**Research Report**

# Agent-Based Modeling of Collaborative Problem Solving

**Yoav Bergner**

**Jessica J. Andrews**

**Mengxiao Zhu**

**Joseph E. Gonzales**

# ETS Research Report Series

RESEARCH REPORT

# Agent-Based Modeling of Collaborative Problem Solving

Yoav Bergner,[1] Jessica J. Andrews,[1] Mengxiao Zhu,[1] & Joseph E. Gonzales[2]

1 Educational Testing Service, Princeton, NJ
2 University of California, Davis

Collaborative problem solving (CPS) is a critical competency in a variety of contexts, including the workplace, school, and home. However, only recently have assessment and curriculum reformers begun to focus to a greater extent on the acquisition and development of CPS skill. One of the major challenges in psychometric modeling of CPS is collecting large-scale data on teams and processes. In this study, we explore the use of agent-based modeling (ABM) to model the CPS process, test the sensitivity of outcomes to different population characteristics, and generate simulated data that can provide a novel means by which to refine and develop psychometric models. Methods of adapting trait-based stochastic processes to a specific task are described, and preliminary results are presented.

Collaborative problem solving (CPS) is a critical competency in a variety of contexts, including the workplace, school, and home. In fact, competency in collaboration has been identified as one of the most important skills for the 21st century workforce (Burrus, Jackson, Xi, & Steinberg, 2013). As students transition into the workforce, they will often be expected to work in teams to solve complex problems, make decisions, and generate novel ideas, each of which requires cooperating and communicating effectively with others and resolving potential conflicts. Despite the importance and relevance of this 21st century skill for the above-mentioned contexts, only recently has assessment and curriculum reform begun to focus to a greater extent on the acquisition and development of CPS skill (Bennett & Gitomer, 2009; National Research Council, 2011).

One of the major challenges in CPS research is collecting large-scale data on teams and processes. Moreover, without relatively large sample sizes and fine-grained interaction logs, it is almost impossible to develop quantitative psychometric models of CPS ability. In conditions in which real data are difficult to come by, simulation can sometimes be a reasonable direction forward. In this study, we explore the use of agent-based modeling (ABM) to simulate the CPS process. ABM can be used to test the sensitivity of outcomes to different population characteristics and interaction rules. Additionally, ABM can be used as a generative model for synthetic collaboration process data, which can provide a novel means by which to refine and develop psychometric models.

The organization of this paper is as follows: We first review substantive issues in CPS. We consider the applicability of ABM briefly in general and then focus on a specific simulated task. We then describe the operationalization of traits in our simulation and present sample results, followed by a discussion of future work.

## Background on Collaborative Problem Solving

Competency in CPS is defined as "the capacity of an individual to effectively engage in a process whereby two or more agents attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills, and efforts to reach that solution" (Organisation for Economic Co-operation and Development, 2013, p. 6). This definition identifies the group processes necessary for effective CPS, including establishing and maintaining a shared understanding, identifying and implementing effective problem-solving strategies, and organizing the group to afford effective information sharing practices. Work across a variety of disciplines has explored how group processes such as these can engender learning and optimal performance in group contexts (e.g., Andrews & Rapp, 2015; Dillenbourg & Traum, 2006; Fawcett & Garton, 2005; Mesmer-Magnus & DeChurch, 2009; Stasser & Stewart, 1992). For example,

*Corresponding author:* Y. Berger, E-mail: ybergner@ets.org

educational research has investigated instructional methods that use collaborative activity across a variety of grade levels and subject areas. Such research has demonstrated positive effects of collaboration for a number of outcomes associated with learning and social, emotional, and psychological well-being (Gaudet, Ramer, Nakonechny, Cragg, & Ramer, 2010; Gillies, 2004; Jeong & Chi, 2007; Slavin, 1983). D. W. Johnson, Maruyama, Johnson, Nelson, and Skon (1981) reviewed 122 studies comparing cooperative learning methods with individual and competitive learning methods. The results showed that cooperation facilitated higher achievement better than did competitive and individual learning, with these results consistent across all subject areas and age groups. More recent reviews have reported similar results, demonstrating the robustness of collaboration as an effective method for promoting learning (Bowen, 2000; Goodyear, Jones, & Thompson, 2014; D. W. Johnson, Johnson, & Stanne, 2000).

Organizational research has also explored the effectiveness of collaboration as a means by which to improve performance in organizational contexts, which often uses teams for a range of tasks such as decision-making and problem solving. Much of this work has been guided by an input-process-outcome (I-P-O) theoretical model of teams (Hackman, 1987; McGrath, 1964). In this framework, inputs—preconditions such as member personalities, task structure, and environmental complexity—lead to processes—communication and coordination—that in turn lead to outcomes—team performance or viability. Specifically, there have been investigations into what makes some teams more effective than others, emphasizing inputs that contribute to effective team outcomes as well as the mediating processes by which such inputs impact team outcomes (e.g., Zhu, Huang, & Contractor, 2013). For example, Jehn, Northcraft, and Neale (1999) have shown that informational diversity (i.e., differences in knowledge and perspectives) can positively impact team performance, but this relationship is mediated by task conflict. Furthermore, research has demonstrated how diversity in group-member personality can influence performance outcomes. Variability in agreeableness and neuroticism can negatively affect performance (Mohammed & Angell, 2003); however, teams with higher average agreeableness tend to exhibit better performance (Barrick, Stewart, Neubert, & Mount, 1998; Bell, 2007), perhaps because agreeable members are more likely to engage in the positive interpersonal processes that have been shown to facilitate performance (LePine, Piccolo, Jackson, Mathieu, & Saul, 2008). Diversity in extraversion (Mohammed & Angell, 2003; Neuman, Wagner, & Christiansen, 1999) and higher cognitive ability among team members can positively affect performance outcomes as well (Barrick et al., 1998; Devine & Philips, 2001; Stewart, 2006).

Particular attention has been given to identifying and measuring the team processes necessary in facilitating performance outcomes (Brannick, Prince, Prince, & Salas, 1995; Liu, Hao, von Davier, Kyllonen, & Zapata-Rivera, 2015; Morgan, Salas, & Glickman, 1993; von Davier & Halpin, 2013). Some of this work has focused on ways to assess these processes in a more time-efficient manner with the use of computer simulations (e.g., O'Neil, Chung, & Herl, 1999). For example, O'Neil, Chung, and Brown (1995) designed a computer simulated negotiation task in which distributed three-person teams were assessed according to their demonstration of teamwork skills associated with adaptability, coordination, decision-making, communication, leadership, and interpersonal skills. Teamwork processes were assessed according to communication among group members with the use of predetermined messages available in the computer simulation. Results showed that teamwork processes could be measured reliably and efficiently as teams used the predetermined messages; however, only decision-making was found to be significantly related to team performance outcomes.

Conceptions of the process component of the I-P-O model have been expanded to include not only members' actions, but other mediating mechanisms referred to as emergent states such as trust, team confidence, collective cognition, and cohesion (Ilgen, Hollenbeck, Johnson, & Jundt, 2005). For example, González, Burke, Santuzzi, and Bradley (2003) have shown that task cohesion mediates the relationship between team efficacy and team effectiveness. A direct positive relationship between multiple components of cohesion and performance have also been exhibited (Mullen & Copper, 1994; Zaccaro & McCoy, 1988), and this positive relationship can be moderated by the extent to which a task requires back and forth exchange between group members. Specifically, the relationship between cohesion and performance is stronger when groups engage in more interdependent tasks (Beal, Cohen, Burke, & McLendon, 2003). Group cohesion can also aid in the development of shared mental models which in turn facilitate group performance (Fiore & Salas, 2004; Mathieu, Heffner, Goodwin, Salas, & Cannon-Bowers, 2000).

Concerning the outcome component of the I-P-O model, group performance has sometimes been conceptualized in terms of whether synergy is achieved or whether performance at the group level is beyond what the individual group members are capable of accomplishing separately, and this sort of gain in performance is attributed to group interaction (Larson, 2010). Two forms of synergy are distinguished: Weak synergy, also referred to as gain (Szumal, 2000), includes

instances in which group performance exceeds that of the average individual working independently, whereas strong synergy refers to instances in which group performance exceeds the individual performance of the best group member (Larson, 2007). There is relatively extensive empirical evidence demonstrating the existence of weak synergy (e.g., Laughlin, Gonzalez, & Sommer, 2003; Sniezek, 1989; Tindale & Larson, 1992); fewer studies have demonstrated strong synergy (but see Laughlin, Bonner, & Miner, 2002; Reagan-Cirincione, 1994; Tindale & Sheffey, 2002). Research has further examined factors that can facilitate the types of interactions that lead to synergy, including how much more capable the best group member is relative to other group members (Meslec & Curșeu, 2013), the size of the group (Laughlin, Hatch, Silver, & Boh, 2006), and the mode of communication utilized by group members (Credé & Sniezek, 2003).

Whether group performance is superior to that of individual constituents appears to depend on the outcome measure, on the size of groups, and on the particular task. Prior to the adoption of a distinction between weak and strong synergy, Hill (1982) noted that group performance, although typically superior to that of the average individual, was often inferior to that of the best individual or to the potential achievable in a statistical pooling model. Laughlin et al. (2006) found that groups tend to perform better than the best individual on letters-to-numbers problems, but this pattern did not emerge for dyads. Quite a few studies have shown that groups recall less information than the pooled recall of individuals working independently (e.g., Andersson & Rönnberg, 1995; Barber, Rajaram, & Aron, 2010; Basden, Basden, & Henry, 2000).

## Agent-Based Modeling

Despite the abundance of research, as described above, examining how collaboration can facilitate learning and performance, much of the research relies on observational and qualitative methods or experiments with small numbers of small groups. Quantitative analysis, for example application of a temporal point process or hidden Markov models, is hampered by a shortage of suitable data. An alternative is to use simulation models for individual behavioral studies (Bonabeau, 2002). ABM can be used to explore complex and dynamic relationships in groups (Gilbert, 2007). In an ABM, each agent is an autonomous entity who makes decisions based on rules and parameters of the environment. This characteristic makes ABM an appropriate tool to simulate the activities of individuals, the interactions between the agents, and the interactions between agents and the environment. In particular, such simulations allow researchers to carefully control factors such as problem-solving strategies or individual personality characteristics and study the effect of variations in these variables by resetting different values through successive runs of the simulation (Holland & Miller, 1991). Thus, computer simulations enable researchers to run thousands or millions of trials and quantify within these models (a) the process dynamics, given the rules and initial conditions, and (b) the sensitivity of outcomes to different rules and/or initial conditions. Results from such computer experiments may be useful in determining the conditions under which collaboration is most successful.

A number of projects have indeed begun to demonstrate how ABM can afford such investigations. For example, Larson (2007) modeled the performance of three-person groups engaging in a value-seeking problem in which the problem solver has to choose between sets of solutions that vary in their underlying value. Problem-solving strategy and amount of communication among group members were variables in the model. Results of the simulation showed that groups often outperformed their average member, thus demonstrating weak synergy; however, groups heterogeneous in terms of problem-solving strategy demonstrated strong synergy, particularly when agents communicated their solution alternatives to group members.

Researchers in computational sociology used ABM to explore how network structures affect prosocial behavior (Macy & Willer, 2002). Four properties of network structures were shown to either facilitate or inhibit cooperation among members: relational stability, network density, homophily, and transitivity. Crowder, Robinson, Hughes, and Sim (2012) used ABM to explore how individual level, team level, and task level variables influence team performance in an engineering environment. Individual level variables included competency, motivation, availability, and response rate; team level variables included communication, shared mental models, and trust; and task level variables included difficulty and work flow.

Related studies include simulating the development of shared mental models in teams (Dionne, Sayama, Hao, & Bush, 2010), the process of collaborative product development and design in organizations (Zhang, Li, Zhang, & Schlick, 2013; Zhang et al., 2009), and human-robot teaming structures for military operations (Giachetti, Marcelli, Cifuentes, & Rojas, 2013). Such work has taken into account a number of variables such as social network structure, heterogeneity of agents' domain of expertise, mutual interest (Dionne et al., 2010), cultural differences (Horii, Jin, & Levitt, 2005), goals (Kraus, Sycara, & Evenchik, 1998), and cognitive load (Fan & Yen, 2011).

In building on prior ABM work, we describe first steps toward using an ABM simulation to model the effect of cognitive and noncognitive traits of individual agents on team performance. In addition to testing the sensitivity of outcomes to population characteristics, we would also like to use the simulation to generate simulated process data for development and training of statistical process models. This is perhaps a novel use of ABM. Because we wish to use the process data and not just the outcomes of our ABM, it is not sufficient to provide rules that lead to emergent complexity. Our rules have to be clearly trait- and goal-driven such that logs of the agents' actions are reasonably interpretable. Finally, a longer term goal is also to optimize team composition using the emergent intelligence from the self-organizing process in ABM (Holland, 1975; S. Johnson, 2002). For instance, rather than assign agents to teams, the simulation may allow agents to negotiate their team membership according to affinity rules.

## Simulation Methodology

To construct a suitable collaborative task in this study, we used as a template the Subarctic Survival Situation (2013; see also the Desert Survival Situation in Human Synergistics International, 2012), which is a relatively widely used team exercise (for example, at Harvard Business School) designed to illustrate team dynamics to participants. We modeled a similar process, essentially a collective ranking task, using ABM simulation. The successful outcome of this task depends on individual abilities as well as personality traits and team communication.

The Subarctic Survival Situation task is used here to maintain realism in our simulation design. We see the ABM simulation more generally as a generative model of team dynamics that includes personality variables — rule-based representations of talkativeness, agreeableness, and communication skill — as well as domain knowledge and critical thinking. In the ABM, agent personality variables manifest themselves in rules that determine how agents adjust their own beliefs (here, rank-order of list) based on interactions with other agents in the group (round by round of information sharing). Our goal was to design a minimal set of rules that resulted in high variability of team performance. For example, a team may equilibrate on a suboptimally ranked list because the best member lacked assertiveness or because the worst member was assertive while the other members were unchallenging, or the team may not equilibrate at all. The outcome of the agent-based generative model is both a sensitivity analysis of outcomes to the initial conditions and a set of simulated process data logs recording the agents' moves in each turn. The process data in turn might be used as training data for the development of models that predict group outcomes.

The ABM simulation in this study was implemented using NetLogo (Wilensky, 1999), which is free and open source software developed and maintained by the Center for Connected Learning and Computer-Based Modeling at Northwestern University. NetLogo is flexible enough to permit programming any agent-based systems within its syntax. It also provides an interactive graphical interface, making it easy to test different simulation settings quickly.

## Traits and Rules

We chose to operationalize the following agent traits in the first instantiation of the simulation: *talkativeness*, an agent's willingness to show one's own list for comparison; *agreeableness*, a willingness to consider changing one's list order when presented with an alternative; and *critical thinking*, an ability in evaluating a competing list ordering. In an ABM, traits are encoded through rules that trigger actions taken during turns. Consider the following rule, which would apply to an agent at the beginning of a given turn:

$$\text{IF talkative THEN show list} \tag{1}$$

A deterministic rule such as (1) will not simulate our intuitive sense of human behavior, which is that talkativeness increases the probability of showing rather than dividing agents into those who do or do not show on every turn. Moreover, we tend to model traits in terms of parametric distributions. We considered two options for converting trait distributions into action probabilities, which we describe below.

One possibility is to model agent traits as random variables from a beta distribution. The motivation is that this distribution has support on the interval (0,1), such that values of the trait draws can be interpreted directly as probabilities for an agent's actions. For example, consider the densities shown in Figure 1.

Both densities have a mode at 0.8, though the B(5,2) distribution (dashed line) is less sharply peaked (the means are of course different: 0.714 compared to 0.765). If agent talkativeness is modeled as a random variable drawn from this broader
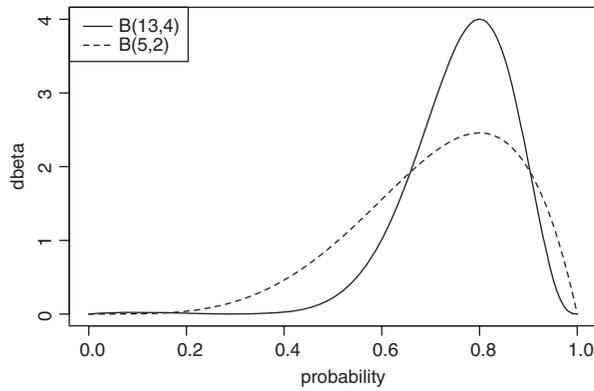
**Figure 1** Beta distributions with shape parameters (13,4), solid line, and (5,2), dashed line.

distribution, then most agents would have a probability of 0.6–0.9 of showing, but a minority of agents would indeed be un-talkative, with a showing probability of 0.2–0.5 on a given turn. A team of four or five agents might thus, for example, have one or two un-talkative members. Using the B(13,4) distribution instead, the range of talkativeness would be more restrictive.

The probabilistic turn for an agent with talkativeness $T$ would be programmed as follows: Draw a random number $x$ from $U(0,1)$ and apply the rule

$$\text{IF } x < T \text{ THEN show list.} \tag{2}$$

The approach described above meets our needs, but it suffers from a lack of intuitiveness about the shape parameters of the beta distribution. Simply put, the mean and variance are not trivial functions of the shape parameters. For a random variable from $B(\alpha, \beta)$, they are given by,

$$E[X] = \frac{\alpha}{\alpha + \beta}, \qquad Var[X] = \frac{\alpha\beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}.$$

In designing simulated experiments, use of beta distributions represents a conceptual difficulty for picturing the trait distributions.

An alternative approach is to model traits using normal distributions $N(\mu, \sigma)$, so that the shape parameters are intuitive, and convert trait values to probabilities using, say, a logistic link function,

$$p(x) = \frac{1}{1 + e^{-x}}.$$

For example, using logistic transforms of variables drawn from $N(1.2, 0.57)$ and $N(0.94, 0.92)$, the densities are qualitatively similar to those of Figure 1, though the shapes are slightly different (see Figure 2).
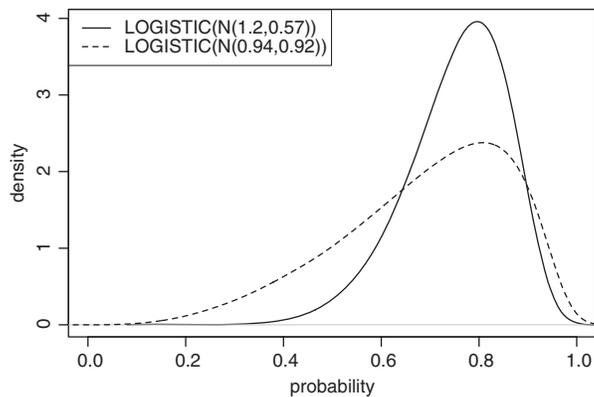


**Figure 2** Densities arising from logistic transforms of two normal distributions.
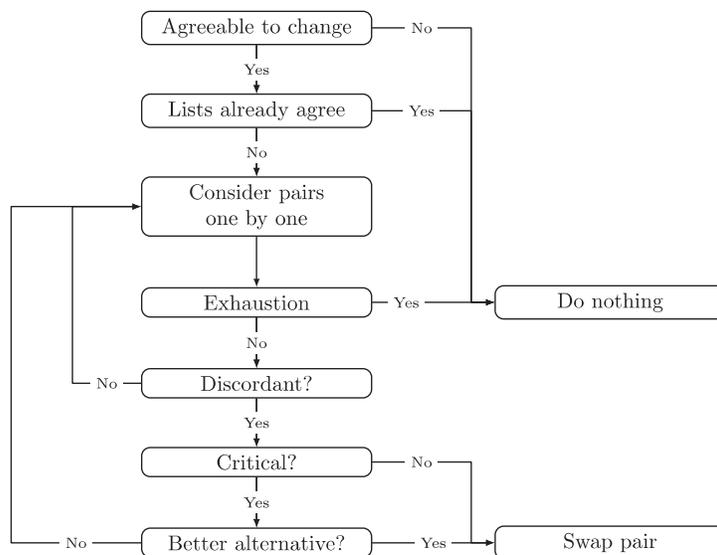
**Figure 3** Schematic of decision process for an agent when comparing lists to a shown (reference) list.

Because normally distributed traits are more intuitive, we chose the second approach described above. The action rule on a given turn for an agent with trait amount $T$ again involves drawing a random variable $x$ from $U(0,1)$ and then applying

$$\text{IF } x < p(T) \text{ THEN show list.} \qquad (3)$$

This way it is more straightforward to simulate a uniform population by setting the variance (or standard deviation) to zero. For example, if we want each agent to show a list each turn (or 99.9% of the time), we can set the talkative distribution to $N(7,0)$.

The worked example above in fact represents the simplest and first decision rule in our model, which is the turn-taking rule with respect to showing one's list. Turns are sampled randomly among the team members, and talkativeness dictates whether or not an agent's list will be considered by the others. If shown, then the remaining agents are prompted to decide whether or not to make updates to their lists. The update rule as currently implemented is shown schematically in Figure 3. The process is described below.

The first rule-based decision concerns whether the agent is agreeable to change on the current turn (as with talkative-ness, a stochastic draw with threshold depending on the agent's agreeableness). If so, the agent then checks whether the list being shown is already identical to the agent's list. If not, then a process of comparing the two lists begins pair by pair in random order, searching for a discordant pair of items, that is two items whose relative ordering differs in the two lists. We have included a global parameter called *knowledge sharing* to represent the idea that during this systematic search, realistic agents might not exhaustively consider all possible pairs. (The intuition might be that two lists may not be dif-ferent enough to overcome inertia.) When knowledge sharing is at its maximum value of 1, then the agent will inspect all possible pairs until a discordant pair is identified or no pairs remain. If knowledge sharing is set to 0.5, then the agent may stop after comparing only half of the pairs (again, in a random order). The proportion of pairs inspected exceeding the knowledge sharing threshold defines the exhaustion criterion in Figure 3. If the agent identifies a discordant pair, then the decision about whether to swap the order within the agent's list depends on the agent's critical thinking on this turn (stochastic draw depending on critical-thinking trait). An uncritical decision will swap the order regardless of whether the order is in fact better. For a critical decision, the agent recognizes the true order and will only swap if the shown list presents the correct relative ordering.

This final decision rule is the "invisible hand" that guides the collaborating agents toward a better outcome. Note that in the current instantiation, the showing agent does not have to make a strong argument (this communication skill is planned for a next iteration). Rather, the other agents will simply know that a competing ordering is better if they use a critical criterion on that turn. An agent who is consistently critical may thus exhaustively determine the competing list to be inferior on all discordances. An uncritical agent will swap two items at the first discordant pair.
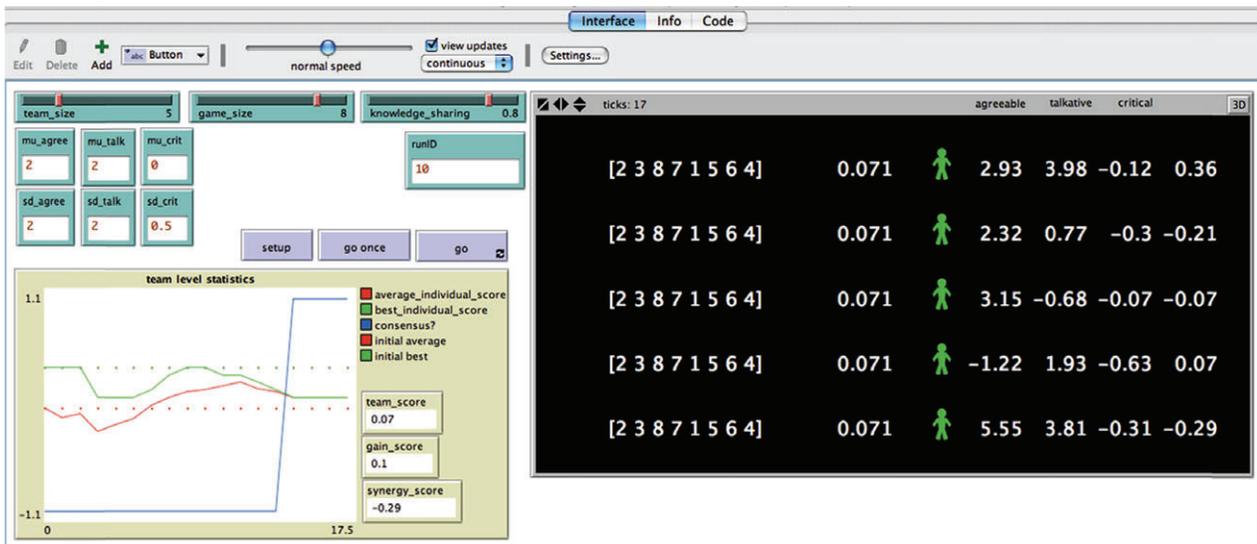
**Figure 4** Screenshot of NetLogo graphical interface at the end of a single run.

## The Simulation Environment

A screenshot of the NetLogo graphical user interface (GUI) for our simulation is shown in Figure 4. Use of the GUI is not required to run the simulation, but the GUI provides some visualization features. Simulation parameters can be entered via the teal sliders and fields in the upper left of the screenshot. These include global parameters such as the team size, game size (length of list), and distribution parameters for the random traits of the agents. With the simulation running in real time, the updating procedure may be visualized in the NetLogo "world" (black region) on the right side of the screenshot. There, the particular agent trait values (fixed once drawn at the beginning of each run) and current rank lists (dynamically changing) are shown.

Some measure of distance between each agent's rank list and the ground-truth list (i.e., a score) is desired in order to track the evolution of rankings over time and compare final scores to initial scores. (The ground-truth list of size $N$ is always 1, 2, … , $N$). In the Subarctic Survival Situation, Szumal (2000) summed over absolute differences between the individual and the expert rank for each item to determine an individual score. This formula has undesirable properties in a simulation study, because the range of the score depends on the number of items in the list, as does its expected value for two random orderings. We used Kendall's $\tau$ rank-correlation coefficient, which is the normalized difference between the number of concordant pairs and the number of discordant pairs in two rank orderings. The value of $\tau$ ranges from 1 (if the rank orderings are the same) to $-1$ (if one rank ordering is the exact reverse of the other), and its expected value for two random rank orderings is 0.

A running plot of the following team-level statistics is shown in the beige region on the bottom left of the screenshot in Figure 4: the average individual score (red line), the best individual score (green line), and an indicator of whether consensus has been reached (blue line). If consensus is reached, then the following quantities are also computed (shown inset in the beige region): team_score = score of consensus list, gain_score = team_score – initial average score, and synergy_score = team_score – initial best individual score. These quantities are analogous to those described in Szumal (2000) for the Subarctic Survival Situation with the difference that we use a rank-correlation coefficient ($\tau$) rather than an item-level sum over absolute differences.

In addition to the GUI output, each run of the simulation produces a process data file containing the input parameters, the agent traits, and a sequence of observable actions. An example process log is excerpted in Figure 5, where agents have been given names from the letters of the Hebrew alphabet (Aleph, Bet, Gimmel, Daled, Hay, … ):

In the excerpt shown in Figure 5, Agents Aleph and Bet are highly talkative ($T = 4.37$ and $5.24$, respectively); hence, in the first few turns, they dominate the observed debate by alternatively showing lists, while the other agents do not show. Agents Bet and Hay are highly agreeable and not very critical, thus both are observed changing their ordering often, even when the result is an inferior list.

```
New Run: 13 with team_size 5, game size 8, and knowledge_sharing 0.5
Agent traits:
Agent Bet A: 3.82 T: 5.24 C: -0.42 K: 0.61 initial list: [5 4 3 6 1 2 7 8]
Agent Daled A: 2.54 T: -1.07 C: 0.06 K: 0.46 initial list: [2 5 7 3 8 6 4 1]
Agent Hay A: 7.13 T: 0.81 C: -0.07 K: 0.36 initial list: [8 4 6 5 1 3 2 7]
Agent Gimmel A: 1.29 T: -0.5 C: -0.36 K: 0.43 initial list: [3 8 4 6 5 2 1 7]
Agent Aleph A: 0.04 T: 4.37 C: 0.83 K: 0.39 initial list: [6 5 1 7 8 4 2 3]

Agent Bet is showing [5 4 3 6 1 2 7 8]
Gimmel changed its ranking from [3 8 4 6 5 2 1 7] (-0.143) to [1 8 4 6 5 2 3 7] (0.071)
Hay changed its ranking from [8 4 6 5 1 3 2 7] (-0.286) to [7 4 6 5 1 3 2 8] (-0.214)
Daled changed its ranking from [2 5 7 3 8 6 4 1] (-0.071) to [2 5 7 8 3 6 4 1] (-0.143)
Agent Aleph is showing [6 5 1 7 8 4 2 3]
Daled changed its ranking from [2 5 7 8 3 6 4 1] (-0.143) to [2 5 7 8 3 6 1 4] (-0.071)
Gimmel changed its ranking from [1 8 4 6 5 2 3 7] (0.071) to [1 5 4 6 8 2 3 7] (0.286)
Bet changed its ranking from [5 4 3 6 1 2 7 8] (0.214) to [1 4 3 6 5 2 7 8] (0.571)
Hay changed its ranking from [7 4 6 5 1 3 2 8] (-0.214) to [8 4 6 5 1 3 2 7] (-0.286)
Agent Aleph is showing [6 5 1 7 8 4 2 3]
Gimmel changed its ranking from [1 5 4 6 8 2 3 7] (0.286) to [1 5 4 6 8 3 2 7] (0.214)
Daled changed its ranking from [2 5 7 8 3 6 1 4] (-0.071) to [2 5 7 8 4 6 1 3] (-0.143)
Hay changed its ranking from [8 4 6 5 1 3 2 7] (-0.286) to [8 4 6 5 1 7 2 3] (-0.357)
Bet changed its ranking from [1 4 3 6 5 2 7 8] (0.571) to [1 4 3 6 7 2 5 8] (0.5)
Agent Bet is showing [1 4 3 6 7 2 5 8]
Hay changed its ranking from [8 4 6 5 1 7 2 3] (-0.357) to [8 4 6 5 1 2 7 3] (-0.286)
Daled changed its ranking from [2 5 7 8 4 6 1 3] (-0.143) to [2 6 7 8 4 5 1 3] (-0.214)
Gimmel changed its ranking from [1 5 4 6 8 3 2 7] (0.214) to [1 5 4 6 8 2 3 7] (0.286)
Agent Hay is showing [8 4 6 5 1 2 7 3]
Bet changed its ranking from [1 4 3 6 7 2 5 8] (0.5) to [1 4 6 3 7 2 5 8] (0.429)

Daled changed its ranking from [2 6 7 8 4 5 1 3] (-0.214) to [2 6 8 7 4 5 1 3] (-0.286)
Gimmel changed its ranking from [1 5 4 6 8 2 3 7] (0.286) to [1 3 4 6 8 2 5 7] (0.5)
...
```

**Figure 5** Excerpt of simulated process log.

## Results From a Sample Experiment

We ran a simulation study using a fixed team size of five agents, a game size of eight, and three values for each of four parameters in a Talkativeness (3) × Agreeableness (3) × Critical_Thinking (3) × Knowledge_Sharing (3) design with 500 replications for each parameter setting (the 40,500 total runs took about 9 hours to complete on a 2.9 GHz Intel Core i5-4570 T CPU with 4 GB RAM using a 32-bit Windows 7 OS). Overall results are shown in Tables 1–3, with these tables split according to the three knowledge sharing conditions that were implemented. Marginal summaries for each of the four variables are shown in Table 4. Outcomes recorded include the percentage of teams that reached consensus, the average number of rounds to consensus (where applicable), and the average (and standard deviation) of the gain and synergy scores. To clarify, a full round is defined as one turn through each of the agents. A turn has two parts: the agent either shows or elects not to show; if the agent does show, then each of the other agents goes through the decision process of Figure 3. Note that the number of rounds, defined this way, is not recoverable from the log files, because it requires knowing the inner state of the agent (i.e., choosing not to show). One design difference between our simulation and the template human task is the existence of a stopping criterion other than consensus. For our simulation, 200 total rounds (i.e., timeout) or three rounds in which no changes occurred (i.e., stall) were stopping criteria.

The parameter values in the tables were selected based on initial exploration with the principal goal of tuning to plausible outcome regions. For example, we wanted to avoid agents moving totally randomly, on one extreme, or lock-step toward perfect orderings, on the other extreme. Three values were used for the population means of each of the three agent-level trait variables: talkativeness, agreeableness and critical_thinking (variances were fixed in each case). The global variable, knowledge_sharing, which governs how exhaustively agents search through shown lists, was also varied.

Results across Tables 1–3 show that as values for knowledge sharing increased (i.e., as agents more exhaustively searched through shown lists for discordant pairs), instances of moderate to high gain and synergy scores also increased. Additionally, across each condition of knowledge sharing, there were higher gain scores when moderate to high talkativeness and agreeableness values were present. Gain scores were less sensitive to critical thinking values. Critical thinking values did, however, seem to influence synergy scores. In particular, low critical thinking values often had low synergy

**Table 1** Overall Summary for Experiment With Varied Agent-Level Trait Values and Knowledge_Sharing = 0.2

| mu_talk (SD = 2) | mu_agree (SD = 2) | mu_crit (SD = 2) | Rounds (SD) | Gain score (SD) | Synergy score (SD) | Consensus % | Stall % | Timeout % |
|---|---|---|---|---|---|---|---|---|
| −2 | −2 | −2 | 33.75 (39.57) | 0 (0) | 0 (0.02) | 0.2 | 98 | 1.8 |
| 0 | −2 | −2 | 79.01 (58.42) | 0.01 (0.06) | −0.01 (0.08) | 4.2 | 87 | 8.8 |
| 2 | −2 | −2 | 109.81 (61.62) | 0.03 (0.13) | −0.01 (0.11) | 12.2 | 68.4 | 19.4 |
| −2 | 0 | −2 | 40.96 (31.63) | 0.01 (0.08) | −0.01 (0.08) | 6.2 | 93 | 0.8 |
| 0 | 0 | −2 | 55.31 (34.03) | 0.07 (0.2) | −0.02 (0.18) | 26.8 | 72.4 | 0.8 |
| 2 | 0 | −2 | 58.05 (36.08) | 0.15 (0.28) | −0.04 (0.27) | 56.4 | 42.4 | 1.2 |
| −2 | 2 | −2 | 30.14 (16.39) | 0.04 (0.15) | −0.02 (0.16) | 19.6 | 80.4 | 0 |
| 0 | 2 | −2 | 33.25 (18.43) | 0.18 (0.29) | −0.03 (0.27) | 61 | 38.6 | 0.4 |
| 2 | 2 | −2 | 27.28 (12.1) | 0.33 (0.33) | 0.03 (0.32) | 89 | 11 | 0 |
| −2 | −2 | 0 | 27.48 (31.82) | 0 (0) | 0 (0) | 0 | 99.2 | 0.8 |
| 0 | −2 | 0 | 66.7 (55.86) | 0.01 (0.1) | 0.01 (0.06) | 1.8 | 91.8 | 6.4 |
| 2 | −2 | 0 | 93.99 (60.03) | 0.02 (0.12) | 0.01 (0.08) | 4 | 82.4 | 13.6 |
| −2 | 0 | 0 | 38.68 (31.76) | 0.01 (0.05) | 0 (0.03) | 1.6 | 97.4 | 1 |
| 0 | 0 | 0 | 58.88 (41.55) | 0.09 (0.24) | 0.04 (0.15) | 17 | 80.6 | 2.4 |
| 2 | 0 | 0 | 64 (45.53) | 0.24 (0.37) | 0.13 (0.26) | 35.8 | 60.8 | 3.4 |
| −2 | 2 | 0 | 31.41 (19.65) | 0.06 (0.2) | 0.03 (0.13) | 11.2 | 88.8 | 0 |
| 0 | 2 | 0 | 35.77 (23.69) | 0.31 (0.38) | 0.15 (0.26) | 47.4 | 52.4 | 0.2 |
| 2 | 2 | 0 | 29 (19) | 0.54 (0.38) | 0.29 (0.3) | 76.8 | 22.6 | 0.6 |
| −2 | −2 | 2 | 21.29 (26.74) | 0 (0) | 0 (0) | 0 | 99 | 1 |
| 0 | −2 | 2 | 45.02 (40.65) | 0.01 (0.07) | 0 (0.05) | 0.6 | 97.2 | 2.2 |
| 2 | −2 | 2 | 62.96 (50.69) | 0.01 (0.08) | 0 (0.04) | 1 | 93 | 6 |
| −2 | 0 | 2 | 28.96 (23.34) | 0.01 (0.07) | 0 (0.05) | 0.6 | 98.8 | 0.6 |
| 0 | 0 | 2 | 46.69 (31.69) | 0.1 (0.28) | 0.06 (0.18) | 12 | 87 | 1 |
| 2 | 0 | 2 | 47.64 (35.06) | 0.23 (0.41) | 0.15 (0.28) | 25.4 | 73 | 1.6 |
| −2 | 2 | 2 | 28.24 (20.5) | 0.06 (0.23) | 0.04 (0.14) | 7.4 | 92.4 | 0.2 |
| 0 | 2 | 2 | 30.51 (17.56) | 0.36 (0.45) | 0.23 (0.31) | 41.6 | 58 | 0.4 |
| 2 | 2 | 2 | 25.06 (17.36) | 0.63 (0.44) | 0.40 (0.32) | 69 | 30.8 | 0.2 |

*Note.* mu_talk = talkativeness; mu_agree = agreeableness; mu_crit = critical_thinking.

scores, even when the values for other agent-level trait variables were high. Unsurprisingly, low talkativeness and low agreeableness values resulted in low consensus rates.

Note that while gain and synergy scores can be quite different at the team level, on aggregate these measures do not provide complementary information. The correlation between the average gain and average synergy for the 81 different parameter settings in Tables 1–3 is 0.81. It is too early to say whether this observation is a consequence of information loss, due to averaging of nonlinear performance effects or whether it is an indicator that the only difference between weak and strong synergy in our simulation is a matter of scale.

We look at Table 4 for a sensitivity analysis of outcomes to each of the four parameters that were varied. Similar to what was noted from the Tables 1–3, higher values for agreeableness and talkativeness increased gain scores and contributed to modest increases in synergy scores. This finding is consistent with prior work (Barrick et al., 1998; Bell, 2007; LePine et al., 2008). High values for these traits also contributed to higher consensus rates. Agreeableness and talkativeness differentially impacted the number of rounds, however. In particular, higher values for agreeableness decreased the duration of the task whereas higher values for talkativeness increased the duration of the task. Higher values for critical thinking showed modest increases in gain and synergy scores, as with prior work (Barrick et al., 1998; Devine & Philips, 2001; Stewart, 2006), and critical thinking was the only variable to reduce consensus rates, a result that perhaps deserves more attention. Increased knowledge_sharing also contributed to modest increases in gain and synergy scores, as discussed from the tables above, but knowledge_sharing had little impact on consensus rates and the duration of the task.

## Future Work

The results described above are preliminary, but they represent a promising foray into the use of ABM for psychometric considerations of CPS. There are still features of the simulation and experimental designs that we intend to pursue in follow-up work. We would like to add new traits such as communication skill and consensus orientation. For example, the

**Table 2** Overall Summary for Experiment With Varied Agent-Level Trait Values and Knowledge_Sharing = 0.5

| mu_talk (SD = 2) | mu_agree (SD = 2) | mu_crit (SD = 2) | Rounds (SD) | Gain score (SD) | Synergy score (SD) | Consensus % | Stall % | Timeout % |
|---|---|---|---|---|---|---|---|---|
| −2 | −2 | −2 | 35.64 (41.83) | 0 (0.03) | 0 (0.04) | 0.4 | 97.6 | 2 |
| 0 | −2 | −2 | 89.07 (62.55) | 0.04 (0.15) | 0 (0.12) | 14.8 | 73.4 | 11.8 |
| 2 | −2 | −2 | 104.74 (62.37) | 0.1 (0.23) | −0.02 (0.21) | 34.6 | 48.2 | 17.2 |
| −2 | 0 | −2 | 36 (28.62) | 0.02 (0.13) | −0.02 (0.14) | 14.4 | 85 | 0.6 |
| 0 | 0 | −2 | 51.8 (36.19) | 0.15 (0.28) | −0.04 (0.26) | 56 | 43 | 1 |
| 2 | 0 | −2 | 47.05 (36.44) | 0.3 (0.33) | 0.01 (0.34) | 83.4 | 15.4 | 1.2 |
| −2 | 2 | −2 | 26.11 (14.14) | 0.12 (0.24) | −0.01 (0.22) | 39.4 | 60.6 | 0 |
| 0 | 2 | −2 | 23.98 (13.44) | 0.27 (0.33) | −0.02 (0.35) | 86.8 | 13.2 | 0 |
| 2 | 2 | −2 | 20.43 (14.13) | 0.37 (0.34) | 0.05 (0.36) | 95.6 | 4.2 | 0.2 |
| −2 | −2 | 0 | 32.16 (37.67) | 0.01 (0.06) | 0 (0.02) | 1 | 97.6 | 1.4 |
| 0 | −2 | 0 | 76.46 (54.76) | 0.04 (0.18) | 0.01 (0.13) | 8.2 | 84.2 | 7.6 |
| 2 | −2 | 0 | 102.91 (64.18) | 0.1 (0.25) | 0.04 (0.17) | 16.6 | 64.2 | 19.2 |
| −2 | 0 | 0 | 37.01 (31.93) | 0.04 (0.16) | 0.01 (0.11) | 8.6 | 90.2 | 1.2 |
| 0 | 0 | 0 | 52.87 (38.34) | 0.29 (0.38) | 0.14 (0.27) | 46 | 52.2 | 1.8 |
| 2 | 0 | 0 | 50.66 (42.92) | 0.5 (0.41) | 0.27 (0.31) | 69.2 | 28.2 | 2.6 |
| −2 | 2 | 0 | 26.87 (15.97) | 0.14 (0.3) | 0.05 (0.22) | 28.2 | 71.8 | 0 |
| 0 | 2 | 0 | 26.45 (16.41) | 0.46 (0.39) | 0.23 (0.32) | 71.4 | 28.6 | 0 |
| 2 | 2 | 0 | 20.39 (14.15) | 0.70 (0.32) | 0.38 (0.31) | 93.4 | 6.4 | 0.2 |
| −2 | −2 | 2 | 25.52 (26.01) | 0 (0.01) | 0 (0) | 0.2 | 99.8 | 0 |
| 0 | −2 | 2 | 49.44 (43.61) | 0.02 (0.13) | 0.01 (0.09) | 2.6 | 95 | 2.4 |
| 2 | −2 | 2 | 62.53 (48.16) | 0.07 (0.24) | 0.04 (0.16) | 7.4 | 87.8 | 4.8 |
| −2 | 0 | 2 | 31.88 (25.87) | 0.04 (0.17) | 0.02 (0.11) | 5.2 | 94.4 | 0.4 |
| 0 | 0 | 2 | 43.11 (34.44) | 0.24 (0.41) | 0.15 (0.28) | 26 | 72 | 2 |
| 2 | 0 | 2 | 43.49 (36.45) | 0.48 (0.47) | 0.30 (0.32) | 53.2 | 45.2 | 1.6 |
| −2 | 2 | 2 | 24.85 (14.33) | 0.17 (0.36) | 0.1 (0.23) | 20 | 80 | 0 |
| 0 | 2 | 2 | 25.88 (19.42) | 0.57 (0.46) | 0.36 (0.33) | 63.6 | 36.2 | 0.2 |
| 2 | 2 | 2 | 19.35 (17.84) | 0.8 (0.35) | 0.51 (0.28) | 86.4 | 13.4 | 0.2 |

*Note.* mu_talk = talkativeness; mu_agree = agreeableness; mu_crit = critical_thinking.

decision process in Figure 3 could include a conjunctive or disjunctive relationship between critical thinking on the part of the listening agent and communication skill on the part of the showing agent. If this is a between-agent trait interaction, within-agent trait interactions might also be added. For example, high talkativeness combined with low agreeableness can translate into not only passively choosing not to update, but also into actively interrupting a debate round. So far, traits in our agents have been completely uncorrelated random variables. For increased realism, we could use multivariate distributions for domain ability, critical thinking, and communication skill, for example.

In the current ABM, the traits have many times been conceptualized as too black and white or good and bad (e.g., high agreeability vs. low agreeability or high critical thinking vs. low critical thinking). Results of the sample experiment described here show that this conceptualization may not have allowed for the kind of variation that could lead to interesting heterogeneity. Therefore, in a new version of the ABM, we are also incorporating trait scales that differ along more qualitative dimensions. As one example, we are including consensus orientation as an agent-level trait. This trait corresponds with an agent's desire or interest in seeking information from others about their solutions as opposed to an agent's interest in informing others of the agent's own solution. As a result, agents with a high consensus orientation should be more willing to solicit information from others while agents with a low consensus orientation are more interested in presenting their own answers to the group.

The sample experiment also involved groups of agents sampled from unimodal population distributions, but the team optimization problem is one of deliberate design. Team optimization may be explored via conditional distributions from the full population. An alternative experiment with respect to team optimization would be to discretize traits and combine archetypal team members in different combinations.

Finally, an important direction of future work is to consider the information value of output logs from our simulation as data for process models. We hope to be able to inform the simulation design by practical considerations for process data analysis. Furthermore, it will not be clear whether we have achieved the nonlinear, complex, emergent dynamic that is the promise of ABM unless we can show that indeed the outcomes transcend description by summary statistics.

**Table 3** Overall Summary for Experiment With Varied Agent-Level Trait Values and Knowledge_Sharing = 0.8

| mu_talk (SD = 2) | mu_agree (SD = 2) | mu_crit (SD = 2) | Rounds (SD) | Gain score (SD) | Synergy score (SD) | Consensus % | Stall % | Timeout % |
|---|---|---|---|---|---|---|---|---|
| −2 | −2 | −2 | 36.65 (41.48) | 0 (0.06) | −0.01 (0.06) | 2.8 | 95.8 | 1.4 |
| 0 | −2 | −2 | 86.85 (58.41) | 0.06 (0.17) | 0 (0.12) | 19.6 | 68.4 | 12 |
| 2 | −2 | −2 | 110.46 (63.84) | 0.14 (0.27) | −0.01 (0.24) | 44 | 34.4 | 21.6 |
| −2 | 0 | −2 | 36.27 (27.53) | 0.04 (0.16) | −0.02 (0.15) | 20 | 79.8 | 0.2 |
| 0 | 0 | −2 | 47.26 (37.29) | 0.19 (0.31) | −0.03 (0.29) | 63.6 | 34.2 | 2.2 |
| 2 | 0 | −2 | 42.46 (34.21) | 0.31 (0.34) | 0.01 (0.36) | 89.2 | 9.8 | 1 |
| −2 | 2 | −2 | 23.97 (14.51) | 0.11 (0.25) | −0.04 (0.23) | 47.6 | 52.4 | 0 |
| 0 | 2 | −2 | 21.52 (10.17) | 0.3 (0.32) | 0.01 (0.34) | 89.4 | 10.6 | 0 |
| 2 | 2 | −2 | 17.84 (9.71) | 0.39 (0.34) | 0.07 (0.36) | 97.4 | 2.6 | 0 |
| −2 | −2 | 0 | 33.09 (40.38) | 0 (0.07) | 0 (0.05) | 1.2 | 97.4 | 1.4 |
| 0 | −2 | 0 | 73.95 (57.52) | 0.07 (0.23) | 0.03 (0.15) | 13 | 79.2 | 7.8 |
| 2 | −2 | 0 | 99.58 (61.53) | 0.17 (0.32) | 0.08 (0.21) | 25.2 | 59.4 | 15.4 |
| −2 | 0 | 0 | 38.27 (30.9) | 0.06 (0.22) | 0.02 (0.17) | 14.8 | 84.4 | 0.8 |
| 0 | 0 | 0 | 52.72 (39.52) | 0.33 (0.4) | 0.17 (0.29) | 49.6 | 48.6 | 1.8 |
| 2 | 0 | 0 | 48.43 (42.21) | 0.58 (0.38) | 0.31 (0.31) | 80.2 | 17 | 2.8 |
| −2 | 2 | 0 | 24.94 (14.01) | 0.19 (0.32) | 0.07 (0.23) | 35 | 65 | 0 |
| 0 | 2 | 0 | 23.46 (13.82) | 0.54 (0.37) | 0.27 (0.32) | 81.4 | 18.6 | 0 |
| 2 | 2 | 0 | 18.01 (15.19) | 0.73 (0.28) | 0.40 (0.28) | 96.2 | 3.6 | 0.2 |
| −2 | −2 | 2 | 24.53 (27.39) | 0 (0.06) | 0 (0.03) | 0.6 | 99 | 0.4 |
| 0 | −2 | 2 | 51.93 (45.74) | 0.05 (0.2) | 0.03 (0.13) | 5.6 | 91.2 | 3.2 |
| 2 | −2 | 2 | 64.93 (51.67) | 0.1 (0.28) | 0.06 (0.18) | 11 | 83.4 | 5.6 |
| −2 | 0 | 2 | 28.21 (19.85) | 0.07 (0.23) | 0.04 (0.16) | 8.4 | 91.6 | 0 |
| 0 | 0 | 2 | 41.4 (33.28) | 0.33 (0.44) | 0.19 (0.28) | 37.6 | 60.8 | 1.6 |
| 2 | 0 | 2 | 36.57 (32.67) | 0.55 (0.47) | 0.35 (0.34) | 59.8 | 38.8 | 1.4 |
| −2 | 2 | 2 | 23.33 (16.47) | 0.22 (0.38) | 0.13 (0.25) | 27.6 | 72.4 | 0 |
| 0 | 2 | 2 | 21.91 (13.82) | 0.65 (0.42) | 0.40 (0.3) | 73.4 | 26.6 | 0 |
| 2 | 2 | 2 | 16.88 (12.23) | 0.86 (0.29) | 0.54 (0.26) | 92 | 8 | 0 |

*Note.* mu_talk = talkativeness; mu_agree = agreeableness; mu_crit = critical_thinking.

**Table 4** Marginalized Mean Outcomes for the Four Varied Parameters

| Parameter | Value | Rounds (SD) | Gain score (SD) | Synergy score (SD) | Consensus % | Stall % | Timeout % |
|---|---|---|---|---|---|---|---|
| *mu_Crit* | **−2** | 49.1 (47.33) | 0.14 (0.27) | −0.01 (0.24) | 44 | 53 | 4 |
|  | **0** | 47.56 (45.84) | 0.23 (0.36) | 0.12 (0.25) | 35 | 62 | 3 |
|  | **2** | 36.00 (34.19) | 0.25 (0.41) | 0.15 (0.27) | 27 | 71 | 1 |
| *mu_Talk* | **−2** | 30.60 (28.26) | 0.05 (0.2) | 0.01 (0.14) | 12 | 88 | 1 |
|  | **0** | 48.56 (43.32) | 0.21 (0.36) | 0.09 (0.27) | 38 | 59 | 3 |
|  | **2** | 53.50 (51.46) | 0.35 (0.41) | 0.16 (0.32) | 56 | 39 | 5 |
| *mu_Agree* | **−2** | 62.98 (57.84) | 0.04 (0.17) | 0.01 (0.12) | 9 | 84 | 7 |
|  | **0** | 44.62 (35.66) | 0.20 (0.35) | 0.08 (0.27) | 36 | 63 | 1 |
|  | **2** | 25.07 (16.68) | 0.38 (0.41) | 0.17 (0.33) | 61 | 39 | 0.1 |
| *Knowledge sharing* | **0.2** | 46.29 (42.28) | 0.13 (0.3) | 0.05 (0.21) | 23 | 75 | 3 |
|  | **0.5** | 43.95 (43.66) | 0.22 (0.37) | 0.10 (0.28) | 38 | 59 | 3 |
|  | **0.8** | 42.42 (43.73) | 0.26 (0.39) | 0.11 (0.29) | 44 | 53 | 3 |

*Note.* mu_Talk = talkativeness; mu_Agree = agreeableness; mu_Crit = critical_thinking.

## References

Andersson, J., & Rönnberg, J. (1995). Recall suffers from collaboration: Joint recall effects of friendship and task complexity. *Applied Cognitive Psychology, 9*(3), 199–211.

Andrews, J. J., & Rapp, D. N. (2015). Benefits, costs, and challenges of collaboration for learning and memory. *Translational Issues in Psychological Science, 1*(2), 182–191.

Barber, S. J., Rajaram, S., & Aron, A. (2010). When two is too many: Collaborative encoding impairs memory. *Memory & Cognition, 38*(3), 255–264.

Barrick, M. R., Stewart, G. L., Neubert, M. J., & Mount, M. K. (1998). Relating member ability and personality to work-team processes and team effectiveness. *Journal of Applied Psychology, 83*(3), 377–391.

Basden, B. H., Basden, D. R., & Henry, S. (2000). Costs and benefits of collaborative remembering. *Applied Cognitive Psychology, 14*(6), 497–507.

Beal, D. J., Cohen, R. R., Burke, M. J., & McLendon, C. L. (2003). Cohesion and performance in groups: A meta-analytic clarification of construct relations. *Journal of Applied Psychology, 88*(6), 989–1004.

Bell, S. T. (2007). Deep-level composition variables as predictors of team performance: A meta-analysis. *Journal of Applied Psychology, 92*(3), 595–615.

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K–12 assessment: Integrating accountability testing, formative assessment and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43–61). New York, NY: Springer.

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences of the United States of America, 99*(suppl 3), 7280–7287.

Bowen, C. W. (2000). A quantitative literature review of cooperative learning effects on high school and college chemistry achievement. *Journal of Chemical Education, 77*(1), 116–119.

Brannick, M. T., Prince, A., Prince, C., & Salas, E. (1995). The measurement of team process. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*(3), 641–651.

Burrus, J., Jackson, T., Xi, N., & Steinberg, J. (2013). *Identifying the most important 21st century workforce competencies: An analysis of the occupational Information network (O* NET)* (Research Report No. RR-13-21). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2013.tb02328.x

Credé, M., & Sniezek, J. A. (2003). Group judgment processes and outcomes in video-conferencing versus face-to-face groups. *International Journal of Human-Computer Studies, 59*(6), 875–897.

Crowder, R. M., Robinson, M., Hughes, H. P., & Sim, Y.-W. (2012). The development of an agent-based modeling framework for simulating engineering team work. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 42*(6), 1425–1439.

Devine, D. J., & Philips, J. L. (2001). Do smarter teams do better: A meta-analysis of cognitive ability and team performance. *Small Group Research, 32*(5), 507–532.

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.

Dionne, S. D., Sayama, H., Hao, C., & Bush, B. J. (2010). The role of leadership in shared mental model convergence and team performance improvement: An agent-based computational model. *The Leadership Quarterly, 21*(6), 1035–1049.

Fan, X., & Yen, J. (2011). Modeling cognitive loads for evolving shared mental models in human–agent collaboration. *Systems, Man, and Cybernetics, Part B: Cybernetics, 41*(2), 354–367.

Fawcett, L. M., & Garton, A. F. (2005). The effect of peer collaboration on children's problem-solving ability. *British Journal of Educational Psychology, 75*(2), 157–169.

Fiore, S. M., & Salas, E. (Eds.). (2004). *Team cognition: Understanding the factors that drive process and performance*. Washington, DC: American Psychological Association.

Gaudet, A. D., Ramer, L. M., Nakonechny, J., Cragg, J. J., & Ramer, M. S. (2010). Small-group learning in an upper-level university biology class enhances academic performance and student attitudes toward group work. *PLoS ONE, 5*(12), 1–10.

Giachetti, R. E., Marcelli, V., Cifuentes, J., & Rojas, J. A. (2013). An agent-based simulation model of human-robot team performance in military environments. *Systems Engineering, 16*(1), 15–28.

Gilbert, N. (2007). *Agent-based models*. Thousand Oaks, CA: Sage.

Gillies, R. M. (2004). The effects of cooperative learning on junior high school students during small group learning. *Learning and Instruction, 14*(2), 197–213.

González, M. G., Burke, M. J., Santuzzi, A. M., & Bradley, J. C. (2003). The impact of group process variables on the effectiveness of distance collaboration groups. *Computers in Human Behavior, 19*(5), 629–648.

Goodyear, P., Jones, C., & Thompson, K. (2014). Computer-supported collaborative learning: Instructional approaches, group processes and educational designs. In J. M. Spector, M. D. Merrill, J. Elen, M. J. Bishop (Eds.), *Handbook of research on educational communications and technology*, (pp. 439–451). New York, NY: Springer.

Hackman, J. R. (1987). The design of work teams. In J. W. Lorsch (Ed.), *Handbook of organizational behavior* (pp. 315–342). Englewood Cliffs, NJ: Prentice-Hall.

Hill. (1982) Group versus individual performance: Are N+1 heads better than one? *Psychological Bulletin, 91(3),* 517–539.

Holland, J. H. (1975). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence.* Ann Arbor: University of Michigan Press.

Holland, J. H., & Miller, J. H. (1991). Artificial adaptive agents in economic theory. *The American Economic Review, 81*(2), 365–370.

Horii, T., Jin, Y., & Levitt, R. E. (2005). Modeling and analyzing cultural influences on project team performance. *Computational & Mathematical Organization Theory, 10*(4), 305–321.

Human Synergistics International. (2012). *Desert Survival Situation.* Retrieved from http://www.humansynergistics.com/Products/ TeamBuildingSimulations/SurvivalSeries/DesertSurvivalSituation

Ilgen, D. R., Hollenbeck, J. R., Johnson, M., & Jundt, D. (2005). Teams in organizations: From input-process-output models to IMOI models. *Annual Review of Psychology, 56,* 517–543.

Jehn, K. A., Northcraft, G. B., & Neale, M. A. (1999). Why differences make a difference: A field study of diversity, conflict and performance in workgroups. *Administrative Science Quarterly, 44*(4), 741–763.

Jeong, H., & Chi, M. T. H. (2007). Knowledge convergence and collaborative learning. *Instructional Science, 35*(4), 287–315.

Johnson, D. W., Johnson, R. T., & Stanne, M. B. (2000). *Cooperative learning methods: A meta-analysis.* Minneapolis: University of Minnesota Press.

Johnson, D. W., Maruyama, G., Johnson, R., Nelson, D., & Skon, L. (1981). Effects of cooperative, competitive, and individualistic goal structures on achievement: A meta-analysis. *Psychological Bulletin, 89*(1), 47–62.

Johnson, S. (2002). *Emergence: The connected lives of ants, brains, cities, and software.* New York, NY: Simon and Schuster.

Kraus, S., Sycara, K., & Evenchik, A. (1998). Reaching agreements through argumentation: A logical model and implementation. *Artificial Intelligence, 104*(1–2), 1–69.

Larson, J. R. (2007). Deep diversity and strong synergy modeling the impact of variability in members' problem-solving strategies on group problem-solving performance. *Small Group Research, 38*(3), 413–436.

Larson, J. R. (2010). *In search of synergy in small group performance.* New York, NY: Psychology Press.

Laughlin, P. R., Bonner, B. L., & Miner, A. G. (2002). Groups perform better than the best individuals on letters-to-numbers problems. *Organizational Behavior and Human Decision Processes, 88*(2), 605–620.

Laughlin, P. R., Gonzalez, C. M., & Sommer, D. (2003). Quantity estimations by groups and individuals: Effects of known domain boundaries. *Group Dynamics: Theory, Research, and Practice, 7*(1), 55–63.

Laughlin, P. R., Hatch, E. C., Silver, J. S., & Boh, L. (2006). Groups perform better than the best individuals on letters-to-numbers problems: Effects of group size. *Journal of Personality and Social Psychology, 90*(4), 644–651.

LePine, J. A., Piccolo, R. F., Jackson, C. L., Mathieu, J. E., & Saul, J. R. (2008). A meta-analysis of teamwork processes: Tests of a multi-dimensional model and relationships with team effectiveness criteria. *Personnel Psychology, 61*(2), 273–307.

Liu, L., Hao, J., von Davier, A. A., Kyllonen, P., & Zapata-Rivera, J.-D. (2015). A tough nut to crack: Measuring collaborative problem solving. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on computational tools for real-world skill development* (pp. 344–359). Hershey, PA: IGI-Global.

Macy, M. W., & Willer, R. (2002). From factors to actors: Computational sociology and agent-based modeling. *Annual Review of Sociology, 28,* 143–166.

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *Journal of Applied Psychology, 85*(2), 273–283.

McGrath, J. E. (1964). *Social psychology: A brief introduction.* New York, NY: Holt, Rinehart and Winston.

Meslec, N., & Curşeu, P. L. (2013). Too close or too far hurts: Cognitive distance and group cognitive synergy. *Small Group Research, 44*(5), 471–497.

Mesmer-Magnus, J. R., & DeChurch, L. A. (2009). Information sharing and team performance: A meta-analysis. *Journal of Applied Psychology, 94*(2), 535–546.

Mohammed, S., & Angell, L. C. (2003). Personality heterogeneity in teams which differences make a difference for team performance? *Small Group Research, 34*(6), 651–677.

Morgan, B. B., Jr., Salas, E., & Glickman, A. S. (1993). An analysis of team evolution and maturation. *Journal of General Psychology, 120*(3), 277–291.

Mullen, B., & Copper, C. (1994). The relation between group cohesiveness and performance: An integration. *Psychological Bulletin, 115*(2), 210–227.

National Research Council. (2011). *Assessing 21st century skills.* Washington, DC: National Academies Press.

Neuman, G. A., Wagner, S. H., & Christiansen, N. D. (1999). The relationship between work-team personality composition and the job performance of teams. *Group & Organization Management, 24*(1), 28–45.

O'Neil, H. F., Chung, G. K. W. K., & Brown, R. S. (1995). *Measurement of teamwork processes using computer simulation* (CSE Technical Report No. 399). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

O'Neil, H. F., Chung, G. K. W. K., & Herl, H. E. (1999). *Computer-based collaborative knowledge mapping to measure team processes and team outcomes.* (CSE Technical Report No. 502). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Teaching.

Organisation for Economic Co-operation and Development. (2013). *PISA 2015 draft collaborative problem solving framework*. Retrieved from https://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf

Reagan-Cirincione, P. (1994). Improving the accuracy of group judgment: A process intervention combining group facilitation, social judgment analysis, and information technology. *Organizational Behavior and Human Decision Processes, 58*(2), 246–270.

Slavin, R. E. (1983). When does cooperative learning increase student achievement? *Psychological Bulletin, 94*(3), 429–445.

Sniezek, J. A. (1989). An examination of group process in judgmental forecasting. *International Journal of Forecasting, 5*(2), 171–178.

Stasser, G., & Stewart, D. (1992). Discovery of hidden profiles by decision-making groups: Solving a problem versus making a judgment. *Journal of Personality and Social Psychology, 63*(3), 426–434.

Stewart, G. L. (2006). A meta-analytic review of relationships between team design features and team performance. *Journal of Management, 32*(1), 29–55.

Subarctic Survival Situation™. (2013). Retrieved from http://www.subarcticsurvival.com/

Szumal, J. L. (2000). How to use group problem solving simulations to improve teamwork. In M. Silberman & P. Philips (Eds.), *Team and organization development sourcebook*. New York, NY: McGraw-Hill.

Tindale, R. S., & Larson, J. R. (1992). Assembly bonus effect or typical group performance? A comment on Michaelsen, Watson, and Black (1989). *Journal of Applied Psychology, 77*(1), 102–105.

Tindale, R. S., & Sheffey, S. (2002). Shared information, cognitive load, and group memory. *Group Processes & Intergroup Relations, 5*(1), 5–18.

von Davier, A. A., & Halpin, P. F. (2013). *Collaborative problem solving and the assessment of cognitive skills: Psychometric considerations* (Research Report No. RR-13-41). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2013.tb02348.x

Wilensky, U. (1999). *NetLogo*. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL. Retrieved from http://ccl.northwestern.edu/netlogo/

Zaccaro, S. J., & McCoy, M. C. (1988). The effects of task and interpersonal cohesiveness on performance of a disjunctive group task. *Journal of Applied Social Psychology, 18*(10), 837–851.

Zhang, X., Li, Y., Zhang, S., & Schlick, C. M. (2013). Modelling and simulation of the task scheduling behavior in collaborative product development process. *Integrated Computer-Aided Engineering, 20*(1), 31–44.

Zhang, X., Luo, L., Yang, Y., Li, Y., Schlick, C. M., & Grandt, M. (2009). A simulation approach for evaluation and improvement of organisational planning in collaborative product development projects. *International Journal of Production Research, 47*(13), 3471–3501.

Zhu, M., Huang, Y., & Contractor, N. S. (2013). Motivations for self-assembling into project teams. *Social Networks*, *35*(2), 251–264.