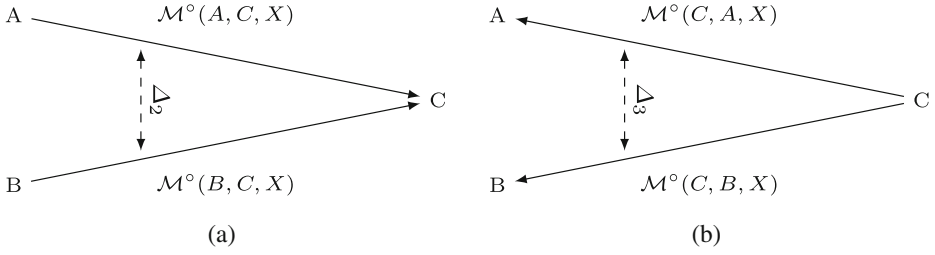


Fig. 2 Mutual modelling in a triadic interaction

influences the accuracy of two models that ţs building. A negative answer would support $\mathcal{H}_2$ since Δ_3 could mainly be explained by the fact that A helped more ţo model him than B did.

In addition, the comparison between Δ_2 and Δ_3 could tell us whether the accuracy of mutual modelling depends more upon the modeller's effort ($\mathcal{H}_1$) or the modellee's behaviour ($\mathcal{H}_2$).

Note that we consider the quality of interactions at triad level (A, B, C), neglecting the cases where A and B interact better for instance than B and C, since there was no "private" communication channel in the following studies. We do nonetheless acknowledge that this point could be debated.

The rectangle questions We can go further by comparing self- versus other modelling (Δ_4 in Fig. 3). A large difference would indicate that meta-cognitive skills (self-modelling) and partner modelling skills are rather different skills, while a small Δ_4 could be interpreted as the indications that these are two specific instance of a more general cognitive process. This questions is however not central to this paper, and the value of Δ_4 is only available in one of the reported studies.

We can also question if modelling skills depend upon what aspects are being modeled (X or Y), which would explain vertical differences (Δ_5 in Fig. 3). These differences would allow refining the notion of modelling skills, namely whether there exist some general ability to model partners or whether this is only the abstraction of a beam of more specific skills such as detecting emotions versus identifying references from deictic gestures.

Studies

We report on five studies (Table 1) conducted by different researchers in different contexts between 2000 and 2015. They do not form a consistent research strategy but the fact that some trends emerge despite their diversity constitutes the richness of this line of work.

Study 1: Effect of an awareness tool on $\mathcal{M}^\circ(A, B)$ in a virtual game

We studied the impact of an awareness tool on group performance and mutual modelling (Nova et al. 2007). The availability of an awareness tool was our independent variable. In previous studies, we replayed a video of the game to subjects who surprised us by their ability to remember former states of their mutual model: "I did that because I thought that you would do that". Hence, this experiment focused the representation of each other's action plans. During the game, we asked them to anticipate the next action of their partner as well as to announce their own actions.

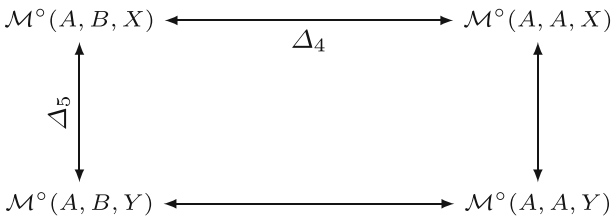


Fig. 3 Meta-cognitive skills (*horizontally*) and domain-dependent modelling (*vertically*)

Table 1 Overview of the studies and main characteristics

	Study 1	Study 2	Study 3	Study 4	Study 5
Task	game in virtual space	game in physical space	building an argument map	building a concept map	building a concept map
Interactions	audio	written	audio / video	audio	audio
Shared editor	3D space	2D map	concept map	concept map	concept map
Group size	dyads	triads	triads	dyads	dyads
Duration (mean)	90min	16min	61min	90min	90min
Awareness tool	partner's position	partner's current/past pos.	-	partner's concept map	partner's scores at pre-test
Dependent variable	team/indiv. performance	team/indiv. performance	team performance	team/indiv. knowledge	team/indiv. knowledge
Independent variable	awareness tool *vs* none	awareness tool *vs* none	audio *vs* audio+video	script *vs* none	awareness tool *vs* none

Experimental setting

SPACEMINERS is a 3D game that involves two players harvesting ore found on asteroids (Fig. 4). To do so, they must launch drones through the space after choosing their initial direction and speed. Once launched, the trajectory of drones is influenced by the gravity of planets and by "trajectory modification" tools. During the experiment, the teams were confronted with three increasingly complex situations. The experiment was 2 hours long: a 30 minute tutorial and 3 levels of 30 minutes. The players were using a regular joystick and communicated with each other through an audio channel.

The independent variable was the availability of an awareness tool that shows to player A the location and gaze direction of player B: in this "awareness" condition, players could switch to the *scout mode* where they could view what their partner was looking at. We hypothesize that this would enable subjects to more accurately infer their teammate's intentions. Each player sat in front of a distinct computer located in different rooms.

Subjects

Thirty-six persons participated in this study, all native French speakers. We formed 18 dyads who did not know each other beforehand. The pairs were randomly assigned to either the control condition (without the awareness tool) or the awareness condition (with the awareness tool).

Variables

Task performance was measured by the score reached by the two subjects at the end of the game (three levels). The effort of mutual modelling was measured as the ratio of time that players would spend in the scout mode (divided by total time), which is the time during which players are not performing their own actions but monitoring their partner's actions.

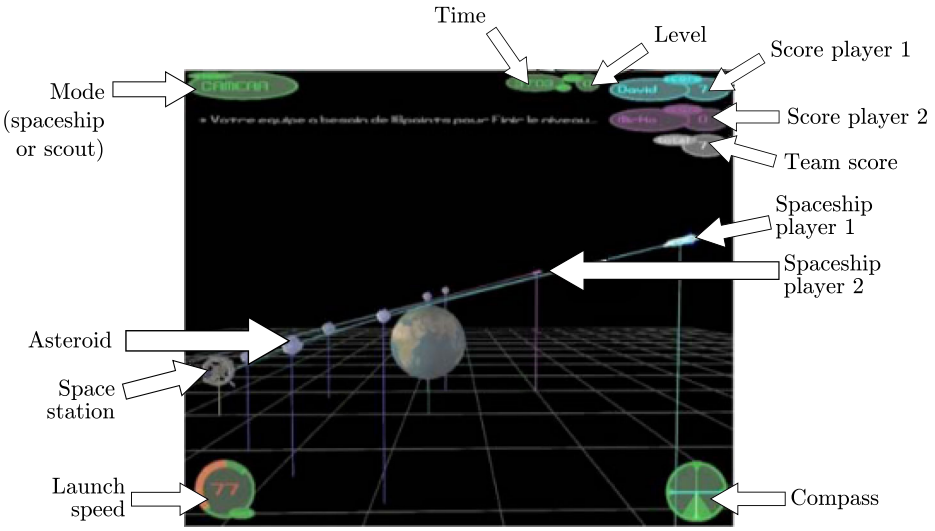


Fig. 4 Screenshot of the SPACEMINERS game

In order to evaluate $\mathcal{M}^{\circ}(A, B)$ during the task, we used two questionnaires (Fig. 5) that were displayed during each of the three games, as a transparent layer appearing over the game display. The first questionnaire concerned the player's intended actions. The second questionnaire asked each player about what he thought her or his partner was intending to do. Some answers were identical in both questionnaires (like "adjusting a shot") while others were reversed ("to guide my partner" versus "to guide me"). This method provides us with a subjective measure of accuracy ($\Delta(\mathcal{M}(A, A, X), \mathcal{M}(A, B, X))$) rather than an objective measure (i.e. the model $\mathcal{M}(A, B)$ is compared to B's next action) because some of the activities proposed by the questionnaire were not observable by the environment (e.g. establishing a strategy). We calculated $\mathcal{M}^{\circ}(A, B, activity))$ as the number of common answers between questionnaires $\mathcal{M}(A, A)$ and $\mathcal{M}(A, B)$ in each game and computed the average value across the 3 levels.

Results

Grounding criterion The grounding criterion was high: the correlation between $\mathcal{M}^{\circ}(A, B)$ and task performance was 0.42, $p = 0.05$. Pairs with an accurate mutual model reached higher scores. A regression analysis confirmed the positive and significant relation between group performance and mutual modelling accuracy ($\beta = 54$, $p = 0.02$).

Study-specific questions The awareness tool permitted higher group performance, but it did not improve the accuracy of the mutual model. Since teams were free to use the awareness tool or not (the *scout* mode), we performed a post-hoc split of players depending on how much time they used it. The split point was the mean of time spent in the scout mode and it led to the constitution of two groups made up of 12 individuals "short time in scout mode" and 24 individuals "long time in scout mode". A two-way analysis of variance conducted on these contrasted groups revealed that pairs in the awareness condition who spent more time in the scout mode reached higher levels of $\mathcal{M}^{\circ}(A, B)(F = 8.02, p = 0.015)$. Of course, a post-hoc split does not support a causal direction. An alternative explanation could be that good modellers are more social and hence appreciate the awareness tool.

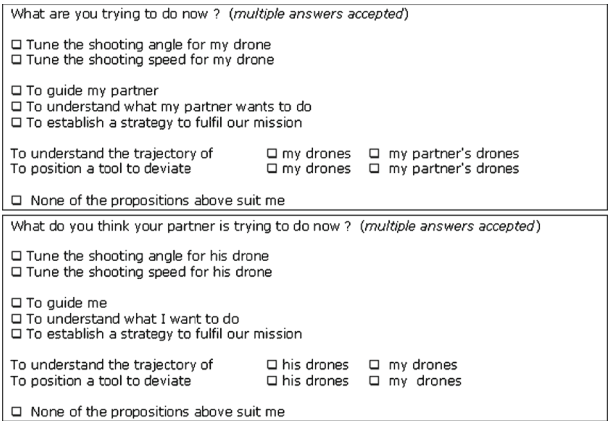


Fig. 5 $\mathcal{M}(A, A)$ and $\mathcal{M}(A, B)$ questionnaires in SpaceMiners (translated from French)

Symmetry question We computed intra-class correlation as described by Kenny et al. (1998) from the answers to the cross-questionnaires. Considering all pairs in both conditions, we found a positive and significant correlation ($r = 0.38$, $p < 0.05$) between $\mathcal{M}^{\circ}(A, B)$ and $\mathcal{M}^{\circ}(B, A)$. Interestingly, this was higher in the control group ($r = 0.44$) than in the experimental group ($r = 0.24$). Actually, $\Delta(\mathcal{M}^{\circ}(A, B), \mathcal{M}^{\circ}(B, A))$, i.e. the absolute differences between the models accuracy, was not significantly different with or without the awareness tool ($F[1, 13] = 0.144$, $p > 0.5$). This result could be explained by the fact that the players without awareness tools communicated more.

The triangle and rectangle questions are not addressed in this study

Discussion

How do we interpret a correlation of 0.38 between $\mathcal{M}^{\circ}(A, B)$ and $\mathcal{M}^{\circ}(B, A)$? It is significant, which supports $\mathcal{H}_2$ and $\mathcal{H}_3$. It is nonetheless far from 1, which implies we cannot discard the individual modelling skills. We collect more evidence in the next studies.

Study 2: Effect of an awareness tool on $\mathcal{M}^{\circ}(A, B)$ in a pervasive game

This study concerns a collaborative game that occurred in physical space (Nova et al. 2006). We studied whether players build an accurate model of the path followed by their partners, assuming that this path would reflect their problem solving strategy. We used an objective measure of $\mathcal{M}^{\circ}(A, B)$: the distance between where A believes B has been walking and where B actually went. The main hypothesis concerned the effect of awareness tools on group performance and on $\mathcal{M}^{\circ}(A, B)$.

Experimental setting

CATCHBOB is a mobile game in which groups of 3 players have to solve a joint task. The game was played on a university campus. Participants had to find a virtual object (*Bob*) and to "catch" it by forming a triangle around it. The players used a Tablet PC that displayed a map of the campus and an indication of their personal distance to Bob. Their annotations on the map were shared with the two other players (A could see what B and wrote). These annotations faded out after a few minutes to avoid covering the full display. The awareness tool also displayed the location of the two other players on the map. Three conditions were considered: the control condition (without tool) and two experimental conditions: synchronous awareness (display of the current position of each player) and asynchronous awareness (display of current position of each player as well as their recent spatial trace).

Subjects

Ninety students participated in this experiment. We only selected students from university campus since knowledge of the campus geography had an impact both on group performance and on mutual modelling: to represent the path of someone across some space is difficult without an *a priori* mental map of this space. We formed groups of students who knew each other. We assigned 10 triads to each of our three experimental conditions. Each condition was made up of approximately 25 % of women, but we did not control gender repartition within each triad.

Variables

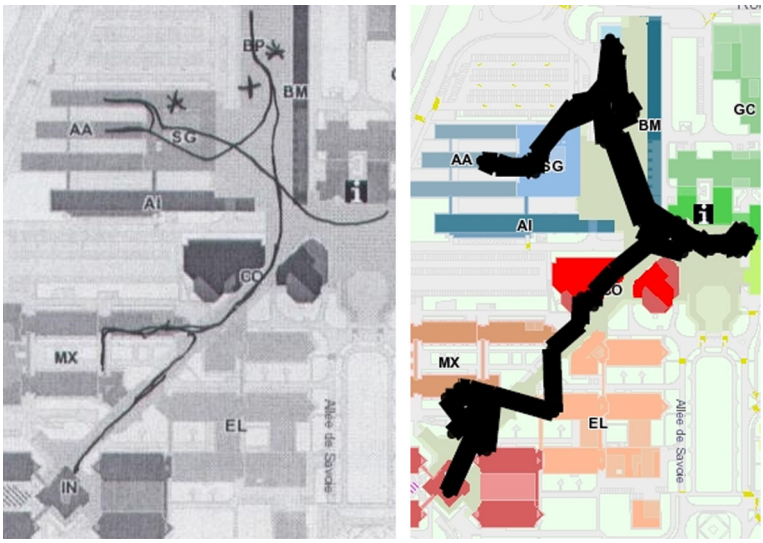
The independent variable was the presence and role of the awareness tool. As a dependent variable, we had the task performance which was the distance covered by the team to catch Bob and $\mathcal{M}^{\circ}(A, B)$. To estimate $\mathcal{M}^{\circ}(A, B)$, we asked players to draw on paper their own path and the path of each of their partners after the game. This enabled us to calculate the number of errors players made while drawing the path of their partners. We compared the path that player A attributed to B with B's real path recorded by the system and the same for A & ąnd B & ąs depicted on Fig. 6.

$\mathcal{M}^{\circ}(A, B)$ is the sum of errors made by A about B's paths. An error was either drawing a place where the partner had not been or not drawing a place where he/she had gone. One could argue that $\mathcal{M}^{\circ}(A, B)$ is biased by the subjects' ability to translate the memory of their trajectories into a map drawing. However, 85% of subjects made no mistake at all when drawing their own path. We therefore consider mistakes in their partners' path as being due to a lack of mutual modelling accuracy instead of being due to spatial reasoning skills.

Results

Grounding criterion The correlation between $\mathcal{M}^{\circ}(A, B)$ and the task performance (path lengths) was low: 0.15. Using a post-hoc split on $\mathcal{M}^{\circ}(A, B)$, we found no significant difference between the performance of the groups with high and low $\mathcal{M}^{\circ}(A, B)$ ($F = 1.45$, $p = 0.24$). Conversely, a post-hoc split of the groups according to their performance did not show any significant differences on $\mathcal{M}^{\circ}(A, B)$ ($F = 1.16$, $p = 0.29$).

Study-specific questions There was no significant difference regarding the task performance. However, and surprisingly, the absence of the awareness tool was related to a higher $\mathcal{M}^{\circ}(A, B)$: players tended to better remember their partners' paths when they could not



(a) A's drawing of B's path. (b) Actual path followed by B.

Fig. 6 Reported and actual path of one of the player, during the CATCHBOB game

constantly monitor their positions. As detailed in the original study (Nova et al. 2005), it appeared that teams without awareness tool made more manual annotations on the map while permanent monitoring has an underwhelming effect.

Symmetry question The correlation between $\mathcal{M}^{\circ}(A, B)$ and $\mathcal{M}^{\circ}(B, A)$ is positive ($r = 0.41$) and significant ($p < 0.01$): the more A makes errors about B, the more B does as well (and vice-versa).

Triangle questions Regarding Δ_2, the correlation between $\mathcal{M}^{\circ}(A, C)$ and $\mathcal{M}^{\circ}(B, C)$ is significant: $r = 0.43$, $p < .001$. Concerning Δ_3, the correlation between $\mathcal{M}^{\circ}(A, B)$ and $\mathcal{M}^{\circ}(A, C)$ is significant as well: $r = 0.30$, $p < .01$.

Discussion

The positive correlation observed in the symmetry question confirms the first study. In this case, this was not expected given the high heterogeneity of spatial skills among adults (Liben et al. 1981). This result therefore supports $\mathcal{H}_2$ and $\mathcal{H}_3$. The results regarding Δ_2 support both $\mathcal{H}_2$ and $\mathcal{H}_3$ but discards $\mathcal{H}_1$: if the skill of the modeller would dominate – as hypothesized by $\mathcal{H}_1$, $\mathcal{M}^{\circ}(A, C)$ and $\mathcal{M}^{\circ}(B, C)$ would tend *not* to be generally correlated. Conversely, Δ_3 supports $\mathcal{H}_1$ and $\mathcal{H}_3$ but discards $\mathcal{H}_2$ (if the modellee's skill were to dominate, $\mathcal{M}^{\circ}(A, B)$ and $\mathcal{M}^{\circ}(A, C)$ would tend *not* to correlate). In summary, various indices support the 3 hypotheses, which implies there is some truth in each of them, but $\mathcal{H}_3$ is the only hypothesis that is not rejected by any index. We may hence, with great caution, conclude that the social perspective ($\mathcal{H}_3$) is moderately reinforced by this second study. Since the correlation values for Δ_2 and Δ_3 are similar, we do not interpret their minor difference as evidence for a stronger role of the modeller ($\mathcal{H}_1$) or the modellee ($\mathcal{H}_2$).

We have also to bring some nuances to the social viewpoint ($\mathcal{H}_3$). The main feature that can be associated to the team level in this experiment is probably not the quality of their verbal interactions per se (they interact mostly by drawing on a shared map), but rather the consistency of the spatial exploration strategy: a clear strategy facilitates memorizing one's partner path. One could argue whether the team strategy can be dissociated from the team interactions quality or constitutes one of its components.

Study 3: Effect of media richness on $\mathcal{M}^{\circ}(A, B)$ in argumentation

The aim of this unpublished[1] study was to evaluate the effect of media richness on $\mathcal{M}^{\circ}(A, B, emotions)$. The hypothesis was that video communication would lead to a better $\mathcal{M}^{\circ}(A, B)$ than audio only since emotions often impact facial expressions.

Experimental settings

Triads had to address an emotional societal debate: authorizing or not adoption by homosexual couples. They worked on-line and had to structure their argumentation with the shared concept map tool TEAMWAVE as illustrated in Fig. 7. Ten groups had only an audio connection while ten groups had audio and video. The video communication was provided by

[1]The data and statistical analyses of this study are available online: https://github.com/chili-epfl/mutual-modelling-emotions-study.

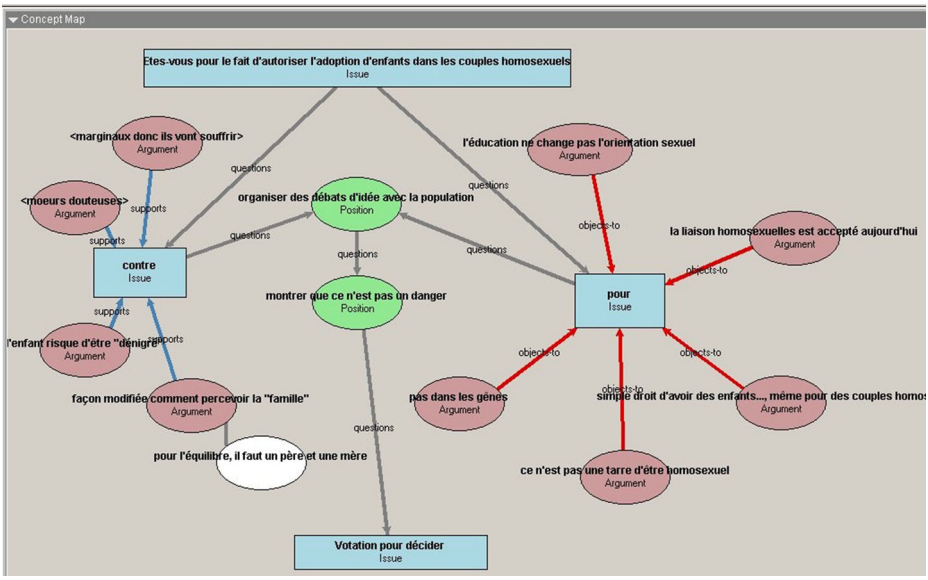


Fig. 7 Example of argumentation graph

a webcam and the software IVISIT. For the audio link, we used microphones, headsets and the BATTLECOM software. In the *audio + video* condition, the screen was divided in three sections. The main part was devoted to the concept map window, and the images of the two peers appeared next to it. In the audio condition, this video zone was left empty so that the size of the concept map was equal in each condition. The subjects were located in the same room, separated by mobile walls. Despite their headsets, non-verbal audio cues (e.g. tapping the floor with feet) were possibly heard by the participants. The task lasted in average 61 minutes.

Subjects

Sixty students (twenty triads) from the University of Geneva participated to this experiment (36 women and 24 men). We formed groups of subjects who knew each other: the task required the discussion of sensitive issues which required to feel quite comfortable with peers. Since groups were formed *a priori*, we did not balance gender in each condition.

Variables

The independent variable was the presence or not of a video link. The dependent variable $\mathcal{M}^{\circ}(A, B)$ was measured subjectively from three questionnaires: in the first one, A described his/her own emotions $\mathcal{M}(A, A)$, while in the two other questionnaire, A described B's and C's emotions. The questionnaire included 18 items (7-point Likert Scale) describing emotions labeled as adjectives: *anxious*, *enthusiastic*, *agitated*, *proud*, *excited*, *quiet*, *calm*, *stressed*, *bored*, *upset*, *relaxed*, *irritated*, *determined*, *hostile*, *active*, etc. $\mathcal{M}(A, B)$ was modelled as a vector of 18 numerical values corresponding to their answers on each

questionnaire items, and $\mathcal{M}^{\circ}(A, B)$ was computed as the distance between the two vectors $\mathcal{M}(A, B)$ and $\mathcal{M}(B, B)$, hence *the smaller the score, the more accurate the model*:

$$\mathcal{M}^{\circ}(A, B, emotions) = \frac{\sum_{emotions} |\mathcal{M}(A, B, e) - \mathcal{M}(B, B, e)|}{18}$$

Results

Grounding criterion The maps produced by teams were ranked by three independent judges on completeness and structure quality (Kendall's $W = 0.474$, limited agreement). We used the average rank as a estimation of the team performance and correlated it with the average of the 6 values of $\mathcal{M}^{\circ}(A, B)$ per team ($\mathcal{M}^{\circ}(A, B)$, $\mathcal{M}^{\circ}(A, C)$, $\mathcal{M}^{\circ}(B, A)$,...). The correlation is 0.22: teams with a good $\mathcal{M}^{\circ}(A, B)$ tend to be better ranked.

Study-specific questions Our hypothesis about media richness is rejected: the average degree of accuracy for $\mathcal{M}(A, B, emotions)$ was 1.25 ($SD = 0.53$) in the audio+video condition and 1.09 ($SD = 0.41$) in the audio alone condition ($t = 1.89, df = 111, p = 0.062$). The smaller distance between $\mathcal{M}(A, B, emotions)$ and $\mathcal{M}(B, B, emotions)$ in the audio condition shows that, in average, the degree of accuracy of $\mathcal{M}(A, B, emotions)$ is *higher* in the audio alone condition.

Symmetry question We computed the absolute differences Δ_1 between $\mathcal{M}^{\circ}(A, B)$ and $\mathcal{M}^{\circ}(B, A)$ over all pair of subjects within a triad (3 values per triad, 20 triads), and compared them with the same differences computed from random gradings (following the same grade distribution as for the experimental data). A t-test on the two sets revealed a significantly lower average difference in the experimental data (mean difference: 0.40 vs 0.54 with random gradings, $t = -3.3, df = 60, p = 0.0016$), which confirms the symmetry of mutual modelling.

Triangle questions Δ_2 is computed in a similar way as the average of the absolute differences between $\mathcal{M}^{\circ}(A, C)$ and $\mathcal{M}^{\circ}(B, C)$ over the 60 subjects. The average difference is 0.42 ($SD = 0.34$), and is not significantly different from the same index computed from random gradings ($t = -1.61, df = 60, p = 0.11$).

For Δ_3, the average absolute difference between $\mathcal{M}^{\circ}(C, A)$ and $\mathcal{M}^{\circ}(C, B)$ over the 60 subjects is 0.40 ($SD = 0.41$) and is significantly lower than chance ($t = -2.62, df = 60, p = 0.01$): a given subject tends to model its two partners with similar degrees of accuracy.

Rectangle question We cannot address the relationship Δ_4 between self and social accuracy here because we do not have an estimation of self-accuracy: subjects describe their own emotions but we have no way to check if they are correct. By measuring $\mathcal{M}^{\circ}(A, B)$ on 18 emotional labels, we can however have a glimpse about Δ_5: how $\mathcal{M}^{\circ}(A, B, X)$ varies according to X. Figure 8 shows the range of modelling errors: the difference between $\mathcal{M}(A, B)$ and $\mathcal{M}(B, B)$, on a scale of 7, is 0.3 in average for the emotion *discasted*, and up to 1.9 for the emotion *calm*. This is probably specific to the variety of scales ($SD = 0.5$ for *discated* versus $SD = 0.7$ for *calm*). Our point is not to interpret this too far, but to show that there are large variations even within one area (perceiving emotions). These variations still question the existence of a general aptitude to model others ($\mathcal{H}_1$).

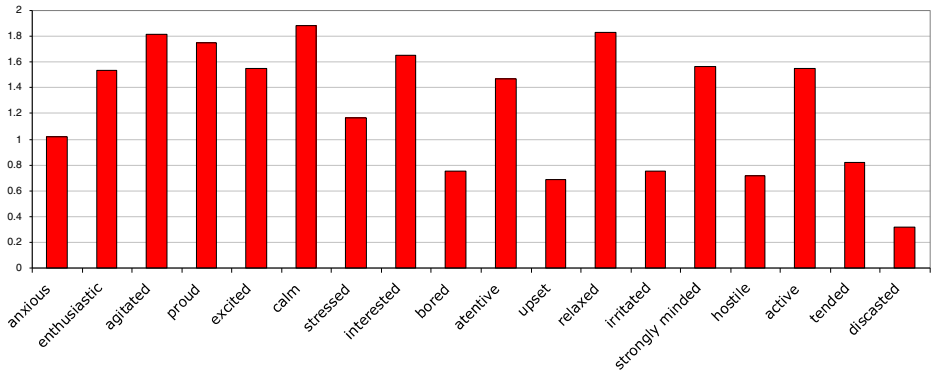


Fig. 8 Average values of $\mathcal{M}^{\circ}(A, B, X)$ where X is one of the proposed emotions (max = 7)

Discussion

In this third study, the accuracy of mutual modelling between two peers tend to be symmetrical, which supports $\mathcal{H}_2$ and $\mathcal{H}_3$. This result is contradicted by the fact that Δ_2 is not significant, which contradicts both $\mathcal{H}_2$ and, to a lower extent, $\mathcal{H}_3$. Finally, Δ_3 supports both $\mathcal{H}_1$ and $\mathcal{H}_3$: this study brings some supports to $\mathcal{H}_3$, but reveals again that $\mathcal{H}_3$ is only part of the explanation.

Like the previous one, this study leads us to refine what we mean by "quality of interactions" in $\mathcal{H}_3$. We expected that the video channel would help peers building a more accurate model of each other's emotions. The results show the opposite: peers in the audio-only condition built more accurate models, probably because they concentrated more on the shared concept map. This confirms other studies that revealed that viewing what one's partner sees (shared graphical editor) is more important than seeing each other (Gaver et al. 1993; Anderson et al. 1997). Therefore, what is meant by the quality of interactions in $\mathcal{H}_3$ is more than the linguistic features of dialogue but includes the way these interactions are articulated to the task.

Study 4: Effects of a script on $\mathcal{M}^{\circ}(A, B)$ in concept mapping

This study investigated the effect of a collaboration script on collaborative learning (Molinari et al. 2008). The script chosen is a JIGSAW: two students receive different but complementary subsets of the knowledge (texts) which have to be integrated to build a shared concept map. This script increases the cognitive effort to build the map, not only to conciliate the viewpoints of each team member but, before that, to find out what the other knows.

Experimental settings

The instructional material consisted of an explanatory text about the neurophysiologic phenomenon of *action potential*. The text was divided into 3 chapters. In the *same information* (SI) condition, the same text was given to both partners. In the *complementary information* (CI) condition, it was divided into two sub-texts, one about the electrical processes of the neuron while the second one about the chemical processes. These two versions were equivalent in terms of number of information pieces.

The peers were located in two rooms equipped with the same computer. The experimental session lasted around 90 minutes and consisted of 6 phases: Participants used two software components, CMAPTOOLS and TEAMSPEAK.

1. As a pre-test, participants were asked to write down all they knew about the neuron and its functioning (5 minutes),
2. Participants were instructed to read a text (12 minutes),
3. Participants were asked to build individually a concept map in order to graphically represent what they learnt from the text (10 minutes),
4. Dyads had to built a concept map during 20 minutes, communicating by audio. The screen layout was structured into three areas (Fig. 9),
5. Participants were invited to individually complete a knowledge test (15-20 minutes),
6. Participants where asked to estimate their own- and their partner's final knowledge in a questionnaire.

Subjects

Fifty-eight first year students from EPFL (47 men and 11 women, mean age: 20.46) were remunerated for participation. Dyads were randomly assigned to one of the two experimental conditions. Gender was balanced over the conditions. Participants did not know each other before the experiment. Students from the Life Sciences faculty were not recruited to avoid high initial background knowledge on the learning domain.

Variables

The independent variable, script versus no-script, was implemented by the difference of texts that individuals had to read. The dependent variables were the post-test scores, used

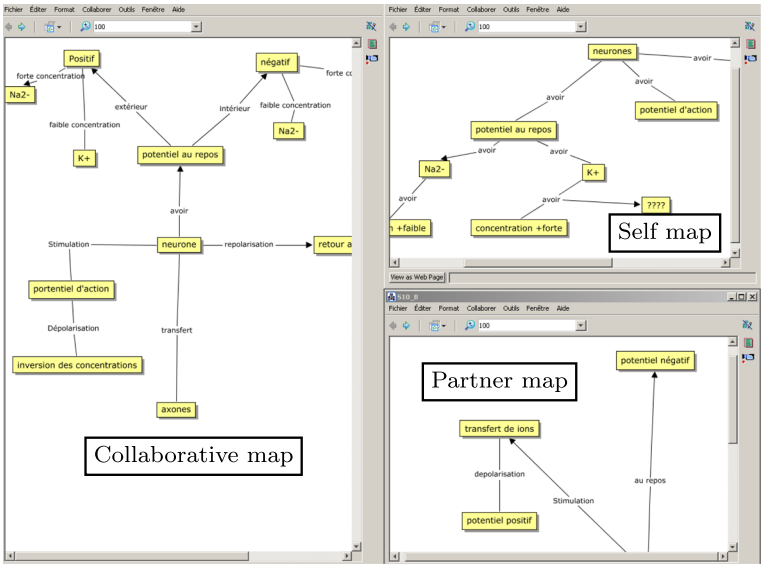


Fig. 9 The group concept map and the individual concept maps in study 4

to assess $\mathcal{M}^{\circ}(A, B, knowledge)$. In phase 6, participants were asked to estimate (7-point Likert scale) their own and their partner's outcome knowledge with respect to each chapter of the learning material. The order of questions about oneself and about the other was counterbalanced across participants.

Results

Grounding criterion In this task, the grounding criterion was low. We did not, however, evaluate task performance (e.g. the quality of the jointly produced concept map) but learning gains. The correlation between $\mathcal{M}^{\circ}(A, B)$ and A's learning gains is not significant ($\beta = 0.08, ns, N = 60$). It is also not significant within each condition.

Study-specific questions We performed a non-parametric Mann-Whitney test on post-test scores for the questions touching the electrical inner working of neurons, and a one-way ANOVA on scores for chemistry-related questions (Levene tests for homogeneity of variances): $p = 0.02$ and ns, respectively. Results did not show any significant difference between the *same information* (SI) condition and the *complementary information* (CI) condition, neither for electrical-related questions ($U = 388.50, z = -0.88, ns$), nor for chemistry-related questions ($F(1, 58) = 0.17, ns$).

The effect of scripts on $\mathcal{M}^{\circ}(A, B)$ was not significant ($F(1, 58) = 0.78, ns$) when considering the absolute difference between $\mathcal{M}^{\circ}(A, B)$ and B's post-test score. However, A tended to underestimate B's score in the SI condition ($M = -2.06$) and to overestimate it in the CI condition ($M = 1.21$) ($F(1, 58) = 6.44, p < 0.01$). Regarding $\mathcal{M}^{\circ}(A, A)$, there was no significant difference between conditions.

Symmetry question The inter-class correlation between $\mathcal{M}^{\circ}(A, B)$ and $\mathcal{M}^{\circ}(B, A)$ is approaching significance ($r = 0.26, F(1, 29) = 1.71, p = 0.075$). It is indeed significant when students read the same text (SI condition: $r = 0.43, F(1, 15) = 2.53, p < 0.05$) but not when they read different texts ($r = 0.13, F(1, 12) = 1.3, ns$).

Rectangle question The correlation between $\mathcal{M}^{\circ}(A, A)$ and $\mathcal{M}^{\circ}(A, B)$ (Δ_4) is globally not significant ($r = 0.05$), and neither it is in each of the conditions: someone good at self-modelling is not necessarily good at modelling someone else and vice-versa. This seems to indicate that partner modelling requires different skills than meta-cognition, despite their similarity at some level of abstraction.

Discussion

In this experiment, the symmetry of mutual modelling is found but in one condition, namely when subjects receive the same information before the task. This condition corresponds to the situation tested in the three previous studies and supports $\mathcal{H}_2$ and $\mathcal{H}_3$.

How do we interpret the fact that the symmetry vanishes when peers receive different texts to read before the task? One explanation would be the difficulty of mutual modelling when peers do not know what the others read, but we found no significance of $\mathcal{M}^{\circ}(A, A)$ between conditions. Since texts were partly overlapping, another explanation is that, in absence of these initial common grounds, mutual modelling requires A to make evident to B what A thinks B does not know about A, which is the second level of modelling described in $\mathcal{H}_2$. A low symmetry means that some peers are better than others at this second level of modelling, which supports the existence of such an individual skill, as stated in $\mathcal{H}_2$.

Study 5: $\mathcal{M}^{\circ}(A, B, knowledge)$ in concept mapping

This study investigates if $\mathcal{M}^{\circ}(A, B, knowledge)$ is related to learning outcomes by comparing teams with or without a Knowledge Awareness Tool (KAT), i.e. a tool that informs A about B's knowledge as measured through a pre-test.

Experimental setting

The peers were located into two different rooms. A complete description of the study is provided in Sangin et al. (2008). The experiment lasted 90 minutes.

It started with the same two first steps as in study 4, followed by:

3. Subjects passed a pre-test, with ten questions per chapter.
4. Participants had 20 minutes to draw a collaborative concept map reporting the content of the texts. They were able to communicate orally through headsets. We used Tobbii eye tracking devices to record their gazes.
5. The post-test included the same items than the pre-test but in a different order.
6. Finally, participants were asked to estimate their partner's knowledge at the post-test for each of the three chapters on a 7-point Likert-like survey.

Subjects

Sixty-four first year EPFL students (46 men, 18 women, mean age: 21.2) participated to the study. They were randomly assigned to conditions and did not know each other before. Like in Study 4, students from Life Science faculty were excluded.

Variables

The participants of the experimental condition group were provided with the Knowledge Awareness Tool on the bottom part of the screen (Fig. 10): each line represents the score obtained by the partner at the pre-test for a chapter. Participants did not see their own score, but they usually started their discussion by exchanging this information.

Results

Grounding criterion A linear regression revealed a positive relation between $\mathcal{M}^{\circ}(A, B)$ and the learning gain of the pair $\{A, B\}$ (calculated as the average of the individual learning gains): $\beta = 0.401, p < 0.001, r2adj. = 0.15$, large effect. This relationship was not significant in the previous experiment which was conducted with the same task (using the same task should produce the same grounding criterion). This is probably due to the fact that $\mathcal{M}^{\circ}(A, B)$ was influenced by the KAT.

Study-specific questions The t-test reported a significant difference between the KAT condition participants ($M = 13.4$ %) and the control group $M = 3.6$ % [$t(1, 60) = 2.73, p < 0.01$, Cohen's $d = 0.7$, medium to large effect]: providing learners with cues about the prior-knowledge of their partner enhances their collaborative learning. As a treatment check, we found a positive and significant correlation between the amount of gazes on KAT (using eye tracking devices) and the learning gains ($r(22) = 0.54, p = 0.01$). A

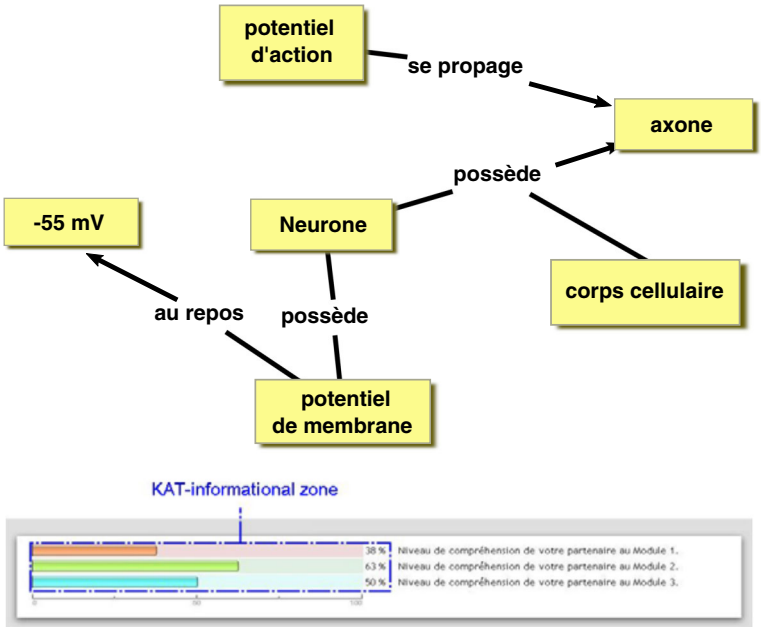


Fig. 10 Screenshot of the KAT condition during the concept-map building phase

detailed analysis revealed that the participants look at KAT to assess their peer's credibility when he/she provided new information.

The KAT has a significant effect on $\mathcal{M}^{\circ}(A, B)$: peers more accurately estimated their partners knowledge ($M = 1.11$) than those in the control condition $M = 0.98$ [$t(1, 60) = 3.19$, $p < 0.01$, Cohen's $d = 0.83$, large effect]. This is a trivial result since the KAT provided them with an initial $\mathcal{M}(A, B)$. However, the participants have to predict the post-test score while the KAT informed them about the pre-test score. Actually, pairs in the KAT condition produced significantly more instances of 3 interesting categories of interactions: (1) utterances asking about the other's knowledge such as "Did you understand how transmission works?" (2) utterances describing one's own knowledge ($\mathcal{M}(A, A)$) such as "I don't remember the Ranvier's thing..." and (3) elaborated utterances with rich contents. These three categories provide different account of $\mathcal{M}^{\circ}(A, B)$: (1) as an effect of A's effort to model B ($\mathcal{H}_1$), (2) as B's effort to give cues to A about his own knowledge ($\mathcal{H}_2$) and (3) as an effect of the quality of interaction ($\mathcal{H}_3$) .

We examined $\mathcal{M}^{\circ}(A, B)$ as potentially mediating the effect of the KAT factor on the relative learning gain. A linear regression confirmed that $\mathcal{M}^{\circ}(A, B)$ was significantly related to the KAT-factor ($\beta = 0.381$, $p < 0.01$, $r2 = 0.15$). The KAT-factor was also significantly and positively related to the RLG ($\beta = .332$, $p < 0.01$, $r2 = 0.11$). We then tested the relation between the independent variable (KAT) and the dependent variable (gains) when controlling for the mediating variable ($\mathcal{M}^{\circ}(A, B)$). A multiple regression showed that the KAT-factor was no longer a significant predictor ($\beta = 0.210$; $p = ns$) whereas the $\mathcal{M}^{\circ}(A, B)$ was still a significant predictor ($\beta = 0.32$, $p < 0.01$). Thus, it can be concluded that $\mathcal{M}^{\circ}(A, B)$ mediated the KAT-factor's effect on the learners' RLG. The Sobel significance test for indirect effects was significant [$z = 1.99$, $p < 0.05$].

Symmetry question We did not find an intra-pair correlation ($ICC = 0.05, ns$), which does not support $\mathcal{H}_2$ and $\mathcal{H}_3$. In addition, the KAT supports the processing of modelling at the first level ($\mathcal{H}_1$) and not at the second level ($\mathcal{H}_2$). Hence the fact that the KAT enhances $\mathcal{M}^{\circ}(A, B)$ brings additional supports to $\mathcal{H}_1$. However, the analysis of verbal interactions finds elements that actually support the three hypotheses.

Discussion

Despite the fact that the learning task was the same as in study 4, the conditions of collaboration (viewing multiple maps or not) and the conditions (scripted or not, awareness tool or not) probably explain differences in terms of mutual modelling.

Synthesis

In the introduction, we mentioned the controversy around the cognitive depth of dialogues: does efficient dialogue require some modeling of what the partner understands or intends to convey, as initially postulated by Clark and Wilkes-Gibbs (1986), or can dialogue simply rely on some shallow syntactic alignment, as objected by Pickering and Garrod (2006). While the shallow hypothesis may be relevant to simple chat situations, we investigated this issue in richer problem solving tasks that are more representative of the tasks assigned to learners in CSCL environments. In four different tasks, we found evidence that partners model each other since the quality of modelling emerged as an intermediate variable, sensitive to several independent variables (like the presence of awareness tools or media richness), and predictive of several dependent variables, such as task performance or learning gains. Therefore, even if we do not bring any definite conclusion to this debate, our results support the CSCL school of thought in which shared understanding or intersubjective meaning making have been a foundational concept (Roschelle and Teasley 1995; Schwartz 1995; Dillenbourg and Traum 2006; Suthers 2006; Stahl 2007).

Second, we questioned whether the accuracy of the partner's model depends on the cognitive skills of each partner or instead results from the quality of the interactions among them. We respectively refer to these hypothesis as $\mathcal{H}_1$ and $\mathcal{H}_3$. Our rationale was that a symmetry of mutual modelling (correlation between the accuracy of each other's model) would favor $\mathcal{H}_3$ over $\mathcal{H}_1$. We also formulated $\mathcal{H}_2$, by which a learner A who is good at modelling B would also help B to repair his inaccurate representations of A.

We found a symmetry of the mutual models on 4 studies out of 5. In study 4, this only applies to the control group (having read the same text before), which corresponds to the situation of the 3 first studies. Evidencing this symmetry constitutes *per se* an interesting result as we are not aware of earlier studies that have established this relationship. Still, the symmetry alone does not allow to discriminate $\mathcal{H}_2$ from $\mathcal{H}_3$. The second hypothesis is questioned by study 3 where çould be modelled accurately by A and not accurately by B. The same hypothesis is however supported by the results of study 2, and indirectly by study 4.

This does not rule out entirely $\mathcal{H}_1$ (i.e. partner modelling is primarily a individual skill) either: even where we found significant correlations, they were all below 0.50, not around 0.90. Hence, even if the quality of the social interaction matters, there is obviously a large part of individual variance within teams.

In other words, these studies do not conclude that one hypothesis is right and the others are wrong, and this is indeed the main contribution of this article. We started from the idea that partner modelling is essentially an individual skill and that we would therefore improve the quality of collaboration by providing awareness tools, as those tools would act as a kind of prosthesis for partner modelling. The role and importance of this individual skill cannot be discarded: everyone has experienced the pleasure of interacting with colleagues who answer precisely to the question we asked them and even guess the reason why we asked this question. Conversely, everyone also experienced the frustration of someone referring to a third person by his name, say Mike Smith, while knowing perfectly that there is no chance that we know this Mike Smith. In fact, the role of this individual skill is not denied by our studies; the novelty of this paper is to show, thanks to symmetry values, that individual skills only account for a part of the accuracy in mutual modelling.

We acknowledge that the difference between these hypotheses is rather theoretical since the process of modelling one's partner $\mathcal{H}_1$ and the process of helping one's partner to model oneself $\mathcal{H}_2$ are mediated by verbal interactions in the team. It is difficult to imagine someone managing a very accurate modeling despite low quality interaction. There is a bidirectional causal link between the accuracy of mutual modelling and the quality of the interactions. This rather artificial distinction does however contribute to the more fundamental discussion of the role of individual and social mechanisms in human cognition. In the field of social cognition, it is commonplace to state that "the whole is greater than the sum of the parts". This refers to the emergence of team properties than cannot be reduced to the set of individual contributions. Our paper illustrates this emergence by showing the symmetry of mutual modelling. In a nutshell, yes there is a non-negligible component of mutuality in modelling one's partner.

These conclusions must be presented with multiple disclaimers. First, they heavily rely on correlations; hence we cannot identify causal links. Second, we faced difficult methodological issues. Providing learners with on-task questionnaires introduces a bias: they will pay more attention to their partners in the remaining time. Providing them with "after-task" questionnaires implies mnemonic and rationalization biases. The nature of mutual modelling implies methodological challenges that call for new measurement methods. We have promising results for using eye tracking methods to address this challenge. Third, our results emerge from a post-hoc reinterpretation of studies that addressed different research questions (media richness, awareness tools, scripts,...). This diversity makes our results difficult to integrate as they appear partly contradictory. Nonetheless, this diversity also incidentally provides some generalizability: mutual modelling has been investigated in different contexts (virtual space versus real space), with different groups sizes (pairs and triads) and different tasks.

These limitations and our difficulty to provide clear-cut conclusions comes from the fact that this series of experiments was not planned a priori. This paper relies on the post-hoc comparison of experiments conducted across various contexts. The overall conclusion is that this research question would deserve a specific research agenda for 3 reasons: (1) it raises fundamental theoretical dilemma on the social nature of cognition, (2) it raises methodological challenges and (3) it could provide empirical grounds to design decisions for CSCL environments. The eye tracking tools we developed and the notation we used for referring to different components of mutual modelling could pave the road for elaborating a systematic agenda for research on mutual modelling. Let us repeat that the expressions we used, such as $\mathcal{M}(A, B)$, do not imply we have a mechanical view of modelling, but were simply useful ways to talk about mutual modelling.

Acknowledgments The experiments were developed with the help of Mirweis Sangin, René Glaus, Patrick Jermann, Fabien Girardin, Marc-Antoine Nüssli, Thomas Werhle, Yvan Bourquin and Jeremy Goslin. We also thank Kshitij Sharma and Łukasz Kidziński for their help with data analysis. The main funding has been provided from a NSF Grant grant #102511-106940.

References

Anderson, A.H., O'Malley, C., Doherty-Sneddon, G., Langton, S., Newlands, A., Mullin, J., Fleming, A.M., & Van der Velden, J. (1997). The impact of VMC on collaborative problem solving: An analysis of task performance, communicative process, and user satisfaction., Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, pp 133–155. Computers, cognition, and work.

Aronson, E., Blaney, N., & Stephan, C. (1978). Sikes J, The jigsaw classroom. Sage Publications, Snapp M.

Baker, M., Hansen, T., Joiner, R., & Traum, D. (1999). The role of grounding in collaborative learning tasks. Collaborative learning: Cognitive and computational approaches 31–63.

Blaye, A., & Light, P. (1995). Collaborative problem solving with HyperCard: the influence of peer interaction on planning and information handling strategies. In: Computer supported collaborative learning, Springer, pp 3–22.

Brennan, S.E. (1991). Conversation with and through computers. *User Modeling and User-Adapted Interaction*, *1*(1), 67–86.

Cherubini, M., Van Der Pol, J., & Dillenbourg, P. (2005). In *Grounding is not shared understanding: Distinguishing grounding at an utterance and knowledge level CONTEXT'05, the Fifth International and Interdisciplinary Conference on Modeling and Using Context* (pp. 11–23).

Chi, M.T., Bassok, M., Lewis, M.W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive science*, *13*(2), 145–182.

Clark, H.H., & Brennan, S.E. (1991). Grounding in communication. *Perspectives on socially shared cognition*, *13*(1991), 127–149.

Clark, H.H., & Marshall, C.R. (2002). Definite reference and mutual knowledge. Psycholinguistics: critical concepts in psychology 414.

Clark, H.H., & Schaefer, E.F. (1989). Contributing to discourse. *Cognitive science*, *13*(2), 259–294.

Clark, H.H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39.

Dillenbourg, P., & Hong, F. (2008). The mechanics of cscl macro scripts. *International Journal of Computer-Supported Collaborative Learning*, *3*(1), 5–23.

Dillenbourg, P., & Traum, D. (2006). Sharing solutions: Persistence and grounding in multimodal collaborative problem solving. *The Journal of the Learning Sciences*, *15*(1), 121–151.

Dillenbourg, P., Baker, M.J., Blaye, A., & O'Malley, C. (1995). The evolution of research on collaborative learning. Learning in Humans and Machine: Towards an interdisciplinary learning science 189–211.

Doise, W., Mugny, G., & Perret-Clermont, A.N. (1975). Social interaction and the development of cognitive operations. *European journal of social psychology*, *5*(3), 367–383.

Gaver, W.W., Sellen, A., Heath, C., & Luff, P. (1993). One is not enough: Multiple views in a media space. In *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems* (pp. 335–341).

Hutchins, E., & Palen, L. (1997). Constructing meaning from space, gesture, and speech. In *Discourse, Tools and Reasoning* (pp. 23–40): Springer.

Kenny, D.A., Kashy, D.A., Bolger, N., & etal (1998). Data analysis in social psychology. *The handbook of social psychology*, *1*(4), 233–265.

Kobbe, L., Weinberger, A., Dillenbourg, P., Harrer, A., Hämäläinen, R., Häkkinen, P., & Fischer, F. (2007). Specifying computer-supported collaboration scripts. *International Journal of Computer-Supported Collaborative Learning*, *2*(2-3), 211–224.

Liben, L.S., Patterson, A.H., & Newcombe, N. (1981). Spatial representation and behavior across the life span. Academic Press.

Molinari, G., Sangin, M., Nüssli, M.A., & Dillenbourg, P. (2008). Effects of knowledge interdependence with the partner on visual and action transactivity in collaborative concept mapping. In *Proceedings of the 8th International Conference on International Conference for the Learning Sciences - Volume 2* (pp. 91–98).

Moreland, R.L. (1999). Transactive memory: Learning who knows what in work groups and organizations. Shared Cognition in Organizations: The Management of Knowledge.

Nova, N., Girardin, F., & Dillenbourg, P. (2005). Location is not enough!: an empirical study of location-awareness in mobile collaboration. In *IEEE International Workshop on Wireless and Mobile Technologies in Education* (pp. 21–28).

Nova, N., Girardin, F., Molinari, G., & Dillenbourg, P. (2006). The underwhelming effects of automatic location-awareness on collaboration in a pervasive game. In *Cooperative Systems Design: Seamless Integration of Artifacts and Conversations-Enhanced Concepts of Infrastructure for Communication* (pp. 224–238).

Nova, N., Wehrle, T., Goslin, J., Bourquin, Y., & Dillenbourg, P. (2007). Collaboration in a multi-user game: impacts of an awareness tool on mutual modeling. *Multimedia tools and Applications*, *32*(2), 161–183.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, *38*(1), 1–4.

Pickering, M.J., & Garrod, S. (2006). Alignment as the basis for successful communication. *Research on Language and Computation*, *4*(2-3), 203–228.

Roschelle, J., & Teasley, S.D. (1995). The construction of shared knowledge in collaborative problem solving. In *Computer supported collaborative learning* (pp. 69–97): Springer.

Sangin, M., Molinari, G., Nüssli, M.A., & Dillenbourg, P. (2008). How learners use awareness cues about their peer's knowledge?: insights from synchronized eye-tracking data. In *Proceedings of the 8th international conference on International conference for the learning sciences-Volume 2, International Society of the Learning Sciences* (pp. 287–294).

Schober, M.F. (1993). Spatial perspective-taking in conversation. *Cognition*, *47*(1), 1–24.

Schwartz, D.L. (1995). The emergence of abstract representations in dyad problem solving. *The Journal of the Learning Sciences*, *4*(3), 321–354.

Slugoski, B.R., Lalljee, M., Lamb, R., & Ginsburg, G.P. (1993). Attribution in conversational context: Effect of mutual knowledge on explanation-giving. *European Journal of Social Psychology*, *23*(3), 219–238.

Stahl, G. (2007). Meaning making in CSCL: Conditions and preconditions for cognitive processes by groups. In *Proceedings of the 8th iternational conference on Computer Supported Collaborative Learning* (pp. 652–661).

Suthers, D.D. (2006). Technology affordances for intersubjective meaning making: A research agenda for cscl. *International Journal of Computer-Supported Collaborative Learning*, *1*(3), 315–337.

Traum, D., & Dillenbourg, P. (1996). Miscommunication in multi-modal collaboration. In *AAAI Workshop on Detecting, Repairing, and Preventing Human–Machine Miscommunication* (pp. 37-46).

Webb, N.M. (1991). Task-related verbal interaction and mathematics learning in small groups. Journal for research in mathematics education 366–389.

Wegner, D.M. (1987). Transactive memory: A contemporary analysis of the group mind. In *Theories of Group Behavior* (pp. 185–208): Springer.