3-2013

Reciprocal trust mediates deep transfer of learning between games of strategic interaction

Ion Juvina
Wright State University

Muniba Saleem
University of Michigan - Dearborn

Jolie M. Martin
University of Minnesota - Twin Cities

Cleotilde Gonzalez
Carnegie Mellon University, conzalez@andrew.cmu.edu

Christian Lebiere
Carnegie Mellon University, cl@cmu.edu

Published In
Organizational Behavior and Human Decision Processes, 120, 2, 206-215.

Reciprocal trust mediates deep transfer of learning between games of strategic interaction

Abstract

We studied transfer of learning across two games of strategic interaction. We found that the interpersonal relation between two players during and across two games influence development of reciprocal trust and transfer of learning from one game to another. We show that two types of similarities between the games affect transfer: (1) deep similarities facilitate transfer of an optimal solution across games; (2) surface similarities can either facilitate or hinder transfer depending on whether they lead players toward an optimal or sub-optimal solution in the target game. Learning an optimal solution in a context of interdependence between players is associated with development of reciprocal trust, which in turn mediates transfer of learning across games. The results can be used to inform the design of training exercises to develop strategic interaction skills.

Keywords: reciprocal trust, transfer of learning, games of strategic interaction, cooperation, surface and deep similarities

INTRODUCTION

Many organizations employ trust exercises in order to increase cooperation within teams, workgroups, departments, and firms. Training may include taking groups of employees on "Outward Bound"-type outdoor adventures or activities that require teamwork, such as an exercise where one individual must fall backwards and trust a team member to catch him. An implicit assumption in these exercises is that the development of reciprocal trust and cooperation among team members in a very specific, ad-hoc situation will transfer to organizational life once the training is over. Is this assumption sufficiently grounded in empirical research? Do individuals and organizations achieve successful transfer?

High-profile examples of failed transfer are common. For instance, in the middle of the 2010 healthcare debate, the U.S. President and the House Speaker took a break and played golf for four hours before returning to the negotiation table. It was everyone's hope that playing golf would help the two leaders build a relationship and a sense of camaraderie that would transfer to their negotiation situation and would ultimately increase the chances of bipartisan cooperation in Congress. This was a failed attempt. While there are many potential reasons why, it might also have something to do with the lack of deep similarities between the two situations and the incentive structure of the golf game that does not promote development of reciprocal trust. Golf is a zero-sum game; it promotes a "win-lose" mindset that is not compatible with the desired bipartisanship in Congress (i.e., a "win-win" solution). If anything, a negative transfer is to be expected in such cases, which is in line with the observed outcome in the healthcare debate.

Examples of negatively transferring learning across games also come from empirical research. Chess masters obtain lower scores than average college students at Centipede[1], presumably because they apply the backward reasoning strategy, which is very effective in Chess (a win-lose game) but suboptimal in Centipede (a win-win game) (Palacios-Huerta & Volij, 2009). Conversely, playing Centipede before Race-to-100[2] decreases performance in the second game, because the optimal strategy in Centipede (a win-win game) is suboptimal in Race-to-100 (a win-lose game) (Levitt, List, & Sadoff, 2011).

In the study reported here, we investigated the conditions for the development of reciprocal trust and successful positive transfer of learning in games of strategic interaction as analytically and experimentally tractable reproductions of real world interactions. In the following section (background), we discuss previous research on the transfer of learning, and highlight the importance of similarities between games and the development of reciprocal trust between players. In a subsequent section, we provide specific hypotheses on how relations between games and between players influence reciprocal trust development and the transfer of learning from one game to another. The details of a laboratory study are presented in the Methods and Results sections. The contributions of this paper to the theoretical literature and its relevance to applied settings are discussed in the final section.

[1] The Centipede game is played between two players who make their moves sequentially. Each player chooses either to take the accumulated pot or pass to the other player. The player who passes takes the risk to make slightly fewer points than the player who takes the pot. However, as long as both players pass, the pot increases providing better payoffs for both players in the long run.

[2] Race-to-100 is also played between two players who alternate choosing numbers within a given range (e.g., 1 to 10). The player who chooses a number that makes the sum of all chosen numbers equal to exactly 100 is the winner.

BACKGROUND

Transfer of Learning Across Games of Strategic Interaction

Researchers often use matrix games—abstract representations of interactions that include a set of players, the options available to each player, and the payoffs that are associated with their actions—to study strategic interaction (e.g., Camerer, 2003; Plott & Smith, 2008). When games are played in sequence, studies find that prior experience in one game can affect behavior in a subsequent game (e.g., Ahn, Ostrom, Schmidt, Shupp, & Walker, 2001; Knez & Camerer, 2000; Schotter, 1998; Van Huyck, Battalio, & Beil, 1991). This transfer of behavior from one game to another has often been referred to as "behavioral spillover" in the Behavioral Economics literature suggesting that choices and behaviors from one game can "spill over" to the subsequent game (Bednar, Chen, Xiao Liu, & Page, 2012).

There is limited research on what factors cause these behavioral transfers or spillovers to occur across games of strategic interaction. Bednar and colleagues (2012) use the concept of *entropy* or strategic uncertainty to explain when behavior is likely to spillover from one game to another. Studies suggest that prevalent strategies in games with low entropy are more likely to be transferred to games with high entropy, but not vice versa (Bednar et al., 2012). In other words, individuals develop strategies for easier games and apply them to more complex games.

Knez and Camerer (2000) studied the transfer of precedent from the Weak Link (WL) to seven-action and three-action Prisoner's Dilemma (PD) games and found that precedents of cooperation in the WL game correlated with higher amounts of cooperation

in the seven-action, but not the three-action PD game. In other words, behavioral transfer across games strongly depended on the presence of superficial, surface similarity (what they call 'descriptive' similarity) between the two games such as having the same number of actions in both games. When the games differed in surface characteristics (e.g., actions were numbered differently in the two games), the transfer of behavior from one game to another did not occur. This finding suggests that in order to understand how people generalize behavior from one strategic interaction to another, it is important to consider if and how similar the two strategic interaction situations are.

Games used in this study

Tables 1 and 2 present the payoff structure of the two games used in this study.

Insert Table 1 and Table 2 around here.

Prisoner's Dilemma (PD) Game: In the Prisoner's Dilemma game, each party is better off competing, regardless of what the other party does. At the same time, both parties are better off when they mutually cooperate (the [1,1] cell in Table 1) than they would be if they mutually competed (Rapoport & Chammah, 1965). Parties can compete to maximize their outcomes – solo competition provides better outcomes for the competitor than mutual cooperation does. However, when PD is repeated indefinitely, competition does not maximize joint gain because it triggers retaliation, which results in perpetual and mutual competition. Thus, in the long-term, a mutually cooperative maximizes both players' payoffs.

Chicken Game (CG): The game of chicken models risky strategic interactions in

which mutual cooperation (the [1,1] cell in Table 2) is unstable as each party has an incentive to unilaterally defect (Halevy, Chou, & Murnighan, 2011). However, because mutual competition leads to the worst possible outcomes for both parties, each party is better off cooperating unilaterally rather than competing when the other party competes. The asymmetric outcomes are equilibria, as they both provide no incentive to deviate unilaterally (Rapoport & Chammah, 1966; Schelling, 1980). Thus, the optimal strategy in Chicken is to do the opposite of one's counterpart. When the game is repeated indefinitely, alternating between the two asymmetric outcomes gives equal payoffs to both players.

Surface and Deep Similarities in CG and PD. Both CG and PD have two symmetric (win-win and lose-lose) and two asymmetric (win-lose and lose-win) outcomes (see Tables 1 and 2). Additionally, the win-win outcome is numerically identical in both games (i.e., [1,1]). Besides these surface similarities, there are significant differences between the two games. In CG, either of the asymmetric outcomes is more lucrative in terms of joint payoffs than the [1,1] outcome. When the game is played for an indefinite time, an alternation between the two asymmetric outcomes can also be considered a win-win outcome. In this sense, CG has two win-win outcomes: a "weakly optimal" one [1,1] and a "strongly optimal" one (an alternation between [-1,10] and [10,-1]). This is not the case in PD, where an asymmetric outcome [10,-10] is inferior in terms of joint payoffs to the win-win outcome [1,1]. Thus, although the win-win solution corresponds to different choices in the two games ([1,1] in PD versus alternation in CG), they share deep similarities in the sense that the win-win solution is, in the long run, superior to the win-lose solution in both games.

Joint learning of risky yet lucrative strategies. In both PD and CG, learning must occur not only at an individual level but also at a dyad level. If learning only occurs in one individual in a dyad, the outcomes are disastrous for that individual, because the best solution also bears the greatest risk. For example, if only one player understands that alternating between the two moves is the optimal solution in CG, the outcome for that player can be a sequence of -1 and -10 payoffs. Only if both players understand the value of alternation, are willing to alternate, and able to synchronize their individual alternations, the result will be a sequence of 10 and -1 payoffs for each players, which in average gives each player a payoff of 4.5 points per round. Thus, the context of interdependence makes unilateral individual learning not only useless but also detrimental. The two players must jointly learn that only a solution that maximizes joint payoff is viable in the long term. This fact is true for both PD and CG in spite of the apparent differences between the two games. In other words, their deep similarity is that both are symmetric non-zero-sum games in which the solution that maximizes joint payoff is superior to all the other possible outcomes, provided that such a solution can be sustained in the long term. However, the jointly optimal solution carries the most risk and is thus unstable in the long term: each player has a short-term incentive to unilaterally depart from the mutually beneficial solution. To ensure that the jointly optimal solution is maintained from one round to another, there must exist a mechanism that mitigates the risk associated with this solution. We discuss this mechanism in the following section.

Reciprocal trust

In situations where there are benefits to individuals that can be maintained through an interaction based on mutual trust, each individual has an incentive to maintain

the relation. It has been suggested that trust relations are self-sustaining once they have been developed (Hardin, 2002). A trust relation develops through gradual risk-taking and reciprocation (Cook, Yamagishi, Cheshire et al., 2005). It is a learning process in which the truster learns to judge the trustworthiness of the trusted and the trusted learns that it pays off to fulfill the truster's trust (Hardin, 2002).

As trust develops, the risk is reduced and the trust relation becomes more stable. Trust is also a means to reduce complexity (Luhmann, 1980): a trusting player does not need to continuously make judgments about the other player's trustworthiness and their likely actions (Gambetta & Hamill, 2005), thus their cognitive load is significantly reduced.

Although trust is not a symmetric concept (the truster and trusted roles are qualitatively different), the games' symmetrical nature is conducive to the development of reciprocal trust. Both players are at risk, thus they both need to develop trust and become trustworthy. As a matter of fact, ongoing trust relationships are typically reciprocal, "because a good way to get me to be trustworthy in my dealings with you, when you risk acting on your trust of me, is to make me reciprocally depend on your trustworthiness" (Hardin, 2002, p. 17). A reciprocally trusting relationship is self-reinforcing because each person has incentives to be trustworthy (Coleman, 1990).

In both PD and CG, there are clear incentives for a win-win solution to be established and maintained, but there are significant risks as well. It is likely that achieving a win-win solution will be associated with the development of reciprocal trust, which in turn will sustain this solution for a longer term. However, it is not clear whether reciprocal trust will affect the probability that a different win-win solution will be

achieved in a subsequent game. Trust is a three-part relation: *A* trusts *B* to do *X* (Baier, 1986), but not necessarily *Y*. The similarity between *X* and *Y* must be factored in. If *X* and *Y* are sufficiently similar, the transfer of trust is likely to occur. Thus, the transfer of learning between games might simultaneously depend on the relations between players (Fischer, 2009) and between games.

QUESTIONS AND HYPOTHESES

The main goal of this study was to understand the conditions for developing reciprocal trust and for transferring learning between games of strategic interaction. We hypothesized that the transfer of learning occurs primarily based on an understanding of a game's deep characteristics (e.g., learning the optimal solution based on the correspondence of outcomes between the two players), although surface similarities (e.g., numerically similar choices) may influence transfer across games in the short-term. We also hypothesized that reciprocal trust will mediate transfer based on deep similarities between games.

The surface and deep similarities between these two games can lead to different kinds of learning transfer patterns. Learning across the two games could occur because of surface (hypotheses 1a-1c) or deep (hypotheses 2a-2b) similarities. Of course, the third possibility is a combination of these two kinds of similarities (hypothesis 3). Hypothesis 4 is about what could explain the hypothesized deep transfer between the two games (a synopsis of these hypotheses is shown in Table 3).

Surface similarity hypothesis 1a: If the transfer of learning from one game to another is based on surface similarities, we would expect a higher frequency of the [1,1]

outcome in the CG when it is played after, as opposed to before, the PD game.

This hypothesis is derived from the assumption that when playing PD, players will learn that mutual cooperation [1,1] is the optimal solution and thus, will be more likely to select the same option when playing CG. In other words, even though [1,1] is not the most optimal strategy in CG in terms of maximizing joint outcomes, players will transfer their learning from one game to another by finding solutions that are similar in surface characteristics across both and provide a satisfactory outcome.

Surface similarity hypothesis 1b: If the transfer of learning from one game to another is based on surface similarities, we would expect a higher frequency of the alternation outcome in the PD game when it is played after, as opposed to before, the CG game.

This hypothesis is derived from the assumption that when playing CG, players will learn that alternation of [-1,10] and [10,-1] is the optimal solution and will be more likely to select the same strategy when playing PD. In other words, even though alternation is not the optimal strategy in PD, players will transfer their learning from one game to another by finding solutions that are similar in surface characteristics across games.

Surface similarity hypothesis 1c: If the transfer of learning from one game to another is based on surface similarities, we would expect a higher frequency of the [1,1] outcome in the PD game when it is played after, as opposed to before, the CG game.

This hypothesis is derived from the assumption that when playing CG, some players will learn the weakly optimal strategy (i.e., [1,1]) and will be satisfied with this outcome rather than pursuing the strongly optimal strategy (i.e., alternation) in CG. Thus,

if they generalize this learning experience from CG to PD, they will be more likely to select the mutually cooperative solution [1,1] as it shares surface similarity characteristics with the weakly optimal strategy in CG [1,1]. Note that hypotheses 1b and 1c predict opposite patterns in PD based on which surface similarity characteristic participants are learning from. If learning across games is a function of applying the optimal strategy from the first game to the next (even if it is not optimal in the second game), then hypothesis 1b should hold true. If learning across games is a function of applying the numerically consistent strategy from the first game to the next, however, then hypothesis 1c should hold true.

Deep similarity hypothesis 2a: If the transfer of learning from one game to another is based on understanding deep similarities, we would expect higher frequency of alternation in CG when it is played after, as opposed to before, the PD game.

Deep similarity hypothesis 2b: If the transfer of learning from one game to another is based on understanding deep similarities, we would expect higher frequency of mutual cooperation ([1,1]) in PD when it is played after, as opposed to before, the CG game.

These hypotheses are derived from the assumption that successfully finding an optimal solution in one strategic interaction makes one more likely to find an optimal solution in the next strategic interaction, even when the two interactions are structurally different (i.e., there is surface dissimilarity). Specifically, players learn from PD that the optimal strategy is to mutually cooperate and this increases their likelihood of finding the optimal alternation strategy in CG, even though it is a different optimal strategy. Similarly, players learn from CG that the optimal strategy is to alternate and this

increases their likelihood of finding the optimal mutually cooperative strategy in PD, even though it is a different optimal strategy.

Combined effects of surface and deep similarities hypothesis 3: If both surface and deep similarities influence the transfer of learning across games, we predict that transfer from one game to another initially occurs based on surface similarities between the two games, but it is a result of understanding deep similarities between the two in the long term. In other words, we expect the effect of surface similarity on transfer to decrease and of deep similarity on transfer to increase over time as the second game progresses. This hypothesis is derived from the observation that the recognition of deep similarities may be impeded by surface similarities (Holyoak & Koh, 1987). We generalize this observation and hypothesize that surface similarities may facilitate or hinder transfer depending on whether they are congruent or incongruent with deep similarities (Gentner & Loewenstein, 2002; Gentner & Medina, 1998; Gonzalez & Wong, 2012). Thus, we expect a larger transfer effect in the CG-PD order than in the PD-CG order, because both surface and deep similarities point toward the optimal solution in PD, whereas surface similarity points to a suboptimal solution in CG.

Reciprocal trust hypothesis 4: We expect that participants will develop reciprocal trust through their mutual cooperation in the first game. Thus, the higher the level of mutual cooperation in the first game is, the higher the level of reciprocal trust at the end of the first game. Moreover, the trust built after the first game will facilitate mutual cooperation in the second game. The higher the level of reciprocal trust after the first game is, the higher the level of mutual cooperation in the second game will be. Thus, we expect reciprocal trust to be a significant mediator between the optimal outcomes in the

two games, regardless of which game is played first. In addition, as a consequence of learning within and across games, the overall level of reciprocal trust at the end of the second game will be higher than at the end of the first game.

Insert Table 3 around here.

METHODS

Participants

One hundred twenty participants (44 women, 76 men; M_{age} = 22.8, SD_{age} = 2.95) were recruited at Carnegie Mellon University through an online advertisement. Participants were paid a $10.00 base rate for their participation in this study. Additionally, participants were eligible to acquire incentive pay based on their performance in the games as discussed in the procedure section. All participants were paired with anonymous partners when playing the games, leading to 60 pairs.

Design

The hypotheses of interest were tested using a 2 (order of games played: PD-CG and CG-PD) by 2 (games: CG and PD) design. The order of games played was a between-subject factor and the game type was a within-subject factor in this design. There were 60 participants and consequently 30 pairs in each game order condition (PD-CG and CG-PD). Participants played 200 unnumbered rounds of each game (CG and PD).

Procedure

All participants gave informed consent after receiving a general introduction

about what the study entailed. After participants signed the consent form, they were asked to complete a brief demographic questionnaire on a computer. Next, participants were randomly selected to play the Chicken game before (CG-PD) or after (PD-CG) the Prisoner's Dilemma Game. Specifically, the computer program informed participants that they have been randomly matched with an anonymous opponent in the other lab to play an interactive game. The instructions further stated that the "game consists of a series of rounds in which you and your opponent will make simultaneous decisions that affect each of your payoffs." Furthermore, the instructions stated that "both of your payoffs will depend on your own action as well as your opponent's action." After reading through the instructions, participants were paired to play 200 unnumbered rounds of each game, PD and CG, over the Internet. In each round of each game, the two anonymous members of a pair (seated in different rooms) chose simultaneously between buttons labeled Action A and Action B with payoffs as in Tables 1 and 2. The game matrix was visible throughout the game. After each choice, participants received full information about both players' choices and payoffs.

Participants started the experimental session with $0 and were able to accumulate points through each game that were ultimately transformed into incentive pay (one cent per point). If participants accumulated $10 or less through game points at the end of their session, they were paid the $10 base pay. If participants accumulated more than $10 through their game points, they were given the $10 base pay as well as their accumulated incentive beyond the $10.00. The average total amount accumulated in CG and PD was M = 1.97 (Range = -19.91 - 10.06; SD = 6.18) and M = 0.24 (Range = -10.00 - 11.60; SD = 2.71), respectively.

At the completion of each game, participants completed a five-item questionnaire online (see Appendix A) assessing several factors: how satisfied they were with the overall outcome of the game; how fair they thought the opponent's actions were; how fair the participants' actions were towards their opponents; how trustful they were of the opponent; and how trustful of them the opponent was.

RESULTS

To study the transfer of learning across both games, we analyzed the outcomes of a game when it is played after the other game as compared with when the game is played before the other game. We also analyzed the round-by-round dynamics of these outcomes. The statistical significance of the observed effects was tested with the aid of a Linear Mixed Effects analysis (*lmer* analysis from the LME4 package in R). This analysis was preferred over the classical analysis of variance (ANOVA) because the data violated the assumption of normality.

Transfer driven by surface similarities

If learning across games is driven by surface similarities, one would expect the strategy learned in the first game to be applied in the second game as well, even though it may not be appropriate for the second game. This may happen especially in the beginning of the second game.

The outcomes in PD when it was played first are displayed in Figure 1. The asymmetric outcomes ([-10,10] and [10,-10]) are not sustainable, their frequency decreases to almost zero; the [-1,-1] outcome starts high and increases, but its frequency starts to decrease after about 70 rounds; and the [1,1] outcome starts low and increases in

frequency as the game unfolds. By the end, participants have learned that maximizing joint payoff is the only realistic way to maximize individual payoff in the long run. Is this strategy transferred as such to the second game? Is the [1,1] outcome more likely to occur in CG when CG is played after PD than when it is played before?

Insert Figure 1 here

The frequency of the [1,1] outcome in CG by order (CG-PD and PD-CG) and by round is displayed in Figure 2. A LME model with occurrence of [1,1] as a dependent variable (binomial distribution), order, round, and their interaction as fixed factors, and participant as a random factor was used to test the observed effects. There was a main effect of order ($z = 2.21$, $p = 0.027$) and a main effect of round ($z = -8.171$, $p < 0.001$); the interaction between order and round was also significant ($z = -7.196$, $p < 0.001$); indicating that the main effect of order is larger at the beginning of the game and it becomes progressively smaller.

Insert Figure 2 here

These results provide support for hypothesis 1a. Specifically, when CG was played after PD, players were more likely to try to achieve the [1,1] outcome because it appeared to be (on the surface) identical to the optimal outcome from the previous game. However, it is much less optimal in CG. Thus, it decreases in frequency as the game unfolds and both players discover the more optimal alternation strategy. The maximum

difference between the two conditions was seen at the beginning of the CG game, suggesting that the effect of surface similarity decreased over time.

The game outcomes for CG when it is played first are displayed in Figure 3. The [1,1] outcome occurred at a higher frequency in the first half of the game, but then decreased in frequency in the second half. The alternation of the two asymmetric outcomes increased in frequency throughout the game. Participants were more likely to find the weakly optimal solution [1,1] in the beginning, however more and more dyads gradually discovered the stronger optimal solution (the alternation of [10,-1] and [-1,10] outcomes).

Insert Figure 3 here

If the transfer of learning across games is driven by surface similarities, one would expect the strategy of alternating between the two asymmetrical outcomes to be attempted in the second game as well, at least in the beginning. Figure 4 shows the frequency of alternation in PD by order (PD-CG vs. CG-PD) and by round. The main effect of order was non-significant ($z = 1.476$, $p > 0.10$), suggesting that the strongly optimal strategy in CG (alternation) was not transferred as such (based on surface similarities) to PD. The fact that there is a slightly increased frequency of alternations in the beginning of the game indicates that players were exploring different strategies and had not yet settled on a dominant one.

Insert Figure 4 here

This result provides evidence for rejecting hypothesis 1b. The fact that alternation was not transferred to the second game may be attributed to the very weak surface similarity between the alternation strategy in CG (alternating [10,-1] and [-1,10]) and the alternation strategy in PD (alternating [10,-10] and [-10,10]).

There remains the possibility that the [1,1] outcome was transferred as such from CG to PD (hypothesis 1c). Even though the [1,1] outcome is only weakly optimal in CG, it was selected with relatively high frequency (see Figure 3) and it might have been considered optimal by some participants. We will revisit this point in the section about the combined effects of surface and deep similarities.

Transfer driven by deep similarities

If learning across games is driven by deep similarities, one would expect that learning the optimal strategy in the first game would increase the probability of learning the optimal strategy in the second game, even though there is no surface similarity between these strategies. These strategies ([1,1] in PD and alternation in CG) are similar only on an abstract, deep level: they both aim at maximizing joint payoff in a sustainable way, which is realistically possible only if the two players make (approximately) equal payoffs in the long run. On a surface level, these two strategies are very different. The [1,1] strategy in PD requires that players make the same move each trial and they do not switch to the opposite move. In contrast, the alternation strategy in CG requires that players make the opposite move each round and they continuously switch between the two.

The frequency of the alternation outcome in CG by order and by round is seen in

Figure 5. There was a main effect of order (z = -2.014, p = 0.044), indicating a higher level of alternation when CG was played after PD. There was a main effect of round (z = 16.205, p < 0.001), indicating that more and more pairs of participants discovered the alternation strategy as the game unfolded. There was also a significant interaction between order and round (z = 8.5, p < 0.001), indicating that the optimal strategy was learned faster when CG was played second.

Insert Figure 5 here

The frequency of the [1,1] outcome in PD by order and by round is displayed in Figure 6. There was a main effect of order (z = -4.340, p < 0.001), indicating that more pairs of participants discovered the optimal strategy in PD when it was played after CG. There was a main effect of round (z = 10.149, p < 0.001), indicating that more and more pairs of participants found the optimal strategy as the game unfolded. There was also a significant interaction between order and round (z = 12.689, p < 0.001), indicating that the optimal strategy reached a ceiling when PD was played after CG, whereas it increased continuously when PD was played before CG.

Insert Figure 6 here

These results supported hypotheses 2a and 2b. Specifically, learning the optimal strategy in the first game increased the probability of learning the optimal strategy in the second game.

Combined effects of surface and deep similarities

In the case of deep transfer, the main effect of order was smaller in magnitude for CG (z = -2.014, p = 0.044) than for PD (z = -4.340, p < 0.001) (see also Figures 5 and 6). It seems as if CG has a stronger impact on PD than vice versa. A possible explanation for this difference is based on how surface and deep similarities combine with each other to drive the transfer of learning across games. They may have congruent or incongruent effects. Thus, in the PD-CG order (Figure 7a), surface and deep similarities act in a divergent, incongruent way: surface similarity makes it more likely that the [1,1] outcome is selected (see also Figure 2) whereas deep similarities make it more likely that the alternation outcome is selected (see also Figure 5). In other words, transfer based on surface similarity interferes with transfer based on deep similarity. In contrast, in the CG-PD order (Figure 7b), both types of similarities act in a convergent, congruent way: they both increase the probability that the [1,1] outcome is selected. There is no impeding effect of surface similarity on PD because there is no optimal strategy in CG that is similar enough to a non-optimal or sub-optimal strategy in PD. These results are consistent with hypothesis 3.

Insert Figure 7 here.

Joint learning and emergence of reciprocal trust

In addition to game choices, we analyzed the debriefing questionnaires that were administered at the end of each game. Since the answers were highly correlated with each

other for any one individual participant, we summed them up in one composite variable that we call Reciprocal Trust[3] (see correlation matrix in Appendix B). Since the debriefing questions were administered twice (at the end of each game), we refer to them as T1 and T2. The Cronbach's alphas for T1 and T2 were 0.80 and 0.90, respectively.

We calculated correlations between these two trust variables and the variables indicating mutual cooperation in the two games. Spearman's *rho* coefficient was used for correlations because the data failed to meet the normality assumption. We found that the more frequent the win-win outcome ([1,1] for PD and alternation for CG) was in the first game, the more likely the players were to rate each other as trustworthy at T1 ($r = 0.75$, $p < 0.001$ for PD and $r = 0.42$, $p < 0.001$ for CG). In addition, the more trustworthy players rated each other at T1, the more likely they were to enact the win-win outcome in the second game ($r = 0.28$, $p = 0.03$ for CG and $r = 0.47$, $p < 0.001$ for PD). Finally, the win-win solution in the second game predicted high levels of trust at T2 ($r = 0.67$, $p < 0.001$ for CG and $r = 0.88$, $p < 0.001$ for PD).

As expected, the level of reciprocal trust increased from T1 to T2 in the PD-CG condition ($mean_{T1} = 11.08$, $mean_{T2} = 16.07$, $t = -4.98$, $p = 0.000$). In the CG-PD condition, T2 was not significantly different than T1 ($mean_{T1} = 12.47$, $mean_{T2} = 12.18$, $t = 0.38$, $p = 0.70$).

Mediation analysis. The correlations presented above suggest that reciprocal trust is a mediator between the learned optimal outcomes in the two games and can explain the observed transfer of learning across them, regardless of the order in which the games

[3] Only the composite Trust variable will be used here. However, we replicated all the analyses presented here with a Trust variable defined only based on the trust items. The results were not significantly different than the ones presented here.

were played. Mediation analysis was conducted with the aid of the R package "mediation." The bootstrapping method was used with 1000 simulations. This method is known to be robust against violations of some of assumptions, in our case, the normality assumption (Imai, Keele, & Iamamoto, 2010; Imai, Keele, & Tingley, 2010; Tingley, Yamamoto, Keele, & Imai, 2011). We present here only the results of mediation analysis for the case of the deep transfer of learning between the games in both directions (PD-CG and CG-PD). We also conducted mediation analysis for the case of surface transfer, although mediation was not hypothesized in this case. As expected, there was no significant mediation by reciprocal trust in the case of surface transfer.

PD-CG order. The [1,1] outcome in PD significantly predicts the alternation outcome ([-1,10]/[10,-1]) in CG (Adjusted R^2 = 0.18, p=0.00), which supports the hypotheses of deep transfer from PD to CG. The [1,1] outcome in PD significantly predicts reciprocal trust (Adjusted R^2 = 0.62, p=0.00). That is, the more often pairs of players achieve mutual cooperation in PD, the more likely they are to report higher levels of reciprocal trust at the end of the game. Reciprocal trust significantly predicts the alternation outcome in CG (Adjusted R^2 = 0.08, p=0.02). Thus, the prerequisites for mediation analysis are met. However, the mediation effect is non-significant (Mediation Effect = -0.08; 95 % CI = [-0.42, 0.24]).

CG-PD order. The alternation outcome ([-1,10]/[10,-1]) in CG significantly predicts the [1,1] outcome in PD (Adjusted R^2 = 0.07, p=0.03), which supports the hypothesis of deep transfer from CG to PD. The alternation outcome in CG significantly predicts reciprocal trust (Adjusted R^2 = 0.17, p=0.01). That is, the more often pairs of players achieve the win-win solution in CG, the more likely they are to report higher

levels of reciprocal trust at the end. Reciprocal trust significantly predicts the [1,1] outcome in PD (Adjusted R^2 = 0.17, p=0.00). Thus, the prerequisites for mediation analysis are met and the mediation effect is significant (Mediation Effect = 0.23; 95 % CI = [0.08, 0.44]).

Discussion of hypothesis 4. These results are only partially consistent with hypothesis 4. However, the inconsistency can be understood through careful examination of the two games' structure and the interaction between surface and deep similarities across the games.

First, in the CG-PD condition, T2 was not significantly higher than T1 as we had expected. A possible explanation for this finding is that there were two different outcomes in CG that could generate reciprocal trust ([1,1] and alternation), while there was only one outcome in PD that could generate or maintain reciprocal trust ([1,1]). This explanation is supported by the fact that T1 in the CG-PD condition (12.47) is higher than T1 in the PD-CG condition (11.08). Given these circumstances, the fact that T2 did not significantly decrease is a result that partly supports our hypothesis or at least does not contradict it.

Secondly, the mediation effect was not significant in the PD-CG order. As mentioned above when we analyzed the combined effects of surface and deep similarities, surface similarities interfered with deep transfer in this case. Part of the transfer of learning from PD to CG was driven by the fact that the [1,1] outcome was (on the surface) identical in the two games. This surface transfer diluted the strong relation that potentially could have occurred between reciprocal trust and the optimal strategies in both games.

In sum, reciprocal trust between players in game 1 was positively associated with cooperation and reciprocal trust in game 2 as reflected in the correlations for both orders. The lack of a significant mediation effect in one order but not in the other can be understood given the structural and strategic differences in optimal strategies between the two games.

DISCUSSION AND CONCLUSION

The main goal of this study was to understand the conditions for development of reciprocal trust and transfer of learning between games of strategic interaction. We hypothesized that transfer of learning occurs primarily based on an understanding of a game's deep characteristics and it is mediated by development and maintenance of reciprocal trust between players. We tested these predictions by using two specific games of strategic interaction, Prisoner's Dilemma and Chicken.

We found that both surface and deep similarities drive transfer of learning. In particular, we found that surface similarities can be conducive to transfer when they are congruent with the optimal solution in the target game ([1,1] in PD). In contrast, when surface similarities point to a suboptimal solution in the target game ([1,1] in CG), they interfere with the deep transfer of the optimal solution. In other words, surface similarities facilitate transfer in one direction (CG-PD) and hinder it in the opposite direction (PD-CG). Nevertheless, deep similarities were the main drive of transfer across two games in both directions. Gonzalez and Wong (2012) have found similar results with different tasks.

The results presented above would have been remarkable even if they were found in the realm of individual cognition, where transfer of learning from one task to a new,

different task is rarely found (Singley & Anderson, 1989). What makes these results even more impressive is that they occur in a context of interdependence. We found evidence that the transfer of learning between interdependent players is facilitated by reciprocal trust.

The results from this study add to the previous literature in at least two ways. First, results from previous studies exploring learning transfer across sequential games of strategic interaction have found transfer to be dependent on the presence of surface similarities between the games (Knez & Camerer, 2000). Based on these findings, the authors concluded that "transfer is mostly limited to those cases in which activities have similar descriptions, rather than to those with similar strategic structures and different descriptions" (Knez & Camerer, 2000, p. 214). However, our results suggest that individuals and pairs who have learned the optimal strategy within one game of strategic interaction are more likely to choose the optimal strategy in the next game, even when the two games and their optimal solutions share surface and structure dissimilarities. This transfer across games is likely due to an understanding of the games' abstract, deep characteristics that are similar across both.

Future research needs to specifically test how long the effect of transfer of learning lasts. Although the second game was played immediately after the first in our experiment, we expect this transfer of learning to extend beyond the immediate short-term as it was due to an understanding of abstract, deep characteristics and presumably this kind of learning leads to broad conceptual changes (Gentner & Loewenstein, 2002; Gentner & Medina, 1998). Future research should also explore if learning optimal solutions within games of strategic interaction can generalize to finding optimal solutions

in real-world strategic interactions (e.g., negotiations).

Second, this experiment provides a novel test of learning transfer within the context of games of strategic interaction. Most of the previous work focused on understanding the effects of surface and deep similarities has been done in the context of individual learning through examples or analogies. We find that the transfer of learning occurring across games depends not only on the nature and relations between games, but also on the development of reciprocal trust between players. We hope future research attempts to further integrate these two levels of analysis to answer other theoretically interesting questions. For example, an interesting experiment would manipulate surface and deep similarities between individuals and study how they interact with surface and deep similarities between tasks.

Our research also has some practical implications. For example, the learning material within training programs must be designed in such a way to avoid the detrimental effect of surface similarity on transfer. If a game is used to develop skills of strategic interaction in the real world, we might want to avoid designing it to match all specific aspects of the real-world situation. The more similarities between the game and the real world that are in place, the higher the chance that surface similarities may interfere with the transfer driven by deep similarities (as in the PD-CG condition). Thompson, Gentner, & Loewenstein (2000) found that lack of surface similarities was beneficial to analogical reasoning in negotiation experiments.

Moreover, surface similarities can be used to boost deep transfer of learning across tasks. Referring to the same example, if the game of strategic interaction matches the real-world situation only with regard to those features that drive deep transfer, the

desired effect is enhanced (as in the CG-PD condition). From this perspective, an abstract game that only matches the target situation in its essential features is to be preferred over a realistic game. The same reasoning could be applied to the relation between individuals. It has been shown that similarity and mimicry can influence reciprocal trust and cooperation (Fischer, 2009; Van Baaren, Holland, Kawakami, & van Knippenberg, 2004). However, given that trust is context-dependent (Hardin, 2002), it is important that what players discover about each other are the kind of similarities that is conducive to joint learning of the optimal outcome. Some of the similarities or dissimilarities between players might interfere with or distract from the optimal solution.

It is important to emphasize that the results reported here are likely to generalize only to those situations where there is a deep similarity between the optimal strategies in the prime and target games. If the two games have fundamentally different optimal strategies, positive transfer of learning is unlikely. The mediation of transfer through reciprocal trust is likely to generalize only to situations in which there is interdependence between players and there are incentives for both players to continue (repeat) the interaction (game) for a long time. This is obviously not the case when the counterparts are changed at each round and players are unlikely to interact again in the future. A special case is one where counterparts are able to exchange information. In this case, there are incentives for players to be perceived as trustworthy even when they change counterparts after each round. Thus trust may still be a mediator of strategic transfer between games, though to a lesser extent, due to the interpersonal-intergroup discontinuity effect (Wildschut, Pinter, Vevea, Insko, & Schopler, 2003).

Our study cannot disentangle success and failure in the first game as causes of

better performance in the second game. Both success and failure are probably important in learning, but we cannot speculate as to which one is more important because they were not independent and we did not manipulate game difficulty in this study.

In conclusion, the results presented here suggest that learning can transfer across games of strategic interaction that share deep similarities and also facilitate the development of reciprocal trust. Our research suggests that games of strategic interaction can be used as training tools provided that they are adequately designed so that they share deep similarities with the target activity, promote the building of reciprocal trust, and do not share surface similarities that may cause the transfer of a suboptimal solution.

REFERENCES

Ahn, T.K., Ostrom, E., Schmidt, D., Shupp, R., & Walker, J. (2001). Cooperation in PD games: Fear, greed, and history of play. *Public Choice, 106*(1-2), 137-155. doi: 10.1023/A:1005219123532

Baier, A. (1986). Trust and antitrust. *Ethics, 96*(2), 231–60.

Bednar, J., Chen, Y., Xiao Liu, T., & Page, S. E. (2012). Behavioral spillovers and cognitive load in multiple games: An experimental study. *Games and Economic Behavior*, *74*(1), 12-31. doi: 10.1016/j.geb.2011.06.009

Camerer, C. F. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton, NJ: Princeton University Press.

Coleman, J. S. (1990). *The foundations of social theory*. Cambridge, MA: Harvard University Press.

Cook, K. S., Yamagishi, T., Cheshire, C., Cooper, R., Matsuda, M., & Mashima, R. (2005). Trust building via risk taking: A cross-societal experiment. *Social Psychology Quarterly, 68*(2), 121-142. doi: 10.1177/019027250506800202

Fischer, I. (2009). Friend or foe: Subjective expected relative similarity as a determinant of cooperation. *Journal of Experimental Psychology: General, 138*(3), 341-350. doi: 10.1037/a0016073

Gambetta, D., & Hamill, H. (2005). *Streetwise: How taxi drivers establish their customers' trustworthiness*. New York: Russell Sage Foundation.

Gentner, D., & Loewenstein, J. (2002). Relational language and relational thought. In J. Byrnes & E. Amsel (Eds.), *Language, literacy, and cognitive development* (pp. 87–120).

Mahwah, NJ: Erlbaum.

Gentner, D., & Medina, J. (1998). Similarity and the development of rules. *Cognition, 65*(2-3), 263–297. doi: 10.1016/S0010-0277(98)00002-X

Gonzalez, C., & Wong, H. (2012). Understanding stocks and flows through analogy. *System Dynamics Review*. 28(1), 3-27. doi: 10.1002/sdr.470

Halevy, N., Chou, E. Y., & Murnighan, J. K. (2011). Mind games: The mental representation of conflict. *Journal of Personality and Social Psychology, 102*(1), 132-148.. doi: 10.1037/a0025389

Hardin, R. (2002). *Trust and trustworthiness*. New York: Russell Sage Foundation.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory and Cognition, 15*(4), 332–340. doi: 10.3758/BF03197035

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods, 15*(4), 309-334. doi: 10.1037/a0020761

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2010). Causal mediation analysis using R. *Lecture Notes in Statistics, 196*, 129-154. doi: 10.1007/978-1-4419-1764-5_8

Knez, M., & Camerer, C. (2000). Increasing cooperation in Prisoner's Dilemmas by establishing a precedent of efficiency in coordination games. *Organizational Behavior and Human Decision Processes, 82*(2), 194-216. doi: 10.1006/obhd.2000.2882

Levitt, S. D., List, J. A., & Sadoff, S. E. (2011). Checkmate: Exploring backward induction among chess players. *American Economic Review, 101*(2), 975-990.

Luhmann, N. (1980). Trust: A Mechanism for the Reduction of Social Complexity. In N. Luhmann (Ed.), *Trust and power: Two works by Niklas Luhmann*. New York: Wiley.

Palacios-Huerta, I., & Volij, O. (2009). Field centipedes. *American Economic Association, 99*(4),

1619-1635. doi: 10.1257/aer.99.4.1619

Plott, C. R., & Smith, V. L. (2008). *Handbook of experimental economics results*. Amsterdam, the Netherlands: Elsevier.

Rapoport, A., & Chammah, A. M. (1965). *Prisoner's dilemma: A study in conflict and cooperation*. Ann Arbor, MI: University of Michigan Press.

Schelling, T. C. (1980). *The strategy of conflict* (2nd ed.). Cambridge, MA: Harvard University Press.

Schotter, A. (1998). Worker trust, system vulnerability, and the performance of work groups. In A. Ben-Ner & L. G. Putterman (Eds.), *Economics, values, and organization* (pp. 364-407). Cambridge, UK: Cambridge University Press.

Singley, M. K., & Anderson, J. R. (1989). *Transfer of cognitive skill*. Cambridge, MA: Harvard University Press.

Thompson, W. L., Gentner, D., & Loewenstein, J. (2000). Avoiding missed opportunities in managerial life: Analogical training more powerful than individual case training. *Organizational Behavior and Human Decision Processes, 82*(1), 60–75. doi: 10.1006/obhd.2000.2887

Tingley, D., Yamamoto, T., Keele, L., & Imai, K. (2011). mediation: R Package for Causal Mediation Analysis (Version 4.1.2) [Software]. Available from http://CRAN.R-project.org/package=mediation.

Van Baaren, R. B., Holland, R. W., Kawakami, K., & van Knippenberg, A. (2004). Mimicry and prosocial behavior. *Psychological Science, 15*(1), 71-74. doi: 10.1111/j.0963-7214.2004.01501012.x

Van Huyck, J. B., Battalio, R. B., & Beil, R. O. (1991). Strategic uncertainty, equilibrium

selection principles, and coordination failure in average opinion games. *The Quarterly Journal of Economics*, *106*(3), 885–910. doi: 10.2307/2937932

Wildschut, T., Pinter, B., Vevea, J. L., Insko, C. A., & Schopler, J. (2003). Beyond the group mind: A quantitative review of the interindividual-intergroup discontinuity effect. *Psychological Bulletin, 129*(5), 698-722. doi: 10.1037/0033-2909.129.5.698

APPENDIX A

Debriefing questionnaire administered at the end of each game.

1) How satisfied are you with the overall outcome of the game?

1	2	3	4	5
Very dissatisfied	Somewhat dissatisfied	Neutral	Somewhat satisfied	Very satisfied

2) How fair were your opponent's actions towards you?

1	2	3	4	5
Very unfair	Somewhat unfair	Neutral	Somewhat fair	Very fair

3) How trustful were you of your opponent?

1	2	3	4	5
Very distrustful	Somewhat distrustful	Neutral	Somewhat trustful	Very trustful

4) How fair were your actions toward your opponent?

1	2	3	4	5
Very unfair	Somewhat unfair	Neutral	Somewhat fair	Very fair

5) How trustful of you do you think your opponent was?

1	2	3	4	5
Very distrustful	Somewhat distrustful	Neutral	Somewhat trustful	Very trustful

APPENDIX B

Correlation Matrix between the items of the debriefing questionnaire administered after the first and the second game.

	OF1	OT1	SF1	ST1	S1	OF2	OT2	SF2	ST2	S2
OF1										
OT1	0.56									
SF1	0.29	0.50								
ST1	0.55	0.50	0.43							
S1	0.64	0.47	0.27	0.45						
OF2	0.28	0.26	0.19	0.32	0.23					
OT2	0.26	0.28	0.20	0.30	0.23	0.73				
SF2	0.22	0.31	0.34	0.33	0.17	0.58	0.67			
ST2	0.30	0.22	0.10	0.33	0.24	0.86	0.84	0.57		
S2	0.19	0.26	0.18	0.15	0.15	0.75	0.73	0.53	0.75	

Legend:

OF – Opponent Fairness

OT – Opponent Trust

SF – Self Fairness

ST – Self Trust

S – Satisfaction

The numbers 1 and 2 refer to when the item was administered, after the first and the second game, respectively.

TABLES

Table 1: Prisoner's Dilemma payoff matrix. The cells show a pair of outcomes (x, y) where x is the payoff to Player 1 and y is the payoff to Player 2.

		Player 2 Action: A	Player 2 Action: B
Player 1 Action	A	-1, -1	10, -10
	B	-10, 10	1, 1

Table 2: Chicken payoff matrix. The cells show a pair of outcomes (x, y) where x is the payoff to Player 1 and y is the payoff to Player 2.

		Player 2 Action	
		A	B
Player 1 Action	A	-10, -10	10, -1
	B	-1, 10	1, 1

Table 3: Synopsis of hypotheses.

1. Surface similarities	1.a	Higher [1,1] in CG after than before PD
	1.b	Higher alternation in PD after than before CG
	1.c	Higher [1,1] in PD after than before CG
2. Deep similarities	2.a	Higher alternation in CG after than before PD
	2.b	Higher [1,1] in PD after than before CG
3. Combined surface and deep similarities	Surface and deep transfers are congruent in CG-PD and incongruent in PD-CG.	
4. Emergence of reciprocal trust	Reciprocal trust develops during the first game and mediates deep transfer of learning to second game	

Figure

OBHDP Figures

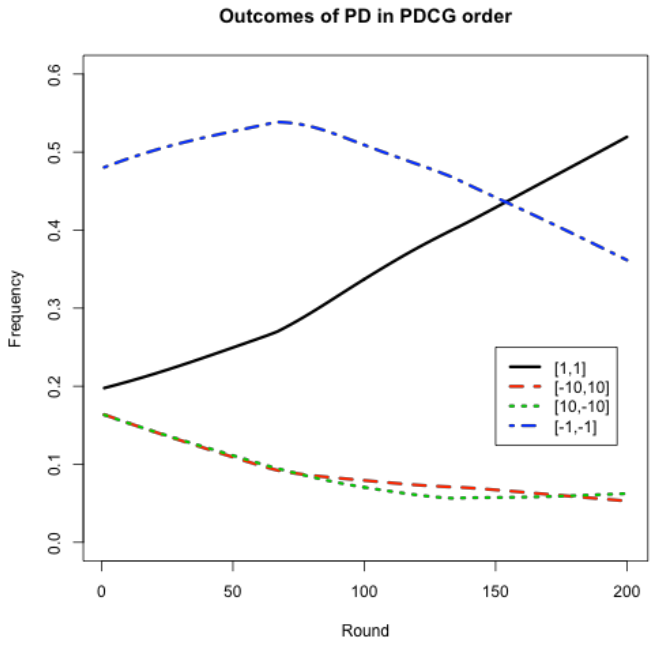

Figure 1. Time course of outcomes in PD when played first.

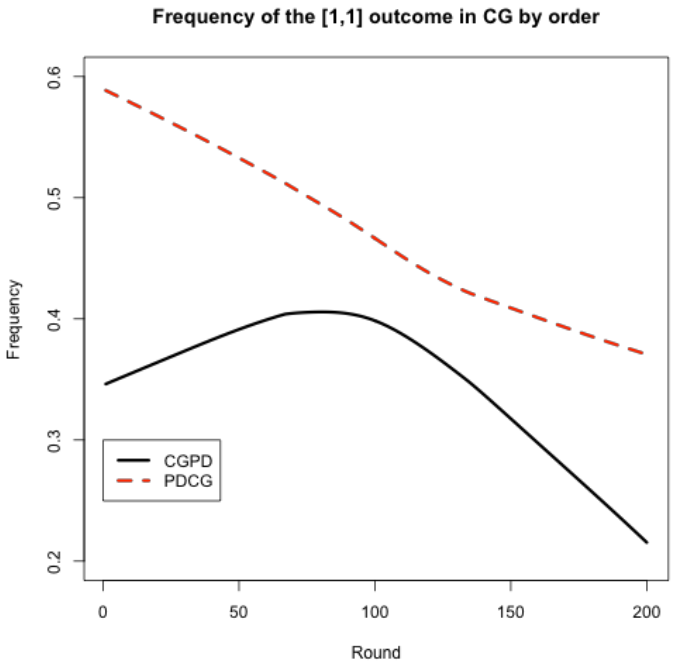

Figure 2. Frequency of the [1,1] outcome in CG by order and round

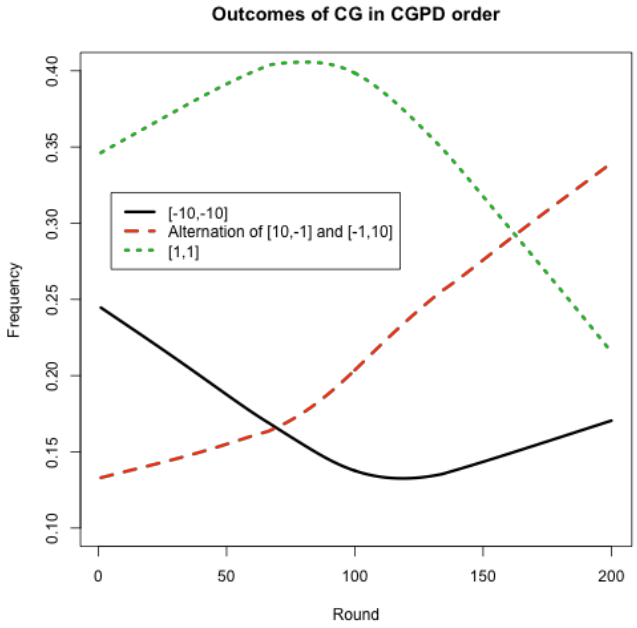

Figure 3. Time course of outcomes in CG when played first

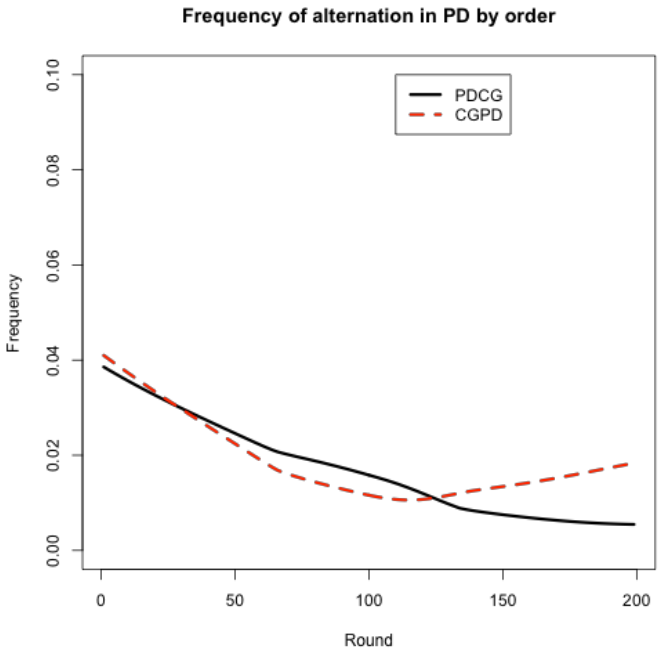

Figure 4. Frequency of alternation in PD by order and round

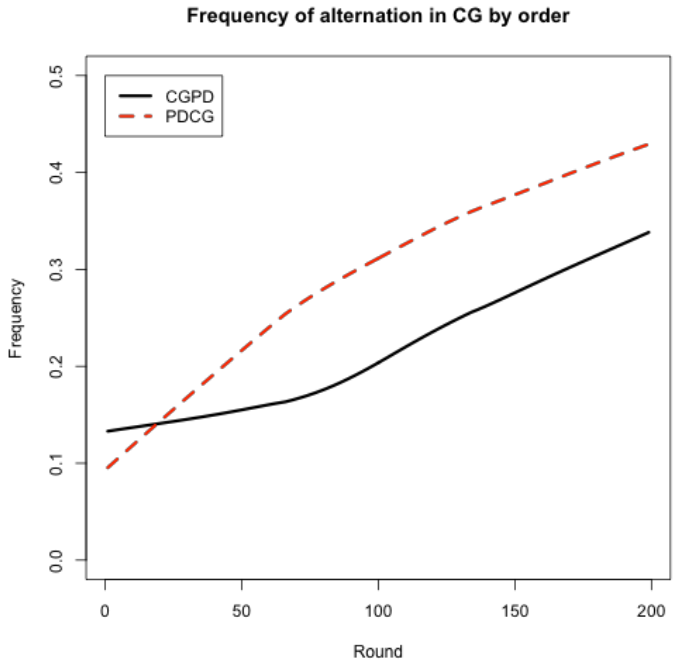

Figure 5. Frequency of alternation in CG by order and round

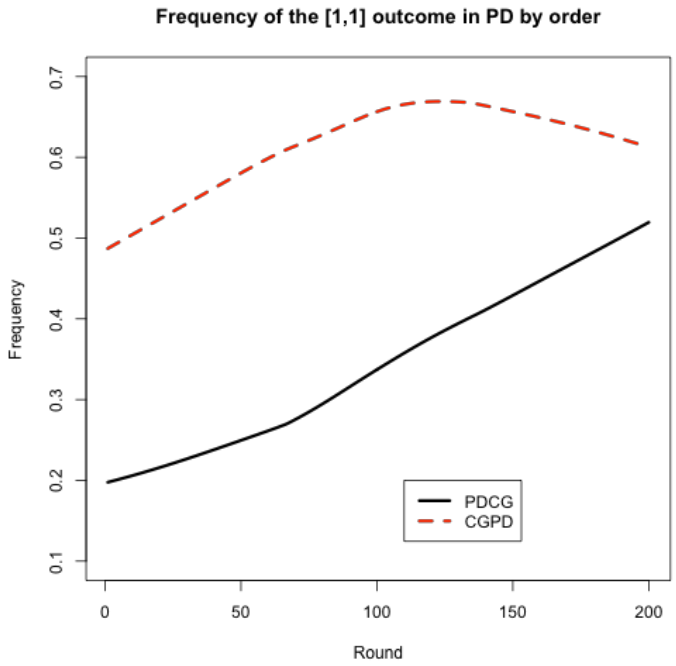

Figure 6. Frequency of the [1,1] outcome in PD by order and round

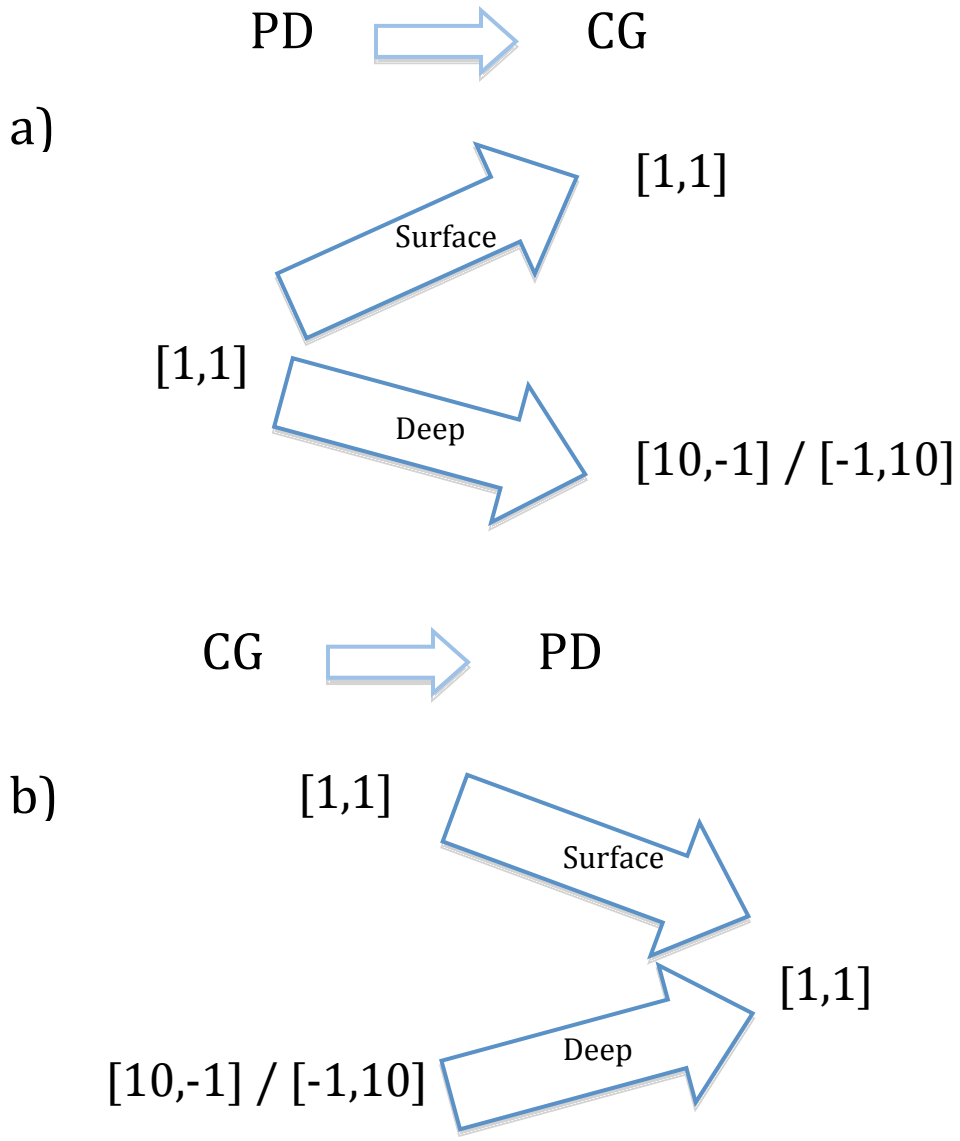

Figure 7. Combined effects of surface and deep-level similarities.