UNIVERSITY COLLEGE LONDON
MSc in Cognitive and Decision Sciences

The Allocation of Responsibility amongst Multiple Causes

Supervisor: Dr David Lagnado

Date: 19/08/2008
Candidate Number: CDS 06
11,730 words

Table of Contents

1 Abstract

How do people assign responsibility to an individual cause or person in a situation of collective responsibility? This dissertation addresses the question by designing an experimental game, the Triangle Game (TG), which participants play in a group with three computer players. The participants' task is to count triangles presented in complex diagrams for a brief period of time. Whether the group wins or loses depends on the accuracy of each player in the group. After each round, participants assign responsibility for the result to each player. Three experimental conditions differ in how the individual judgments are combined to determine the deviation of the group's answer from the correct solution. This deviation determines whether a round is lost or won. For the three experimental conditions, the group's deviation is determined, respectively, by the sum of each player's deviation, the deviation of the least accurate player or the deviation of the most accurate player. The results show that these differences in the underlying causal structure have a significant influence on participants' responsibility ratings. Furthermore, the predictions of different cognitive models are tested. A model for assigning responsibility to individual causes in cases of multiple causation developed by Chockler and Halpern (2003) describes the empirical data best. A second experiment replicates these findings. In an additional step, participants were asked to change the result of each round, for example from a loss to a win, by minimally altering the answers that the players had given in that round. The results show that participants perceive a possible world, where several small changes have been made as more similar to the actual world than a possible world with one big change. The implications of these findings for counterfactual theories of causation are discussed. The main advantage of the TG is that it allows testing psychological theories of responsibility attribution in a formally rigorous manner.

309 words

2 Introduction

How do people attribute responsibility to individual causes in situations, where multiple causes concur to generate a particular outcome? Or, stated somewhat differently, how do people decide to what extent an individual person should be held responsible for an event that several people have brought about collectively? This question, besides being interesting in itself, is of substantial importance for the legal system. How is a business organization or a whole country supposed to be punished for negative outcomes that resulted from its collective course of action? In a time when it is easier and easier for people to connect with each other and when many organizations operate on a global scale, the boundaries between the individual and the collective sometimes appear to dissolve.

To find situations of collective responsibility, one does not even have to go as far as the political or economic sphere. Situations of collective responsibility occur quite naturally and frequently in our everyday life. Hence, we would like to emphasise the general problem by pointing out a few everyday examples. The reader is encouraged to think about how he or she would assign responsibility to individual causes or persons in these examples.

As a first example, consider a penalty shoot-out at the end of a football game. The individual players can be viewed as causes, who together determine the outcome of their team. While it could be said that each player is only responsible for his own shot, which he can either hit or miss, the players together are collectively responsible for the win or loss of the team. Another example is that of a house being built. The house can either be built on time or not. The multiple causes are the different suppliers who need to deliver their materials on time. The house can only be built on time if all suppliers deliver on time. As a final example imagine the following situation: you and your friends would like to go into a very prestigious club. The bouncer at the door, however, insists that your group will only be granted access if at least one of you knows the party's host personally. In this situation, you and your friends

denote the causes, either knowing the host or not, and the outcome is whether your whole group gets in or not.

In each of these situations one can think of various patterns of values that the individual causes could take on and the influence that these patterns would have on people's judgments about the extent to which certain persons or causes are to be held responsible. In subsequent sections, we will refer to these three examples as illustrations of more abstract concepts.

From the examples mentioned above, it already becomes clear that there is a very close relationship between causation and responsibility (see for example Shafer, 2001). Our general intuition is that somebody can only be held responsible for something if her action or omission to act was in some way causally connected to the outcome. In the first part of the introduction, we would like to review how the general relationship between causality, responsibility and blame has been analyzed by different scientific disciplines. In the second part of the introduction, we will look more specifically at the problem of assigning responsibility to individual causes in situations of collective responsibility.

2.1 The Relationship between Causality, Responsibility, and Blame

The question of how causality, responsibility and blame or praise are interconnected is a matter of particular interested for different fields, such as, philosophy, law, psychology and artificial intelligence. We will discuss their take on the problem in turn.

2.1.1 Philosophy

What follows is a brief overview of two significant strands of thinking on causality through philosophy. Unfortunately, the constraints of this study prevent detailed discussion of these theoretical propositions.

Since Aristotle, the concept of causality has been a matter of continuing philosophical debate: most philosophers were and are interested in the metaphysics of causation. What is it that makes something a cause? Different theories have been developed in attempt to explain what causation - the "cement of the universe" (Mackie, 1974) – consists of. Some of them will be elaborated at a later point.

Another important debate that continues to this day focuses on the relationship between causation and responsibility. The question more precisely, is whether a deterministic world is compatible with the idea of moral responsibility. Advocates of determinism believe that it is in principle possible for a supercomputer, which has perfect knowledge about the laws of nature and complete knowledge of the current state of the world to predict the future. This computer could, hence, also predict all the actions that people will perform, even before they are actually born. However, we could argue that somebody is only responsible for an action if he had a choice that is, if he had the opportunity to act differently. Some philosophers have argued that determinism, if it is true, rules out responsibility. On the contrary, others have argued that the idea of a deterministic world is indeed compatible with our understanding of responsibility (Strawson, 1974).

2.1.2 Law

As already mentioned above, the notion of causality is a fundamental concept in law. In order to punish a perpetrator for his actions, a causal connection from his behaviour to the harm being done has to be established. In their seminal book "Causation in the Law", Hart and Honoré (1959) provide an extensive analysis of causal concepts and specify their application in the common law by discussing various legal cases. The coverage of cases discussed and the breadth of underlying causal structures is suggestive of the immense diversity encompassed by the conceptual umbrella of causality. Because of its importance to the legal system it is worth exploring how the notion of causality is commonly defined in legal codes.

In order to establish a causal connection, the law applies two different criteria. First, there is the *factual cause* criterion. To determine whether an event *a* was the cause of an event *b*, that is whether *a* was the factual cause of *b*, the *but for* or *sine qua non* condition is applied. According to this condition, *a* can be said to have caused *b*, if *a* "… is an antecedent but for which the result [*b*] would not have occurred."[1] In a situation where the event *a* represents a gunshot and the event *b* the death of a person, it is valid to say that *a* caused *b*, if it is true that the death (*b*) would not have occurred but for the gunshot (*a*). In short, the *but for* test can be stated as follows. 'But for *a* would *b* still have happened?' In order for *a* to be the cause of *b*, the answer to that question has to be negative. The criterion of the factual cause gives an answer to the question of *whether* a particular action *was* actually the cause of an event *or not*.

The criterion of a factual cause is restricted through a second criterion, namely the *legal cause* or proximate cause criterion. One major problem with the factual cause criterion is that for every event there are infinitely many causes, which pass the *but for* test. For example, one could say that *b*, the death, would not have happened but for *c*, the fact that a manufacturer produced the bullet, with which the gun was loaded. What distinguishes *c* from *a* is the degree of proximity to the event *b*. The criterion of the legal cause limits the number of *but for* causes for which it makes sense to hold people liable thus delimiting which of the potentially infinite number of causes should be *selected* as the legal cause In the experiment conducted for this study, we make it explicit to the subjects who should be considered relevant as a cause. In what follows, we will thus be less interested in the problem of causal selection (see for example Lagnado & Channon, forthcoming).

2.1.3 Psychology

The relationship between causality, responsibility and blame has been an extensive field of psychological research. However as we will see, researchers have mainly focused on the

[1] Modal Penal Code § 2.03(1), 1985

attribution of responsibility and blame on the level of individual causes. We would, first, like to offer a broad overview of two theoretical frameworks that have heavily influenced subsequent research in the field. Second, we briefly discuss a recent and controversial debate concerning the relationship between cognitive processes and moral judgment.

Shaver (1985) was first to develop a *normative* theory of blame attribution. His theory specifies the steps that a rational agent should follow when considering, whether somebody is to blame or not. First, he should assess the causal influence that the person had on the outcome. To what extent was the action of the person a necessary and/or sufficient cause? Second, the agent should evaluate the person's degree of responsibility which depends on various epistemological variables. Did the person perform the action *voluntarily* or was he coerced? Should he have *foreseen* the negative outcome that his action would lead to? Did he *intend* to bring about the outcome? Were there any *mitigating circumstances*? Finally, the assignment of blame should depend on whether the person is able to offer a justification or excuse for his action. Only if the latter is not accepted, a person should be blamed, according to Shaver's theory.

Alicke (2000) has developed a *descriptive* theory of blame assignment trying to capture the influence of peoples' motivations and expectations on their blame attributions. While Shaver's theory clearly specifies the order of assessment from causation to responsibility to blame, Alicke points out the possibility of a reverse process. Emotional reactions to a situation elicit spontaneous evaluations, which might activate the desire to blame someone for a negative event. Alicke calls this cognitive process, *blame-validation processing*. To justify an increased attribution of blame, people might exaggerate the control a person had over the outcome or assume that he should have foreseen the negative consequences of his behaviour. Furthermore, there are also non-motivational factors which trigger blame-validation processing. The well-known correspondence bias leads people to

attribute behaviour to the dispositional constitution of the actor, while neglecting situational characteristics (Gilbert & Malone, 1995).

Very recently the debate concerning the causal direction of the relationship between moral judgments, such as blame or praise, and the underlying cognitive processes, such as judgments about *causal relations* and *mental states*, has been revived. Solan (2003) is interested in explaining the ease with which people are able to blame other people. He argues that this is due to the fact that the cognitive processes involved in blaming are frequently used in our everyday life independently of moral judgments. Essentially, the processes involved in blaming are inferring causality, theorizing about other people's mental states and categorizing outcomes. Solan argues that these processes have originally emerged for the purpose of predicting and controlling our environment. However, when we blame other people, the exact same cognitive processes are used. Thus, the impulse to blame merely consists of a cognitively cheap secondary use of the processes that are almost already in constant use.

Knobe (2005), on the other hand, points out the possibility that our cognitive processes might have been shaped by our impulse to blame. He finds support for this controversial claim in the results of an experiment that he had conducted. In this experiment the exact same action leads to a positive or negative outcome – depending on the conditions of the experiment. The results show that people attribute intention to the actor when his decision led to the negative outcome, whereas they attribute no intention when the outcome was positive. In line with the possibility of an inverse connection between blame, responsibility and causality, people appear to use the valence of the outcome, which influences their impulse to blame, to infer whether an action was performed intentionally or not. Moral considerations, hence, appear to have an influence on people's folk psychology.

2.1.4 Artificial Intelligence

Researchers in artificial intelligence, mainly in the subfield of multiple agent systems, have begun to develop formal theories of responsibility and blame attribution. Mao and Gratch (2006) have developed a computational model of social causality and responsibility. They have basically adopted Shaver's (1985) framework of blame attribution and formalized the concepts involved. Their model possesses a knowledge base and is able to observe both behaviour and speech acts. Knowledge and observation together are used to infer causal and speech information. Depending on that information, attributions on aspects of agency, intention, foreknowledge and coercion are made. The attribution values in turn determine how much responsibility and blame is assigned to a particular agent. In short, their model is able to derive blame attributions from observing and knowing about social interactions. In an empirical test, their model was able to fit participant's ratings concerning blame assignment in a hypothetical scenario.

2.2 The Difference between Individual and Collective Responsibility

So far we have restricted our discussion of the relationship between causality, responsibility and blame to cases of individual causation. Most of the psychological literature has focused on problems of individual causation and assessed how the variation of a person's mental states, such as intention or foresight, influences people's attribution. Far less research has been conducted in the area that we are specifically interested in, that is, situations of collective responsibility, where multiple causes concur to determine the outcome. This part of the introduction will first review the philosophical literature and then discuss the problem of the legal definition of causation involving cases of multiple causation. It will culminate in a short overview over psychological research conducted so far.

2.2.1 Analytic vs. Holistic Approach

A group of people can be collectively responsible for something that they all intended but what they would not have been able to achieve on their own. A fundamental philosophical question is whether collective responsibility can be analyzed in terms of individual responsibility or whether in those situations new primitive principles emerge. We will refer to advocates who hold that collective responsibility is not analyzable into individual responsibility as *holists* and those who believe it is as *analysts*. Examples of collective moral responsibility in the philosophical literature include the collective responsibility of the Germans for the Holocaust, the responsibility of an organization for pollution, the responsibility of the industrial states for the starvation in the Third World countries and the collective responsibility of the whole population for the global warming effect.

Interestingly, the discussion between analysts and holists is akin to the now historical dispute between Structuralists and Gestalt psychologists. The latter group debated whether phenomena of visual perception can be explained by analyzing perceptions into their elementary parts or whether new phenomena on a higher level emerge that can not be explained from its lower level elements. Is the whole more or equal to the sum of its parts?

Gilbert (2006), a holist, argues that any group of persons jointly committed in some way create what she calls a plural subject. People are jointly committed to each other when they share the same intention. Plural subjects create a group view, which consists of attitudes that need to be primitively stated in a we-form. In the words of Thomas Hobbes, the people who form a plural subject are "reducing all their Wills … unto one Will" and thus create "a real Unitie of them all" (Hobbes, 1982/1651, p. 227). A plural subject is more than a mere aggregate of persons. Its collective responsibility is more than the sum of each person's individual responsibility. Collective responsibility, hence, is responsibility assigned to a collective as a single, independent entity.

Analysts do not deny the fact that groups have a very strong influence on people's behaviour on the individual level but they are convinced that individuals are the fundamental unit of analysis. They claim that moral questions always need to be asked on the level of an individual: what should a particular person do in a given situation (Sadler, 2006)? Additionally, Narveson (2002) argues that there are severe conceptual problems with the way collective responsibility is understood by holists. If collective responsibility really is non-analyzable, that is irreducible, then, this would preclude any individual member of a collective from being punished. The punishment should only be applied to the collective as a whole but not to its individuals.

Sverdlik (1986) tries to resolve the conflict by pointing out that when holists speak about collective responsibility they focus on the fact that more than one person can be responsible for a certain outcome. Analysts, on the other hand, focus on the fact that an individual person can only be made responsible for his or her own actions or intentions. He continues to claim that responsibility for action is the fundamental concept. What holists thus need to show is that a person can be held responsible for more than his individual action. Sverdlik argues that the key concept to solve this problem is intentionality. Since a group of persons can intend more than each individual would be capable of doing for himself, they can be collectively responsible for the result. However, each person is still only responsible for his or her individual actions.

Perhaps one of the most straightforward examples of collective responsibility is team sports. Reflected in the notorious statement of coaches that 'there is no I in TEAM', a team can only achieve a win as a collective. While each player in the team might share the intention to win and they are unified through their joint action, each player is, nevertheless, only responsible for his or her own contribution (Miller & Maleka, 2005). A theory, which takes individual responsibility as primitive can explain how multiple people can be responsible for an outcome.

2.2.2 The Problem of Distributing Responsibility

Even if we accept Sverdlik's (1986) attempt to resolve the conflict concerning the relationship between individual and collective responsibility, a fundamental problem remains. What implications does collective responsibility have for its group members? If a collective is responsible, does that imply that each member is responsible, that some members are or that, in fact, none of the members is individually responsible (Gilbert, 2006)? This problem, as already pointed out at the beginning, will be of main interest in this dissertation. It is nicely captured by a quotation from Edward, First Baron Thurlow. "Did you ever expect a corporation to have conscience, when it has no soul to be damned, and no body to be kicked?" (Coffee, 1981, p. 386)

Feinberg's (1968) typology of collective moral responsibility arrangements helps to address the problem systematically. He distinguishes four types of collective responsibility: 1) group liability without fault; 2) group liability with non-contributory fault; 3) contributory group fault, which is collective *and* distributive, and 4) contributory group fault, which is collective *but not* distributive. We shall only be interested in the third and fourth type.

The third type applies to situations where people have acted collectively and each individual's contribution is distinguishable. This type of collective responsibility applies to the three example situations mentioned at the beginning of this introduction. However, this does not imply that the assignment of responsibility in these cases is trivial. The contribution of each member in a group is normally not identical. Apart from that, multiple causes might create diverse situations, differing in the underlying causal structure. Particular actions might be necessary and combined actions sufficient to bring about the result. So even if the contribution of each member in a group is observable and distinguishable the attribution of responsibility and blame and the assignment of punishment remain substantial problems.

The fourth type applies to situations where people have acted collectively but each individual's contribution is not easily identifiable. Most team sports are an example of this

type of collective responsibility. In a football team, for example, each player has a particular role and a certain responsibility assigned to that role. The goal-keeper's responsibility is to keep his box clean whereas the responsibility of a striker is to score goals. The credit or blame for the win or loss is not that easily attributable to each individual player. Should each player receive the full responsibility for the result? Is no individual player to be held responsible for the result? Should players be held responsible according to the role they played or proportional to their causal contribution which as has just been mentioned, is hard to assess? Or, finally, should players be held responsible according to the role they could have played?

The same problem occurs in a class of situations that has received much attention in philosophical papers dealing with causality, namely situations of overdetermination. A common scenario is that of a firing squad.

> *A firing squad, consisting of four marksmen, simultaneously shoots at a prisoner resulting in his death. Each of the marksmen shot the prisoner directly in the head. Each shot on its own would have been sufficient to kill the prisoner.*

Because of the existence of four independent causes, which would each have been sufficient on their own to bring about the effect, the effect is causally overdetermined. How should the responsibility for the death of the prisoner be divided amongst the four marksmen?

At first glance it seems plausible to say that a person should only be held responsible for something if her behaviour made a difference to what happened. If the same would have happened, no matter what she did, she should not be held responsible. The attentive reader has already detected the similarity of this consideration to the criterion of the *factual cause* applied by the law as discussed above. From a philosophical perspective, the *but for* condition is essentially equivalent to the adoption of a counterfactual theory of causation (Lewis, 1973; Mackie, 1974). According to the counterfactual theory of causation an event *a* is the cause of

an event *b*, if and only if the following two conditions are true. First, *a* occurred and *b* occurred and, second, if *a* had not occurred then *b* would not have occurred. The second part of the definition gives the counterfactual theory of causation its name. It relies on considering a situation that did in fact not occur.

Let us now apply this intuitively plausible definition to the firing squad example. Let us denote the shots of the four marksmen with $a_1 - a_4$ and the death of the prisoner with *b*. We can now ask the question, whether a_1, the shot of the first marksmen caused *b*, the death of the prisoner by checking both conditions of the counterfactual definition of causation. The first condition is valid since a_1 occurred and *b* occurred. The second condition, however, is invalid. If a_1 had not occurred, *b* would *still* have occurred. According to this definition, then, a_1 does not make a difference and was thus not the cause of *b*. The same, of course, also applies to a_2, a_3 and a_4. We thus arrive at the very counterintuitive conclusion that neither of the events $a_1 - a_4$ caused *b* further implying that none of the marksmen should be held responsible for the death of the prisoner. We can see that the *but for* test appears to have severe limitations in cases of overdetermination.

Different propositions have been made to solve this problem. Some have argued that, since a death of a person is by no means divisible, each individual is responsible for *all* the wrongdoing of the collective (French, 1984). This policy has also been embraced by the law in several legal cases. Others have argued that although the death of a person is not divisible, the responsibility of several people for the death can be. Cohen (1981) has pointed out that 100 people could each be responsible for one hundredth of a death.

More recently, researchers in artificial intelligence have offered formal solutions to the attribution of responsibility in cases of overdetermination. Turrini, Paolucci and Coate (2006) have formalized responsibility as the power to prevent a harmful state of the world from being brought about. They distinguish between shared responsibility, where each member of the group has the potential to avoid the damage versus collective responsibility, where only a

combination of all members' actions can avoid the damage being done. The power to prevent damage is seen as a microfoundation of the notion of responsibility. In the firing squad example, the individual marksmen are clearly interdependent towards the avoidance of the death.

Chockler and Halpern (2003) have defended the idea that there is a direct relationship between causality and responsibility. They have developed a model which is specifically suited for cases of overdetermination. Their model is cast in the framework of causal models which uses structural-equation modelling to capture the counterfactual dependencies between sets of events (Pearl, 2000; Woodward, 2003). Their theory basically offers a relaxed criterion of counterfactual dependence. Accordingly, an event a_i is a cause of an event b if and only if there is some contingency under which b counterfactually depends on a_i. In other words, if there is a situation where the manipulation of a_i, given that the set of other causes in the model is fixed to some value, results in a change in b, then a_i is a cause of b. According to this definition, the concept of causality is discrete. Something can either be a cause or not. From this modified counterfactual definition of causality they derive their definition of responsibility which is a continuous concept. The degree of responsibility of a cause a_i for an effect b is determined by the equation $\mathrm{Resp}(a_i) = \frac{1}{N+1}$. In this formula, N denotes the *minimal number of changes* that need to be made to obtain a situation where b counterfactually depends on a_i. We can now apply this formula to the firing squad example. How many changes would one need to make to obtain a situation where the death of the prisoner depends on the shot of, for example, the first marksman? One would need to make three changes. In fact they all shot. If one could change the marksmen 2, 3 and 4 from having shot to not having shot then the death of the prisoner would indeed counterfactually depend on marksman 1. Only if marksman 1 shoots in this situation the prisoner dies. The model would thus predict that marksman 1 should be assigned a responsibility of $\mathrm{Resp}(\mathrm{marksman}_1)$

$= \frac{1}{3+1} = \frac{1}{4}$. The same does, of course, apply to the other three marksmen. In short, the model is able to assign a degree of responsibility to individual causes in situations of collective responsibility, where multiple sufficient causes overdetermine a given effect.

2.2.3 Psychological Theories to the Present

Psychological research which assesses the relationship between causality and responsibility in cases of collective responsibility is very sparse. However, a couple of papers have assessed the effects that different sorts of causes, such as necessary, sufficient or indirect causes, have on people's responsibility and liability ratings (Greene & Darley, 1998; Solan & Darley, 2001). Studies have also looked at, whether people's causality ratings are influenced by what could have happened had a person acted differently (Wells & Gavanski, 1989). They have found that if the outcome would have been the same, no matter what action a person had taken, the person is perceived less responsible for the outcome. This effect is present although in both scenarios the person's behaviour *and* mental states were identical.

To the knowledge of the authors, there is only one study which explicitly looks at cases of causal overdetermination and their influence on attribution. Spellman and Kincannon (2001) presented participants of their study with the example of a firing squad with two marksmen. In one scenario both shots were said to be each sufficient for the death of the prisoner. In the other scenario participants were told that each shot was necessary and only the two shots together were sufficient for the death of the prisoner. Although they did not test Chockler and Halpern's (2003) model, it would make clear predictions in this scenario. In the scenario where both are sufficient each should get a responsibility of 1/2. In the scenario where they are both necessary and together sufficient, each of them should get a responsibility of 1. However, contrary to the predictions of the model, Spellman and Kincannon found that participants rated the individual causes in the multiple sufficient condition as more causal than in the multiple necessary condition. They conclude that it is not counterfactual dependence,

which constitutes the main factor of people's perception of causality but rather a notion of causal contribution which they operationalize in probabilistic terms.

Those results are not entirely convincing as participants' intuitions in these artificial and abstract scenarios might not be well calibrated. We thus designed an experiment that we think provides a much more rigorous test of Chockler and Halpern's model's predictions. We were, in part, motivated by the fact that we are not aware of any empirical test of the model's predictions so far. In short, our aim is to find an answer to the following research question.

Research question 1: How do people distribute responsibility amongst multiple causes?

3 Experiment 1

In order to approach our research question, we designed and experimental game, which is called the "Triangle Game" (TG). The TG is played on the computer. The participant's task is to count triangles presented in complex diagrams for a brief period of time. Participants are instructed that they are not playing the game individually but in a group together with three computer players. Whether a particular round in the game is won or lost depends on the accuracy of each player in the group.

Figure 1 shows a diagram that is used as an example stimulus in the instructions. Participants are advised that they should only count the small blue triangles and larger triangles that are made up of blue triangles. The green marked triangles represent valid examples, whereas the red marked triangles would be invalid.

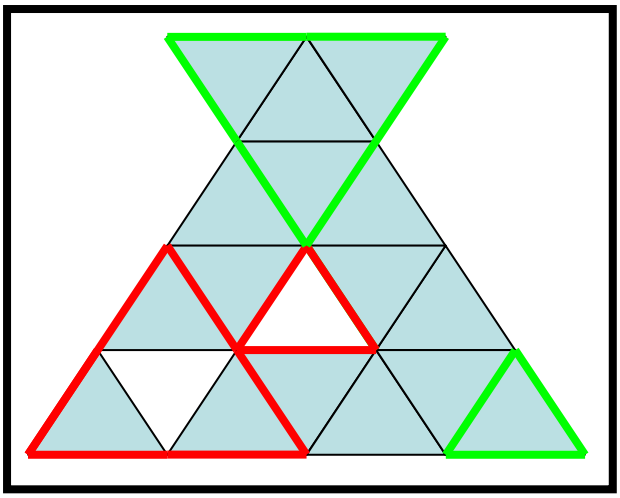


Figure 1

Each round of the game consists of two consecutive steps. In the first step, participants are shown the diagram for 12 seconds (see Figure 2). They have to count the triangles present in the diagram and type in their answer. They then get shown the correct solution and are informed about whether their group won or lost the round. The amount of points the team possesses is changed according to whether they have won or lost. Participants are instructed at the beginning that the main aim of the game is to get as many points as possible.

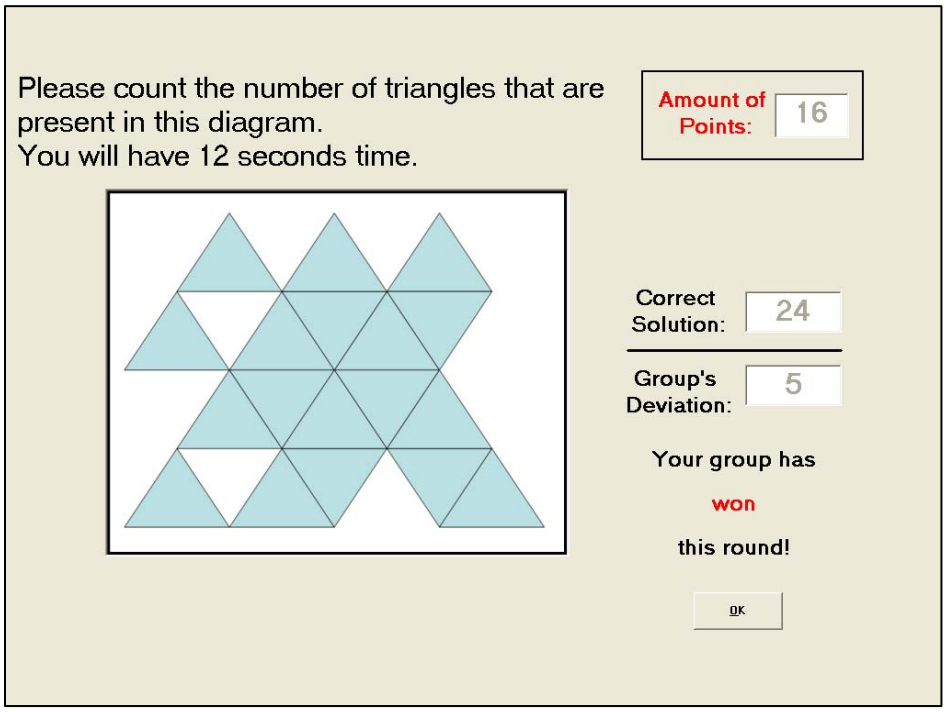


Figure 2

In the second step, participants view a table, which contains the answers of each player in their group including their own answer (see Figure 3). Furthermore, the correct solution is displayed and participants can see for each player how much his or her answer deviated from the correct solution. For each player there is a slider, which participants can use to determine how much they think that player was responsible for the group's loss or win in that particular round. The scale ranges from 0 (not responsible at all) to 10 (very responsible). Participants are also reminded that in the case of winning, high responsibility values indicate that they think that a person should get high credit for his or her answer, whereas in the case of losing, high responsibility values indicate high blame.

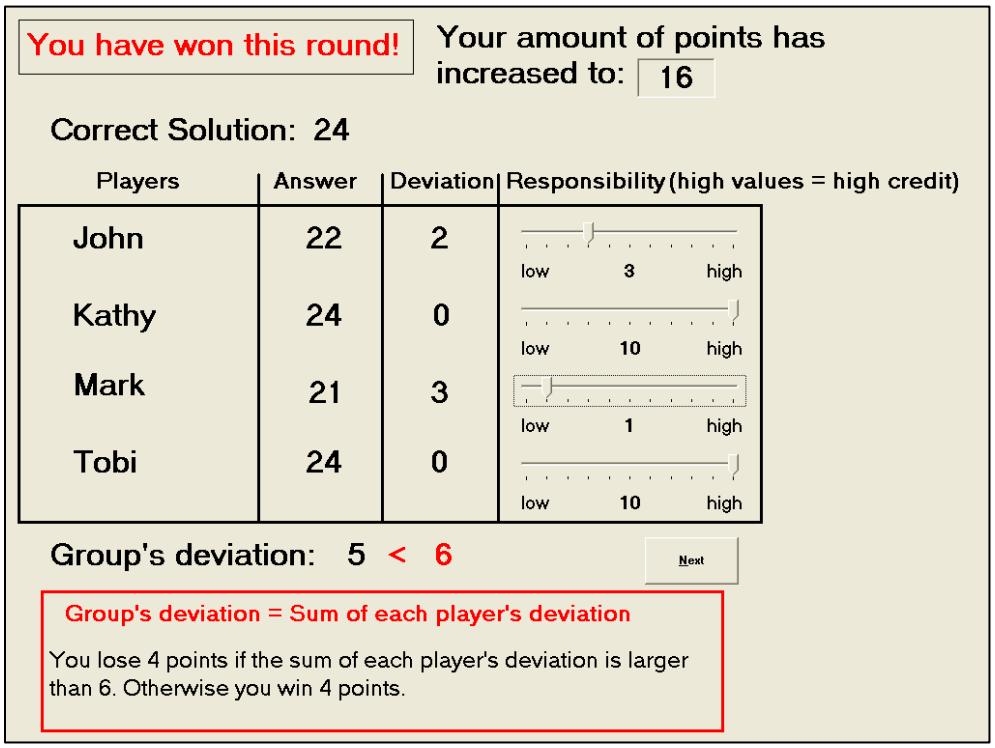


Figure 3

The experiment has three different experimental conditions. The only way in which those differ from each other is in how the deviation of the whole group is calculated from the deviation of each individual player. The deviation of the group determines whether a particular round is lost or won.

In the first condition, the group's deviation equals *the sum* of each player's individual deviation from the correct solution. If this sum is equal to or less than 6 the group wins the

round, otherwise it loses. In the second condition, the group's deviation equals the deviation of the *least accurate* player in that round. If his or her answer's deviation from the correct solution is equal to or less than 2, the group wins, otherwise it loses. In the third condition, the group's deviation equals the deviation of the *most accurate* player. If this player gives the correct solution, that is, his or her deviation is 0, then the group wins that round, otherwise it loses. Two example situations will help to clarify the differences between the three experimental conditions.

Table 1

Player	Deviation
John	0
Kathy	1
Mark	3
Toby	2

Figure 4

Table 2

	Condition 1 (= sum)	Condition 2 (= least)	Condition 3 (= most)
Group Dev	6	3	0
Outcome	Win	Loss	Win

Table 1 shows the deviation of each player's answer from the correct solution. Figure 4 is a reminder of how the group's deviation is calculated and the critical value to which it is compared for the three conditions. Table 2 shows the result that this pattern of deviation would have led to in the different conditions.

In the first condition, the group would have won the round because the group's deviation is determined by the sum of each player's individual deviation. This sum equals 6, which is equal to the critical comparison value. In the second condition, the group's deviation equals the deviation of the least accurate player. In this round, Mark is the least accurate player with a deviation of 3. The group would have lost this round because Mark's deviation exceeds the comparison value of 2. In the third condition, the group's deviation equals the

deviation of the most accurate player. John is the most accurate player in this particular round. Because he was able to give the correct solution the group would have won the round.

Table 3

Player	Deviation
John	2
Kathy	1
Mark	2
Toby	2

Figure 5

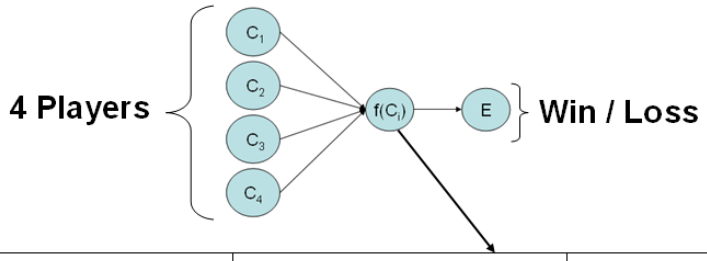
Table 4

	Condition 1 (= sum)	Condition 2 (= least)	Condition 3 (= most)
Group Dev	7	2	1
Outcome	Loss	Win	Loss

In this second example situation, the outcome in each round is exactly opposite to the one in the first example. This is to show that the outcomes in the different conditions are largely independent from each other. Hopefully, the reader is now able to see why the pattern of deviation in Table 3 would have led to the particular outcomes in the different conditions (Table 4).

It is important to note that the general underlying causal structure is identical in each condition. There are always four causes with continuous values, which are determined by the deviation of each individual player. The outcome is always binary in that the group can either win or lose a round. What differs between the experimental conditions is the way in which the contribution of the individual causes is combined to determine the outcome. Table 5 gives an overview over the different forms that the integration function can take on depending both on the experimental condition and on whether a round was won or lost. In designing the experimental conditions, we drew on Steiner's (1973) analysis of group performance on unitary tasks.

Table 5



	Win	Loss
Condition 1 (= sum)	Additive	Additive
Condition 2 (= least)	Conjunctive (AND)	Disjunctive (OR)
Condition 3 (= most)	Disjunctive (OR)	Conjunctive (AND)

For the first condition, the integration function is additive both for wins and for losses. Each player's deviation always adds up to determine the group's deviation. This is similar to the example of the penalty shoot-out mentioned at the beginning.

For the second condition the integration function is conjunctive for winning and disjunctive for losing. In order for the group to win a round the answers of all players need to deviate less than three from the correct solution. The causes are thus combined according to an AND gate. In the case of losing it is sufficient if a single player's answer deviates more than two – an OR gate combination. This causal structure is similar to the house building example. The house can only be built on time if all suppliers deliver their materials on time. The delay of one supplier is sufficient to cause a delay in the overall building time of the house.

The nature of the integration function in the third condition is exactly opposite to the one in the second condition. In the case of winning the integration function is disjunctive, whereas for losses it is conjunctive. The group wins if at least one player gives the correct solution. It loses only if no player was able to get it right. This should remind the reader of the bouncer example, where you and your friends tried to get access to the party. You would only be allowed access if one of you knew the host whereas you would be refused entrance if nobody of your group knew him.

We are convinced that there are several other situations of multiple causation in the real world which would fit one of the three underlying causal structures embedded in our experimental conditions. Our experimental paradigm enables us not only to ask how participants generally distribute responsibility but also to ask a second, more refined question.

Research question 2: Does the way the causes are combined influence the attribution of responsibility?

3.1 Methods

3.1.1 Participants

32 participants took part in the experiment. The sample consisted of 8 women (25 %) and 24 men (75 %). The mean age was $\bar{x} = 26$ ($\sigma = 7$), with a range from 18 to 50. There were 10 participants in the first, 12 participants in the second and 10 participants in the third experimental condition. Most of the participants were friends of the author and were contacted via E-mail. They were advised to follow a link to download the program, run it and send the data file back to the author. 40 out of 83 friends replied, which equals a return rate of 48.2%. 8 cases were not included in the analysis because it was apparent from their results that they did not understand the task correctly. This might have been mainly due to an insufficient understanding of the instructions. The language used in the TG was English and most participants were German.

3.1.2 Instruments and Materials

The TG was programmed by the author with Microsoft Visual Basic 6.0. Participants ran the TG on their individual computers. The diagrams were designed by the author and the correct solution for each diagram was double checked.

3.1.3 Design

The study was run as a 3-factor between-subjects design. Each participant was randomly assigned to one of the three experimental conditions described above.

3.1.4 Procedure

Participants were instructed that the experiment will take about 10 minutes and that they should not stop the game once started. Participants were able to press a cancel button if they preferred to play the game at a later stage. They were then introduced to the nature of the task. The main characteristics of the game were identical for each of the three conditions. The aim of the game in each condition is to maximize the amount of points for the group. The group starts with an initial amount of 16 points and for each round that is won or lost the amount of points is increased or decreased by 4, respectively. Furthermore, participants in each condition play together with the same computer players, whose answers were always the same.

The instructions in each condition were identical except for how the deviation of the group is calculated from the individual player's deviation. Participants were able to remind themselves of how the group's deviation is calculated throughout the game. The rule was shown in a box at the bottom of the screen, when participants assigned responsibility (see Figure 3).

The flow chart in Figure 6 shows the general procedure of the game. After being presented with the instructions, participants play a practice round to become familiar with the structure of the game. During the practice round, little notes pop up helping to explain the different components of each screen. After they have finished the practice round, they start with the real game. Each round consists of two steps: first, the triangle count and then the responsibility attribution as mentioned above.[2] After the tenth round the game stops and participants are shown a final screen where they are informed about the amount of points their

[2] Table 13 and Table 14 in the appendix (9.3) show how well the computer players performed overall. Table 15 and 16 in the appendix (9.4) show participants' performance for different diagrams.

team has achieved. They are also reminded to send back the data file. Additionally, participants have the opportunity to put down any questions or comments concerning the game in a text field. They are told that any questions will be answered as quickly as possible.

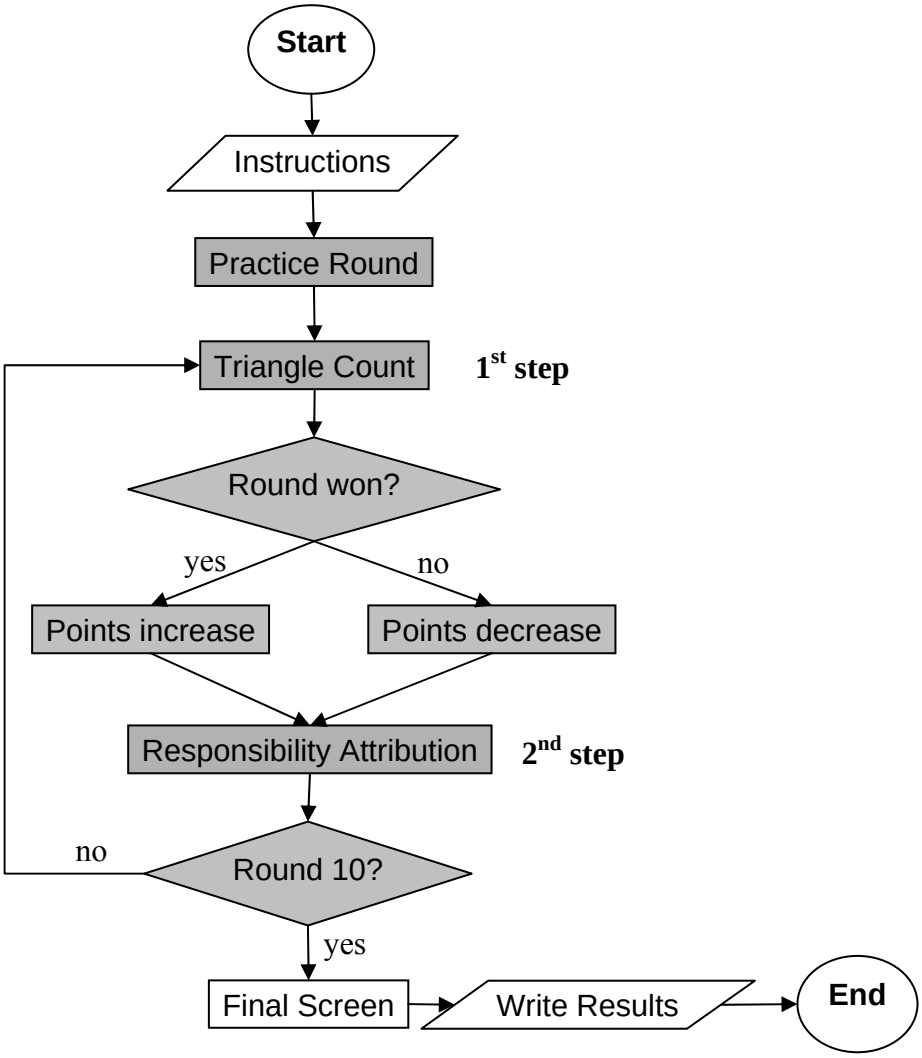


Figure 6

3.2 Results

3.2.1 Responsibility Ratings

First, we would like to address the second research question. Does the way the causes are combined influence the responsibility ratings? In other words, do the different experimental conditions have an influence on participants' responsibility ratings? In order to answer this

question, we will compare the responsibility ratings between the three experimental conditions both for losses and wins.

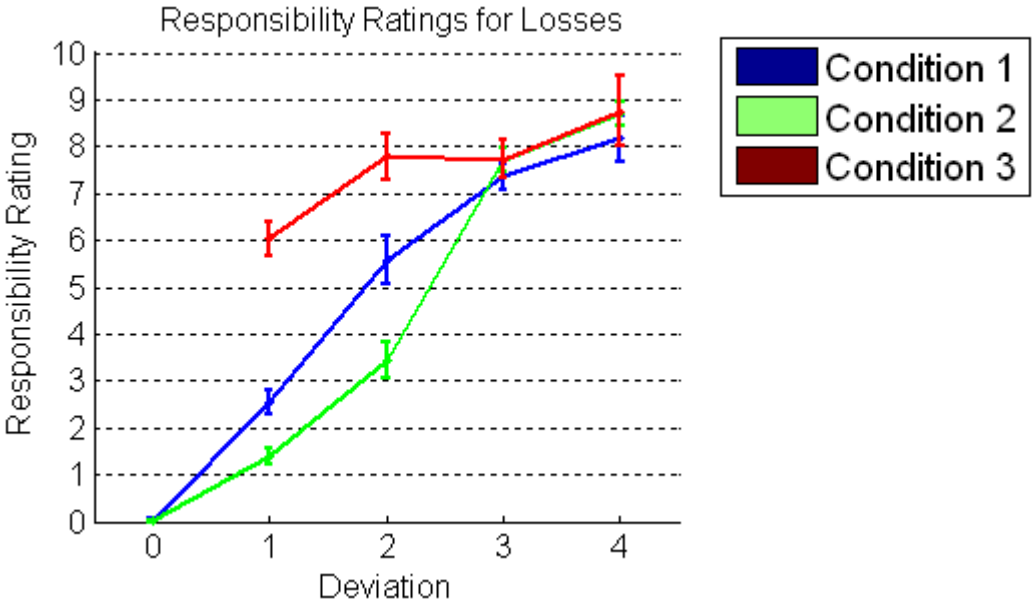


Figure 7

Figure 7 shows the mean responsibility that has been assigned to a player depending on his or her deviation for rounds that the group has lost.[3] The three different lines represent the three experimental conditions.

In the first condition, the mean responsibility ratings increase gradually with an increased deviation. The more a player deviated the more responsibility was assigned to that player. In the second condition, the mean responsibility ratings for a deviation of 1 and 2 are both significantly lower than in the first condition with $t(118) = 4.11$, $p < .001$ and $t(61) = 2.98$, $p < .01$, respectively.[4] Furthermore, there is a big difference between the mean responsibility rating for a deviation of 2 and 3 within condition 2 with $t(157) = -9.32$, $p < .001$. In the third condition, there was of course no case where a player deviated 0 and the group lost the round. This explains why the line starts at a deviation of 1. The mean responsibility for a deviation of 1 and 2 are significantly higher than in condition 1 with $t(113) = -6.84$, p

[3] A different and more detailed way of presenting the data is shown in the appendix (9.1). However, due to the simplicity of the line graphs, we preferred to show them here.

[4] Independent t-tests have been calculated on comparisons of interest. Generally, if the standard error bars of two means do not overlap, the means differ significantly.

< .001 and $t(34) = -3.09$, $p < .01$. For a deviation of 3 and 4 the means of the three conditions do not differ significantly.

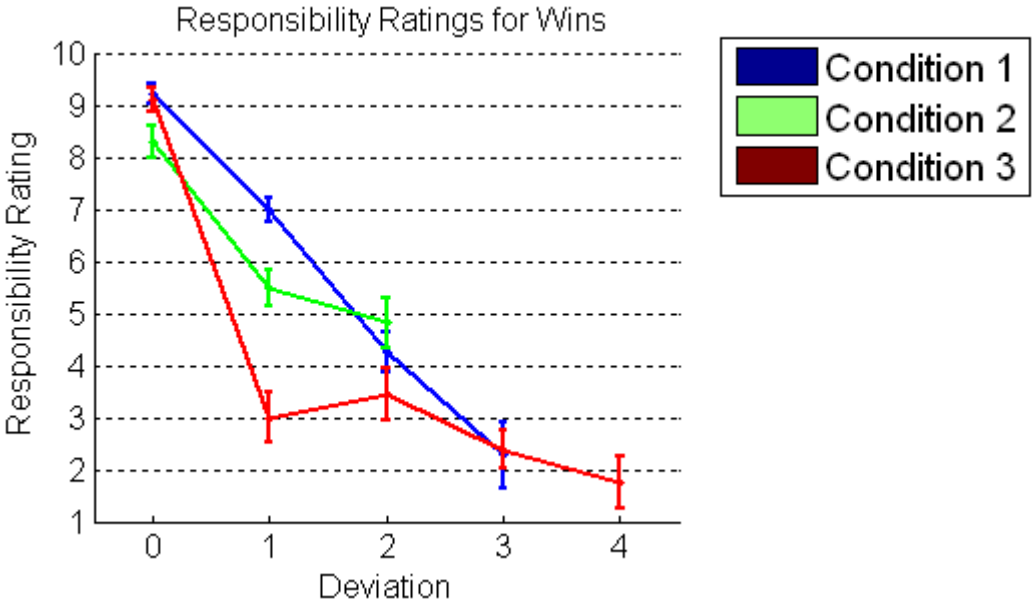


Figure 8

Figure 8 shows the mean responsibility ratings for rounds that the group has won. In the first condition, again, the responsibility ratings decrease gradually relative to an increased deviation. In the second condition, there is no case where a player deviated 3 or 4 and the group won. For a deviation of 1, the mean responsibility rating is significantly lower than in condition 1 with $t(177) = 3.71$, $p < .001$. In the third condition, there is a significant difference between the mean responsibility rating for a deviation of 0 and for a deviation of 1 with $t(164) = 12.57$, $p < .001$. The means for a deviation of 1 and 2, 2 and 3, and 3 and 4 do all not differ significantly. The mean responsibility rating for a deviation of 1 is significantly lower than the mean in condition 2 with $t(142) = 4.35$, $p < .001$.

3.2.2 Model Predictions

We can now try to find an answer to the first research question. How do people allocate responsibility amongst multiple causes? Can we find a cognitive model that explains some of the variance of the participants' responsibility ratings?

3.2.2.1 Matching Model

Participants could use a matching heuristic. When thinking about how much responsibility they should assign to a particular player they could directly match the deviation of a player to his or her responsibility rating. If a round was won, that means that a player should get a responsibility of 10 minus his or her deviation. If a round was lost, one can directly match the deviation to the responsibility rating. Table 6 shows the predictions this model would make for a particular situation.

Table 6

		Condition 1 (= sum)	**Condition 2 (= least)**	**Condition 3 (= most)**
Player	**Dev**	**Win**	**Loss**	**Win**
John	2	8	2	8
Kathy	3	7	3	7
Mark	1	9	1	9
Toby	0	10	0	10

Win if
1. sum ≤ 6
2. least ≤ 2
3. most = 0

3.2.2.2 Basic Counterfactual Model

To decide how much responsibility each player should bear, participants could also apply the kind of counterfactual thinking that is adopted by the law. A person should thus only be made responsible for the outcome of a particular round if the outcome was counterfactually dependent on his answer. If changing his answer would have made no difference to the outcome then he was not the cause and therefore should bear no responsibility for what happened. In short, a player should get a 10 responsibility rating if he could have changed the outcome by changing his answer and a 0 responsibility if he could not.

Table 7

		Condition 1 (= sum)	Condition 2 (= least)	Condition 3 (= most)
Player	**Dev**	**Win**	**Loss**	**Win**
John	2	10	0	0
Kathy	3	10	10	0
Mark	1	10	0	0
Toby	0	10	0	10

Win if
1. sum ≤ 6
2. least ≤ 2
3. most = 0

Table 7 shows the predictions the basic counterfactual model would make for the same situation as discussed above. In the first condition, each player should get a 10 responsibility because each player could have changed the outcome by changing his answer. If only one player had deviated a bit more, then the whole group would have lost that round. In the second condition, it is only Kathy who can make a difference to the outcome by changing her answer. The other players could not have changed the outcome of that round, no matter what answers they had given. Since the outcome is not dependent on their answers they should bear no responsibility for the loss. In the third condition, Toby made the group win by giving the correct solution. If he had not given the correct solution, the group would have lost.

Table 8

		Condition 1 (= sum)	Condition 2 (= least)	Condition 3 (= most)
Player	**Dev**	**Loss**	**Loss**	**Loss**
John	3	0	0	10
Kathy	5	0	0	10
Mark	3	0	0	10
Toby	2	0	0	10

Win if
1. sum ≤ 6
2. least ≤ 2
3. most = 0

As discussed above, the basic counterfactual model has problems dealing with cases of overdetermination. These cases occur quite naturally within the TG. Table 8 shows a different pattern of deviation which quite likely could have occurred for a very difficult diagram. For the first condition, the deviation of the group in that round is 13. There is no player in the

group who could have changed the outcome of that round by changing only his or her answer. Thus, the basic counterfactual model predicts that none of the players caused the loss. They should all bear no responsibility. The same is true for the loss, which would have occurred in the second condition. The loss is overdetermined by the answers of John, Kathy and Mark who each deviated more than 2 from the correct solution. Each of them is not capable of changing the outcome themselves by changing their answer and they are, hence, not responsible for the loss. In condition three, however, each player would have been responsible for the loss. Each of them could have changed the outcome to a win, if they had had a 0 deviation.

3.2.2.3 Modified Counterfactual Model

It is exactly these cases of overdetermination that the modified counterfactual model by Chockler and Halpern (2003) is capable to deal with. Table 9 shows the predictions that this model would make for the situation just described, where the loss in both the first and the second condition is overdetermined.

Table 9

		Condition 1 (= sum)	**Condition 2 (= least)**	**Condition 3 (= most)**
Player	**Dev**	**Loss**	**Loss**	**Loss**
John	3	1/2 (i.e. **5**)	1/3	1
Kathy	5	1/2	1/3	1
Mark	3	1/2	1/3	1
Toby	2	1/2	0	1

Win if
1. sum ≤ 6
2. least ≤ 2
3. most = 0

The formula, which is used to determine how much responsibility each player should bear is $\text{Resp}(C_i) = \frac{1}{N+1}$, where N denotes the minimal number of changes that have to be made to obtain a situation where the result counterfactually depends on the answer of a particular player C_i. For the first condition, how many changes in the answers of the other player would

have been needed so that the outcome would have been dependent on, for example, John's answer? One change would have been sufficient. If Kathy had given the correct solution, thus decreasing the sum of deviation from 13 to 8, then John could have changed the outcome had he given a different answer, namely a deviation of 1 or less. John should therefore get a responsibility of 1/2 or a 5 according to our scale, because one change is needed. The same holds true for each of the other three players. Changing the answer of one other player is always sufficient to establish a counterfactual dependence. In the second condition there are two changes needed to make the outcome counterfactually dependent on John. Only in a situation where the answers of both Kathy and Mark are changed to a deviation of 2 or less, is the outcome counterfactually dependent on John's answer. Because two changes are needed, the formula predicts that John should get a responsibility of 1/3 for the loss. Toby has no responsibility for the loss in this situation, because he could not have made a difference to the outcome by changing his answer. In the third condition, the predictions are identical to the basic counterfactual model. Each player could have changed the outcome and thus they should all get a responsibility of 1, or a 10 on the scale that has been used in the game.

3.2.3 Model Comparison

Now that we have seen what specific predictions these three models make, we can evaluate whether they can predict the empirical responsibility ratings that were given by the participants. Figure 9 shows how well the three afore-mentioned models and a comparison model, which essentially gives random responsibility ratings, fit the data.

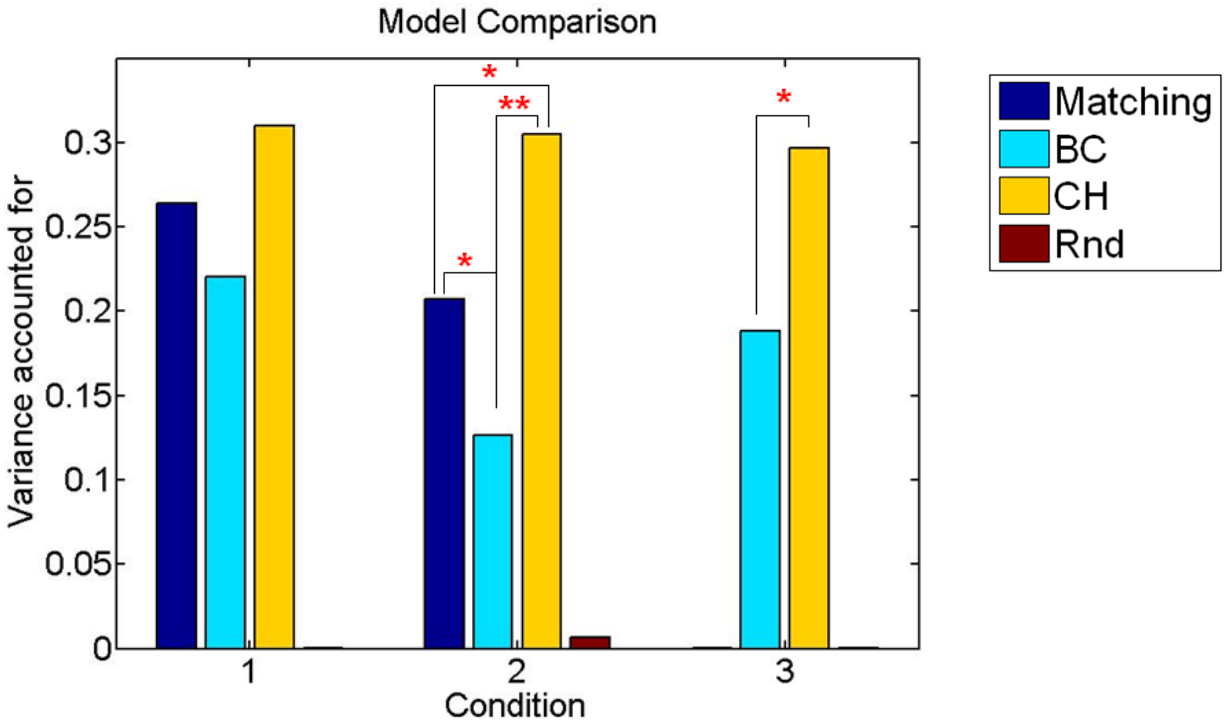


Figure 9

The different bars represent the different models. The height of the bars indicates how much variance of the empirical data is accounted for by the models, that is, the squared correlation between the values that the model predicts and the empirical ratings.[5] To calculate the model fit, a linear regression model with two parameters was used. The first parameter represents the intercept and the second one the slope of the straight line. None of the models contain any free parameters. The model fits are shown for each experimental condition, represented by the 3 different columns on the x-axis. One star indicates that the fit between two models is significantly different on the $\alpha = 5\%$ level and two stars indicate that two models differ on the $\alpha = 1\%$ level.[6]

In the first condition, the fits of the matching model, the basic counterfactual (BC) model and the modified counterfactual model by Chockler and Halpern (CH) do not differ significantly. In the second condition, the CH model is significantly better than both the

[5] The correlation coefficients between the predicted values and the empirical ratings are shown for each model and condition in the appendix (9.2).

[6] Simple Interactive Statistical Analysis (SISA) [http://www.quantitativeskills.com/sisa/], an online statistics program, has been used to calculate whether the model fits differ significantly. An exemplar of the output of the program can be found in the appendix (9.5).

matching model and the BC model. In the third condition, the matching model does not account for any of the variance. The CH model fits the data significantly better than the BC model. The random model does not account for any of the variance in the three conditions.

3.3 Discussion

The second research question, which asked whether the way the causes are combined influences the allocation of responsibility, can be answered with a clear yes. If the underlying causal structure had no influence on participant's responsibility ratings then the line graphs for each condition should have been identical. However, there are significant differences in the mean responsibility ratings between the three experimental conditions for both wins and losses. Most participants are sensitive to the underlying causal structure and the nature of the integration function.

Regarding the first research question of how people distribute responsibility amongst multiple causes, we have seen that their responsibility attributions are best described by the CH model. The CH model fits the empirical responsibility ratings equally well in all three experimental conditions. A very important question, however, remains unanswered. Should the CH model be interpreted as an as-if model or as a process model (Woodward, forthcoming)? That is, should the model be regarded merely as a good description of the empirical data or does it go even further, implying that participants (implicitly) run mental simulations on their causal representations in order to determine how to allocate the responsibility? Do participants think about different answers the players might have given and whether that would have resulted in a different outcome?

Findings of previous studies have suggested that the simulation of potential outcomes and causal attribution are potentially dissociable and should, hence, be viewed as independent separate processes (N'Gbala & Branscombe, 1995; Mandel, 2003). However, these findings are all based on participant's attributions in hypothetical scenarios. Do they also hold in a

more natural, game-like situation? In order to make progress answering this important question we have conducted a second experiment.

4 Experiment 2

The second experiment is largely equivalent to the first one, except for the fact that each round of the game now contains a third step. After participants have assigned responsibility to each player in their group, they are shown the following screen (see Figure 10).

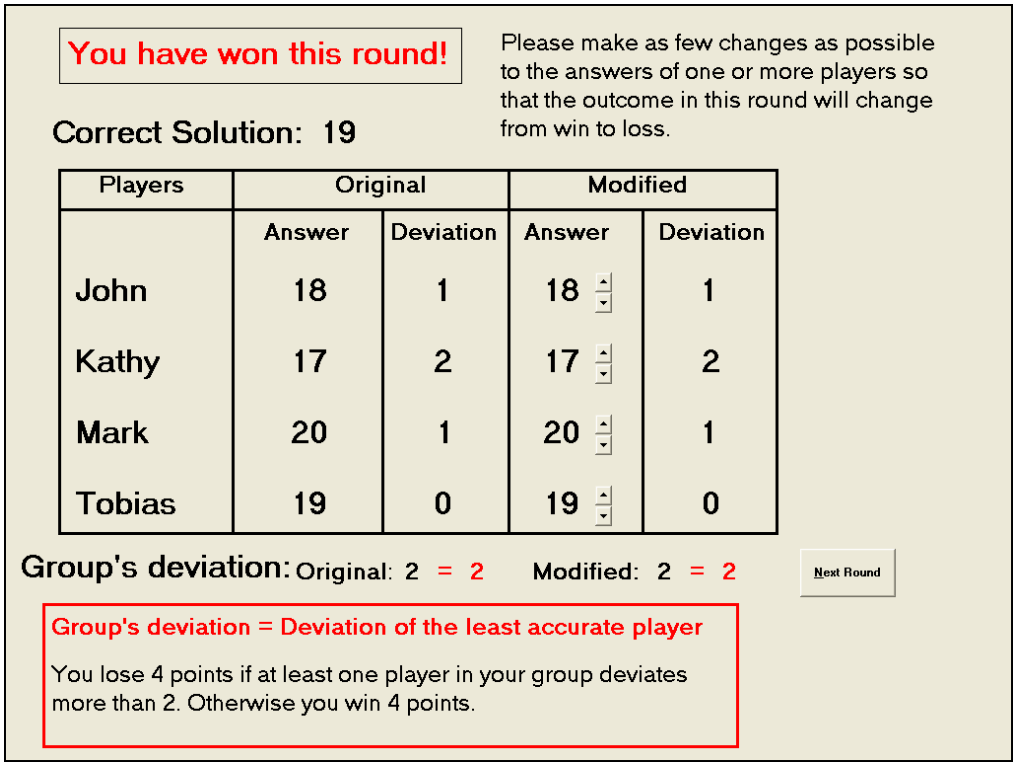


Figure 10

On that screen they see the original answers that each player has given and the value of the original group deviation. Participant's task is, then, to make as few changes as possible to the answers of one or more players, so that the result in that round would have been different. In order to do so, participants can use the arrows facing upwards and downwards next to the answers of each player in the 'Modified' column of the table. The amount of deviation then changes accordingly. When participants have changed the answers so that the result for this

round changed, they are informed that they have successfully changed the outcome (see Figure 11).

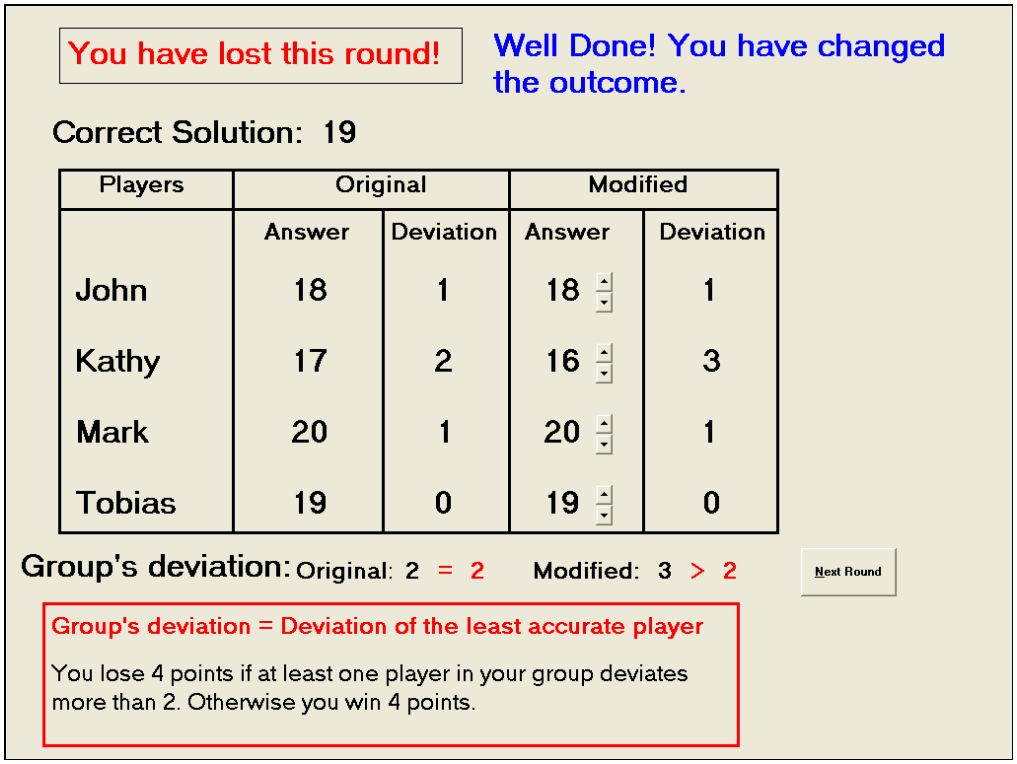


Figure 11

There are a couple of reasons why we have conducted this second experiment. First of all, we wanted to replicate the findings of the first experiment with a different population of subjects. Second, this manipulation made sure that each participant understood the rule of how the deviation of the group is calculated. Although we have seen in the first experiment that the different conditions had a substantial influence on people's responsibility ratings, we did not explicitly test whether participants had really understood how the group's deviation is calculated from each player's individual deviation. Third, we were interested in assessing how much the attribution of responsibility correlated with participant's changes of the player's answers. Intuitively, one would think that the more a participant thought a person was responsible for the outcome, the more this person's answer should be altered to change the outcome. Fourth, we were interested in whether this manipulation would prompt participants' counterfactual thinking. Will participants think more about what could have happened differently, had some players just given different answers?

Apart from these questions we had one more major interest. How would participants understand the notion of a minimal change? This notion is of particular importance for many contemporary counterfactual theories of causation. Lewis' (1973) theory, for example, is cast in a possible world semantics for counterfactuals. Possible worlds are worlds with maximally consistent propositions that could have come into existence had a particular event occurred differently. The central concept in Lewis' semantics is the concept of comparative similarity between worlds. Possible worlds can be weakly ordered according to their similarity to the actual world. One world is closer to actuality if it resembles the actual world more than the second does.

The possible world semantics can be used to analyze the truth of counterfactual claims in the following way. A counterfactual statement is true if "… it takes less of a departure from actuality to make the antecedent true along with the consequent than to make the antecedent true without the consequent." (Menzies, 2001) The counterfactual that a prisoner would not have died had a marksman not shot is valid if a situation where both the shot and the death occurred is closer to our actual world than a situation where the shot occurred but the death did not occur. In this context, Lewis also speaks of miracles that have to be made to transfer a possible world into the actual world. The third step of the TG enables us to ask, whether participants use one large miracle or several small miracles to transfer the actual world into a possible one. In terms of the game, will they make a large change to the answer of just one player or will they make several small changes to the answers of more than one player?

4.1 Method

4.1.1 Participants

37 participants took part in the second experiment. The sample consisted of 18 women (48.6 %) and 19 men (51.4 %). There were 11 participants in the first, 11 participants in the second and 15 participants in the third experimental condition. The mean age was $\bar{x} = 24.8$ ($\sigma = 5.3$),

with a range from 20 to 51. Participants were contacted via the subject pool of University College London. Each participant was entered into a prize draw. Three prizes could be won, whereby the first prize was £100, the second one £30 and the third one £20. Six participants had to be excluded from the sample which initially consisted of 43 participants.

4.1.2 Instruments and Materials

The instruments and materials were identical to the first experiment.

4.1.3 Design

The experimental design was identical to the first experiment.

4.1.4 Procedure

The procedure was identical to the first experiment except for the fact that each round now consisted of *three* consecutive steps. Participants were instructed that the game will take about 15 minutes. Figure 12 shows the modified flowchart of the game.

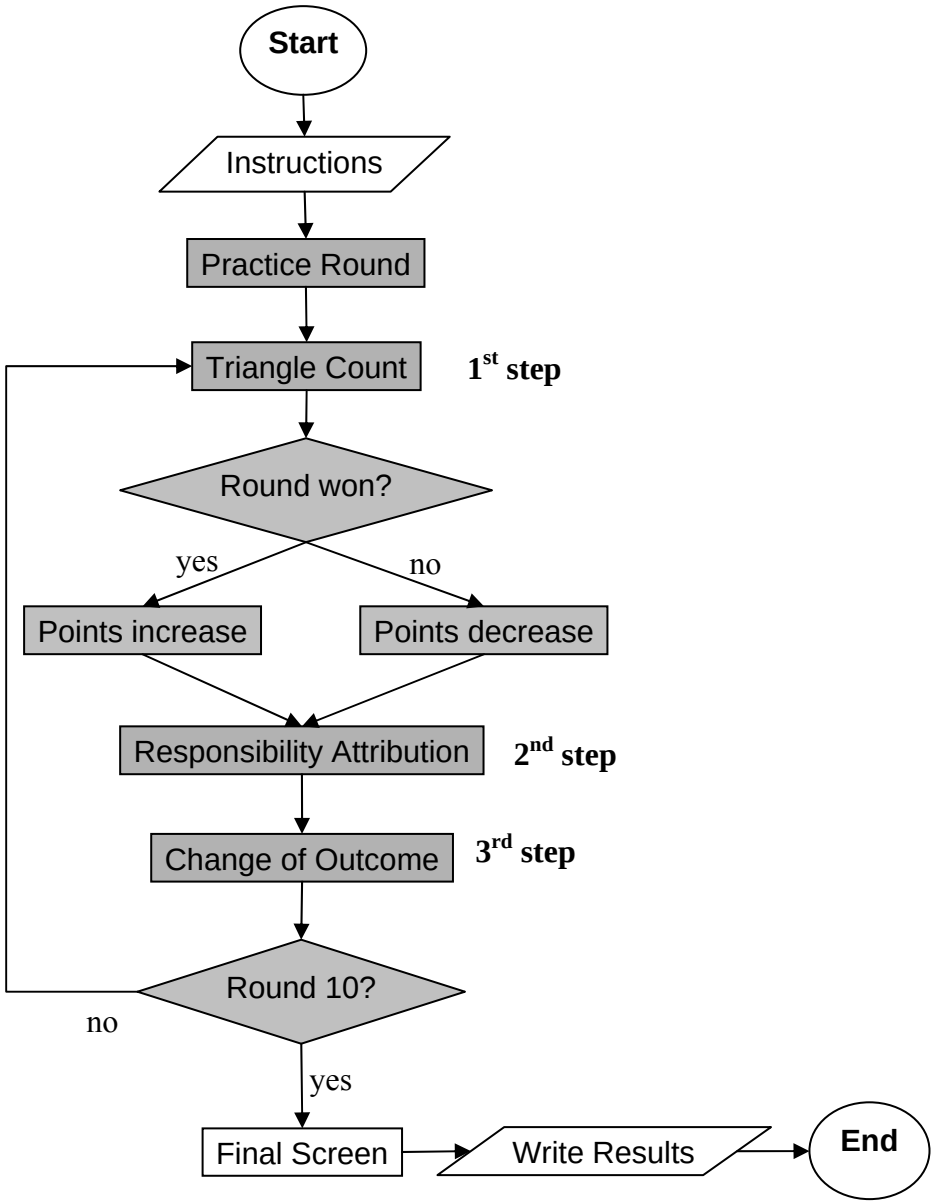
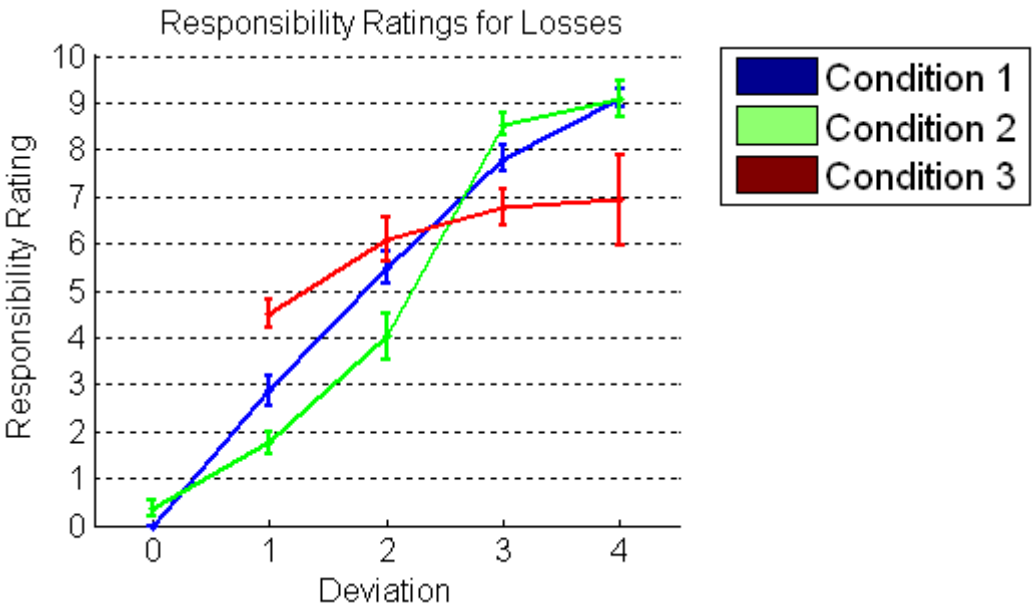


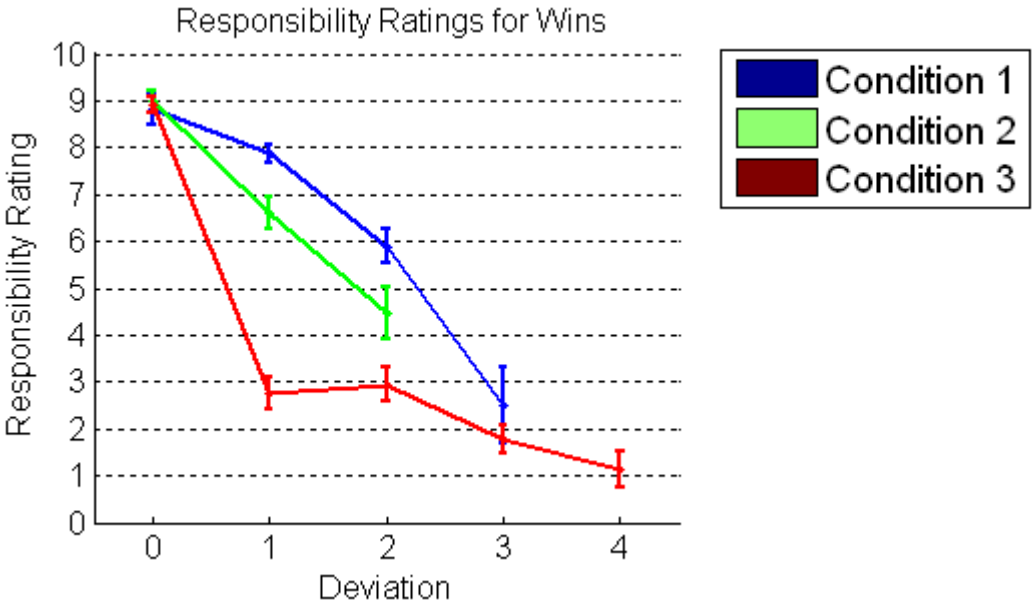
Figure 12

4.2 Results

4.2.1 Responsibility Ratings

Since the results in the second experiment are essentially a replication of the first experiment, as can be seen by comparing the graphs, we will not discuss them separately.


Responsibility Ratings for Losses
Responsibility Rating
Deviation
0
1
2
3
4
5
6
7
8
9
10
Condition 1
Condition 2
Condition 3


Responsibility Ratings for Wins
Responsibility Rating
Deviation
0
1
2
3
4
5
6
7
8
9
10
Condition 1
Condition 2
Condition 3

4.2.2 Model Comparison

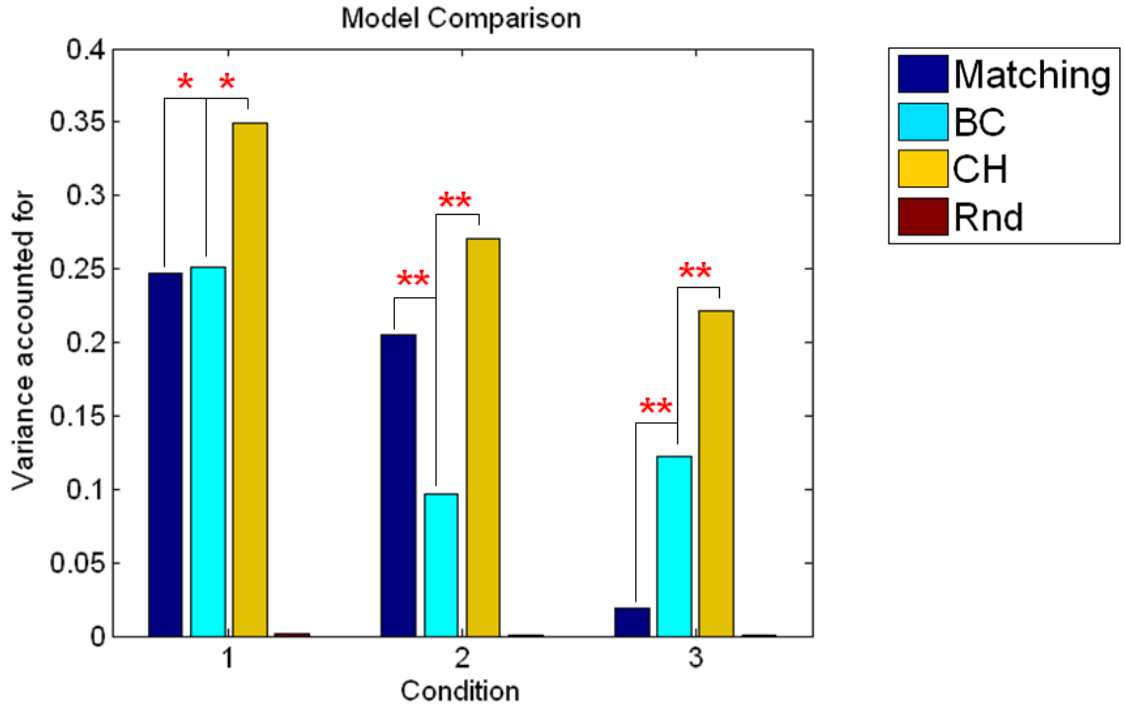


Figure 13

4.2.3 Correlations between Responsibility Rating and Change of Answer

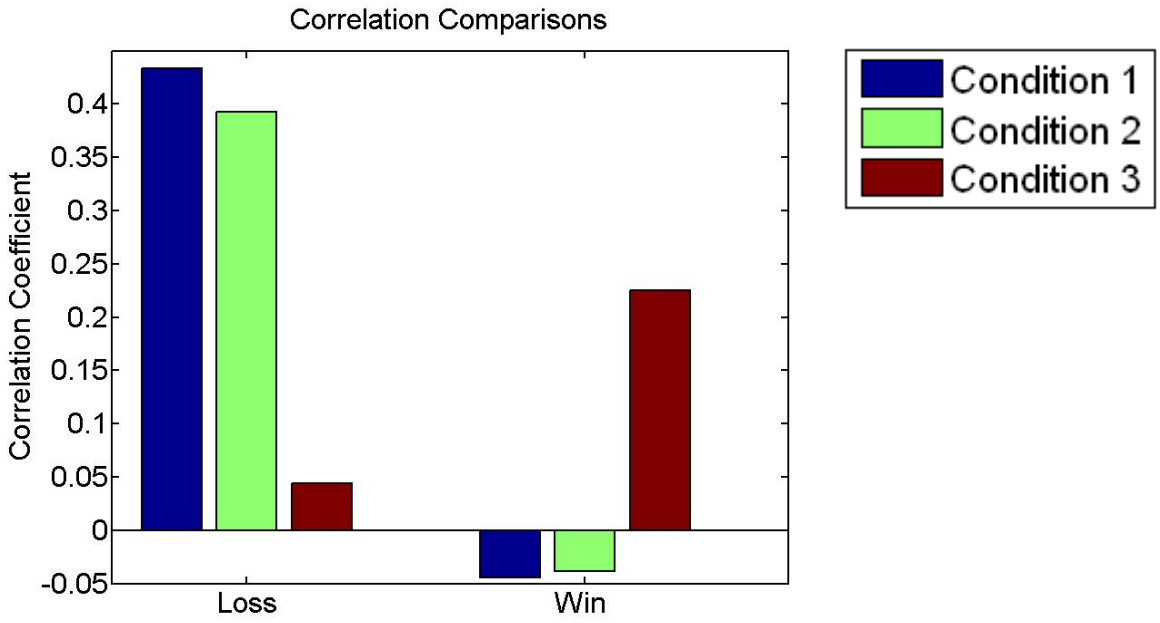


Figure 14

The bars represent the degree of correlation between the responsibility attribution for a particular player in step two of the experiment and how much this player's answer was changed in step three. The first group of bars represents the correlation coefficients for losses.

The second group of bars represents the correlation coefficients for wins. The blue, green and red bars represent, respectively, the first, second and third condition.

For the cases in which the group lost a round, there are high positive correlations between the responsibility ratings in the second step and the degree to which the answer of a player has been changed in the third step. That is, the more a player was rated responsible for the loss the more his or her answer was changed. This correlation, however, does not hold in the third condition.

For cases where the group has lost a round, there is a positive correlation in the third condition between responsibility ratings and the degree to which an answer was subsequently changed. There are marginally negative correlations for both the first and second condition. This means that the less responsible a player was rated for the win the more his or her answer was changed.

4.2.4 Change of Outcome

In order to assess how people interpret the notion of a minimal change, we will only have a look at the results from the first condition. Essentially, the question is whether participants believe that a possible world in which only one big change versus a world in which several small changes have been made resembles the actual world more.

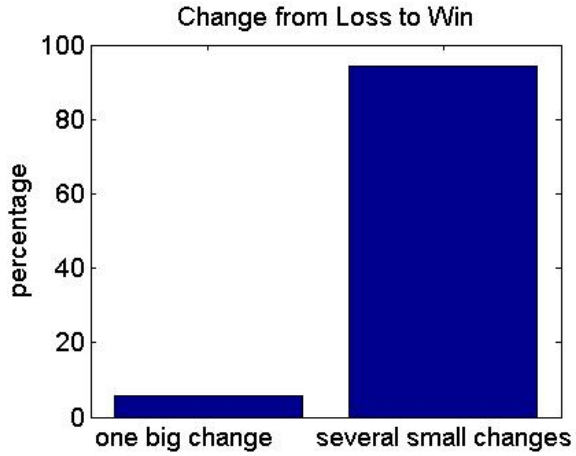


Figure 15

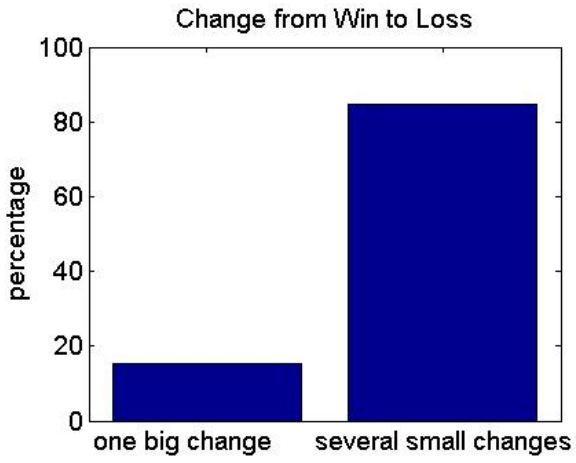


Figure 16

Figure 15 shows the percentages with which participants made one big change versus several small changes, to change the result *from an initial loss to a win*. Thereby, only those rounds were considered, for which changing the answer of one player would have been sufficient to change the result of the round. Furthermore, only those rounds were included in the analysis, for which participants needed to change at least two deviation points.[7] In 5.6% of the cases participants preferred to only change the answer of one player. In 94.4% of the cases participants preferred to change the answers of at least two players.

Figure 16 shows the percentages with which participants made one big change versus several small changes to change the result *from an initial win to a loss*. Again, only those rounds were included in the analysis, for which participants needed to change at least two deviation points.[8] In 15.4% of the cases participants preferred to only change the answer of one player. In 84.6% of the cases participants preferred to change the answers of at least two players.

4.3 Discussion

Generally, the results of the first experiment were replicated in the second experiment. The line graphs for the mean responsibility are very similar to the ones in the first condition. The model comparison shows, again, that the CH model fits the empirical data best in each condition. However, whereas in the first experiment the fit of the model was equal for each condition, there is a slight decrease in the model's fit from condition 1 to condition 2 to condition 3. At this point, we do not have an explanation for these differences. In sum, we can conclude that the introduction of the third step in the experiment did not substantially change the results of people's attributions.

[7] There were 36 cases for which those two conditions where met.
[8] There were 39 cases for which this condition was met.

The correlations between participants' responsibility ratings and their subsequent change of the players' answers show another interesting result. For losses, participants seem to have mainly changed the answers of players with high responsibility ratings, for which we can assume that their answers deviated higher from the correct solution. The more a player deviated the more his or her answer was changed. This result, however, can not be seen for the third condition. There is essentially no correlation between responsibility rating and amount of change. This might be due to the fact that in the third condition, the answers of the players who deviated highly from the correct solution do not need to be changed in order to change the result. In contrast, it is more likely that participants chose to only change the answer of the player with the smallest deviation.

For wins, we see exactly the opposite pattern. There are no to marginally negative correlations for the first and second condition and a positive correlation for the third condition. For the first condition this means that participants had no tendency to change the player's answers of which they thought they were high or low responsible for the win. For the second condition, people only need to change the answer of one player, for example, from a deviation of 2 to 3. The positive correlation in the third condition is due to the fact that in order to change the result from a win to a loss, all players with a 0 deviation need to be changed. Those are the players that have quite likely received a high responsibility rating for the win.

Particularly interesting are the results of people's interpretation of the notion of a minimal change. For both initial wins and losses, we can see that participants were much more likely to make minor changes to the answers of several players than one big change to only one player. This indicates that people regard a possible world with several small changes as more similar to the actual world than a world with one big change.

5 General Discussion

We have presented a new experimental paradigm, the TG, in which multiple players are collectively responsible for the outcome of their team. We have seen that most participants assign responsibility in a systematic manner, taking into account the different ways of how the group's solution is determined. Even for an outcome that is essentially indivisible, namely a win or a loss, participants are willing to divide the responsibility between the players. In both experiments, the CH model accounts for a substantial amount of variance in the empirical responsibility ratings.

The results of the second experiment have shown that participants' answers are not changed by directing their attention to potential counterfactual outcomes. Further experimentation is needed to answer the question of whether people only behave as if they reason counterfactually or whether cognitive processes might directly operate on counterfactual representations, maybe in the form of mentally manipulating causal models. One way to make progress on that question would be to ask participants to think aloud while performing the task. Although it has several drawbacks, the thinking aloud technique is heavily used in research concerning problem solving. It would be interesting to find out, whether any counterfactual thoughts would appear in the thinking-aloud protocols or whether participants are mainly thinking about the actual answers that a player had given.

Another domain, where there is plenty of space for future research, is in developing plausible computational models. Although the CH model does already account for a substantial amount of variance there is nevertheless room for improvement.

Table 10

		Condition 1 (= sum)	**Condition 2 (= least)**	**Condition 3 (= most)**
Player	**Dev**	**Loss**	**Loss**	**Loss**
John	3	1/2 (i.e. **5**)	1/3	1
Kathy	5	1/2	1/3	1
Mark	3	1/2	1/3	1
Toby	2	1/2	0	1

Table 10 shows, again, the predictions that the model would make for a particular pattern of deviation. It predicts that in the first condition, each player should get assigned a 5 responsibility for the loss. For each player, the answer of only one other player needs to be changed to make the outcome counterfactually dependent on his answer. The model does not, however, take into account *how much* a particular player's answer needs to be changed. The intuition is quite clear that Kathy, whose answer deviated 5 should be made more responsible for the loss than, for example, Toby, whose answer only deviated 2. The next step is thus to develop a weighted model, which takes into account the number of deviation points that need to be changed and evaluate how that model would fit the data.[9]

One of the advantages of modelling data is that it allows one to have a very precise look at interindividual differences. Figure 17 shows how well the CH model fits the responsibility ratings of each individual participant in the first experiment.

[9] The results for a preliminary version of a weighted model are shown in the appendix (9.6).

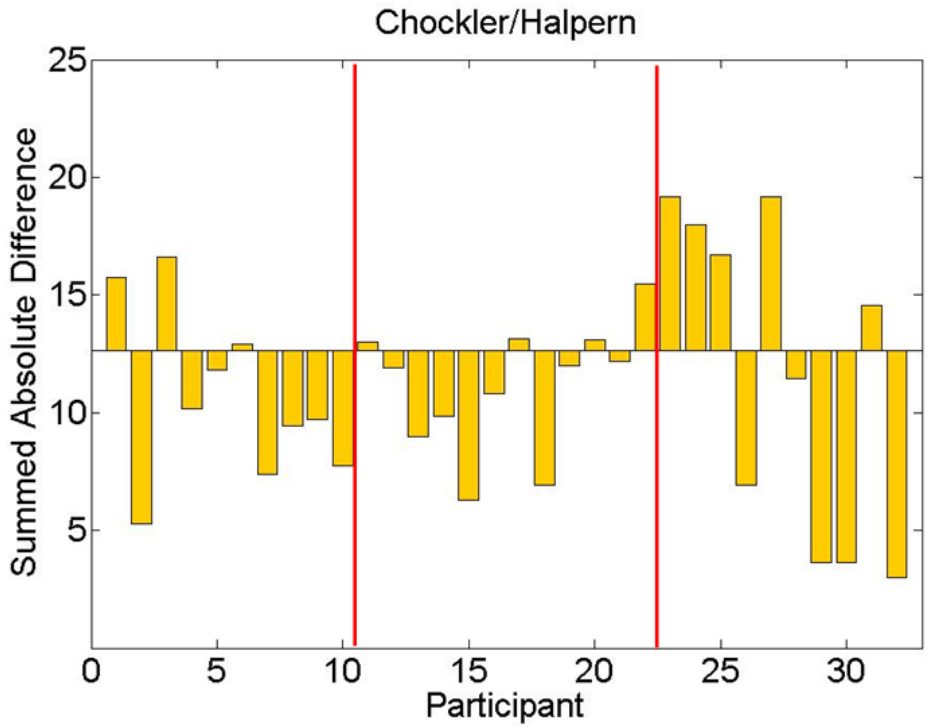


Figure 17

The results for each participant are represented by the bars. The bars indicate by what average the predictions of the model differed from the actual responsibility ratings of that participant. The mean absolute difference for all participants is taken as the baseline. Hence, the closer a bar is to a summed absolute difference of 0 the better this participant's ratings were predicted by the model. The two red lines separate participants in each experimental condition. From left to right, the bars show the model fit per participant from condition 1 to 3. While the interindividual variation of the model fit in condition 1 and 2 seem to be quite low, there appear to be large interindividual differences in the third condition. Some participants' ratings are fit particularly well, while those of others are fit rather poorly. Maybe those participants who are only fitted poorly by the CH model apply a different systematic strategy. If this was the case, it would be interesting to find any interindividual, cognitive differences that could predict the differences in strategy selection.

Besides the aspects of modelling, it would also be interesting to investigate the influence of other variations in the experimental paradigm on the responsibility ratings. One potential variation would be to present the answers of the players successively. Would there

be something like a recency effect in participant's responsibility ratings? Or are the players that first made sure that the round is going to be won or lost conceived as being more responsible than any latter players that had given identical answers? The intuition for that variation comes from the observation that football players, who miss or hit the crucial penalty, are more likely to be blamed or praised than the other players who scored or missed before. Another interesting variation would be, to ask participants for their confidence after they have typed in their answer. Are highly confident players that deviated from the correct solution considered more to blame than players who were not that confident? One could also easily create a situation, where a particular player stands out as the person who makes the team lose. Will this encourage participants to blame this person more for latter "normal" deviations and thus create a phenomenon akin to scapegoating?

A very obvious variation that would be highly worthwhile is to conduct the study in the laboratory. Four participants could be asked to come to the laboratory at the same time. They would be instructed that they are playing the game together and that the money they receive is dependent on the performance of the whole group. Each participant would then be led to a different cubicle, giving them the impression that the computers are connected via network. They would play the exact same TG with the names of the computer players changed to the names of other participants. It is likely that feelings of blame and credit are more naturally elicited in such a situation than they were in the hypothetical game situation. It would be interesting to evaluate, whether participants' answers would be different in this more externally valid situation. Furthermore, we could have a look at other well known phenomena of the social psychology literature. Will participants exhibit a group-serving bias? That is, will they be more likely to attribute responsibility for a groups' win to themselves while attributing responsibility to the whole group for losses (Rantilla, 2000)?

6 Conclusion

We hope that we have convinced the reader that the TG is an exciting, new experimental paradigm that readily lends itself to a variety of fundamental psychological questions and allows, due to its clear formal shape, to assess these questions in a mathematically rigorous manner. We hope that researches in cognitive and social psychology will be attracted by the amenities of the paradigm to fully explore its potential as a powerful research tool.

7 Acknowledgments

I would very much like to thank Dr David Lagnado for being a marvellous supervisor, for encouraging me to think freely and for stipulating the right ideas at the right time. I would also like to thank Joseph Halpern, who helped to clarify the CH model predictions for the TG and Mark Shovman, who suggested ways of graphically representing the data. Further thanks go to Jim Woodward, Martin Speekenbrink and Nick Chater for very helpful comments during the process of developing the TG.

8 References

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, 126, 556-574.

Alicke, M. D. (Forthcoming). Blaming badly. *Journal of Cognition and Culture*

Chockler, H. & Halpern, J. Y. (2003). Responsibility and blame: A structural-model approach. *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, 147-153.

Coffee, J. C. (1981). "No soul to damn: No body to kick": An unscandalized inquiry into the problem of corporate punishment. *Michigan Law Review*, 79, 386-459.

Feinberg, J. (1968). Collective responsibility. *The Journal of Philosophy*, 65, 674-688.

French, P. A. (1984). *Collective and Corporate Responsibility*. Columbia University Press, New York.

Gilbert, D. T. & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin*, 117, 21-38.

Gilbert, M. (2006). Who's to blame? Collective moral responsibility and its implications for group members. *Midwest Studies in Philosophy*, 30, 94-114.

Graham, K. (2006). Imposing and embracing collective responsibility: Why the moral difference?. *Midwest Studies in Philosophy*, 30, 256-268.

Greene, E. J. & Darley, J. M. (1998). Effects of necessary, sufficient, and indirect causation on judgments of criminal liability. *Law and Human Behavior*, 22, 429-451.

Hart, H. L. & Honoré, A. (1959). *Causation in the Law*. Clarendon Press, Oxford.

Hobbes, T. (1982/1651). *Leviathan*. Penguin Books, Harmondsworth.

Knobe, J. (2005). Cognitive processes shaped by the impulse to blame. *Brooklyn Law Review*, 71, 929-937.

Lagnado, D.A., Channon,S. (Forthcoming). Judgments of Cause and Blame: The influence of Intentionality and Foreseeability. *Cognition*.

Lewis, D. (1973). Causation. *Journal of Philosophy*, 70, 556-567.

Mackie, J. L. (1974). *The Cement of the Universe*. Oxford University Press, Oxford. Second edition, 1980.

Mandel. D. R. (2003). Judgment dissociation theory: An analysis of differences in causal, counterfactual, and covariational reasoning. *Journal of Experimental Psychology: General*, 137, 419-434.

Mao, W. & Gratch, J. (2005). Social Causality and responsibility: Modeling and Evaluation. *Lecture Notes in Computer Science*, Springer-Verlag, Berlin.

Menzies, P. (2001). Counterfactual theories of causation. *The Stanford Encyclopedia of Philosophy* (Zalta, E.N., ed.), http://plato.stanford.edu/archives/spr2001/entries/causation-counterfactual/.

Miller, S. & Makela, P. (2005). The collectivist approach to collective moral responsibility. *Metaphilosophy*, 36, 634-651.

N'gbala, A. & Branscombe, N. R. (1995). Mental simulation and causal attribution: When simulating an event does not affect fault assignment. *Journal of Experimental Social Psychology*, 31, 139-162.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge.

Rantilla, A. K. (2000). Collective task responsibility allocation: Revisiting the group-serving bias. *Small Group Research*, 31, 739-766.

Sadler, B. J. (2006). Shared intentions and shared responsibility. *Midwest Studies in Philosophy*, 30, 115-144.

Shaver, K. G. (1985). *The Attribution of Blame: Causality, Responsibility, and Blameworthiness*. Springer-Verlag, New York.

Solan, L. M. (2003). Cognitive foundations of the impulse to blame. *Brooklyn Law Review*, 68, 1003-1029.

Solan, L. M. (2005). Where does blaming come from?. *Brooklyn Law Review*, 71, 939-943.

Solan, L. M. & Darley, J. M. (2001). Causation, contribution, and legal liability. An empirical study. *Law and Contemporary Problems*, 64, 265-298.

Spellman, B. A. & Kincannon, A. (2001). The relation between counterfactual ("but for") and causal reasoning: Experimental findings and implications for jurors' decisions. *Law and Contemporary Problems*, 64, 241-264.

Steiner, I. D. (1972). *Group Processes and Productivity*. Academic Press, New York.

Strawson, P. (1974). *Freedom and Resentment and other Essays*. Methuen, London.

Sverdlik, S. (1987). Collective responsibility. *Philosophical Studies*, 51, 61-76.

Wells, G. L. & Gavanski, I. (1989). Mental simulation of causality. *Journal of Personality and Social Psycholgy*, 56, 161-169.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation.* Oxford University Press, Oxford.

Woodward, J. (forthcoming). Psychological studies of causal and counterfactual reasoning.

9 Appendix

9.1 Responsibility Ratings – Bar Charts

9.1.1 Experiment 1

The following bar charts depict the likelihood that a certain responsibility rating has been given, dependent on the deviation that a player had. For an example, in the first condition when a player deviated 0 and the group lost the round, in 97% of the cases that player was assigned a responsibility of 0 for the loss.

The percentages for the responsibility ratings of 1-4 and 6-9 are cumulated. It is important to note, that cases where the subject deviated 5 or more from the correct solution were very rare. The computer players never deviated more than 4. Hence, the bars for a deviation of 5 rely on only a few cases.

9.1.1.1 Loss

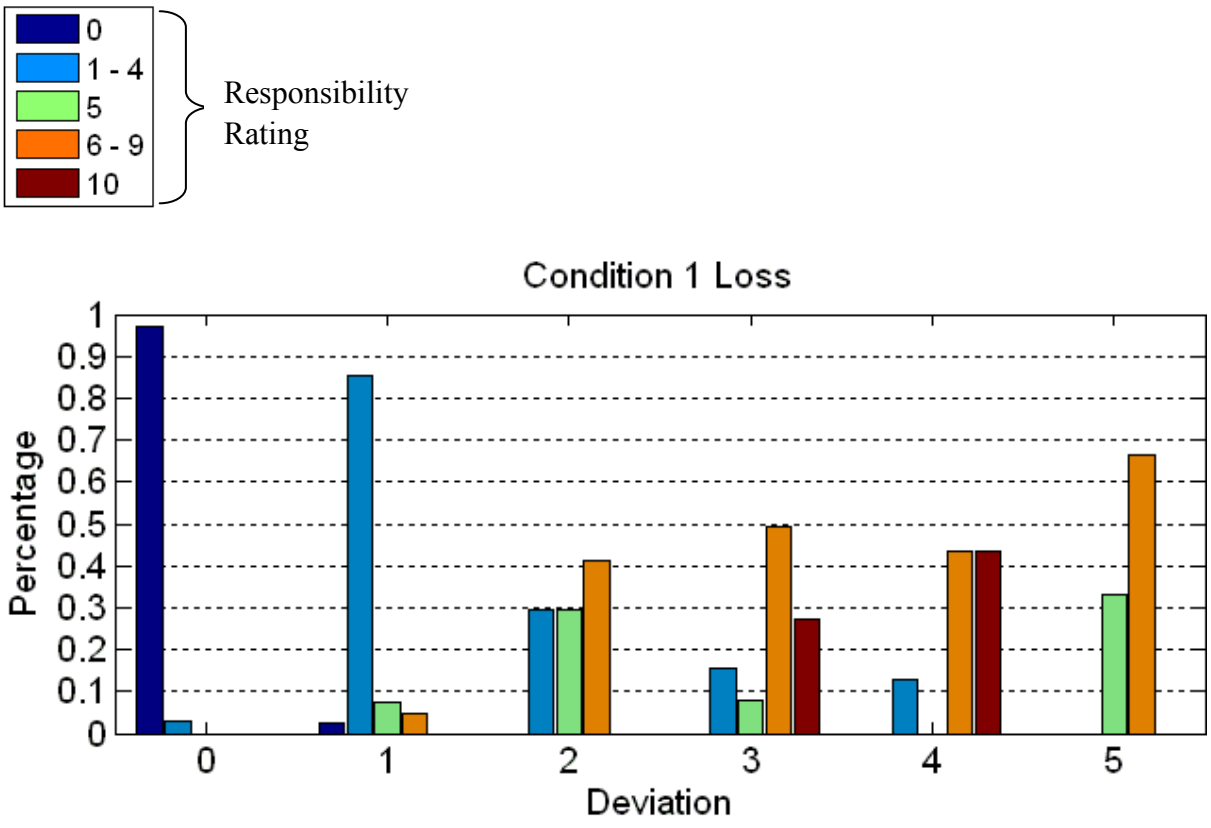


Figure 18

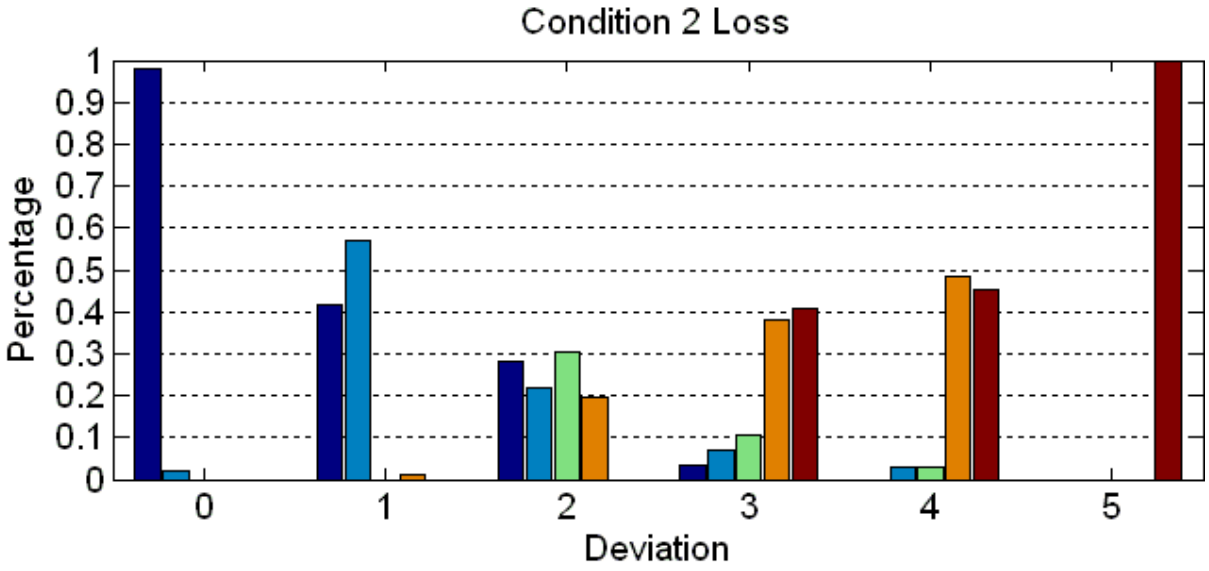


Figure 19

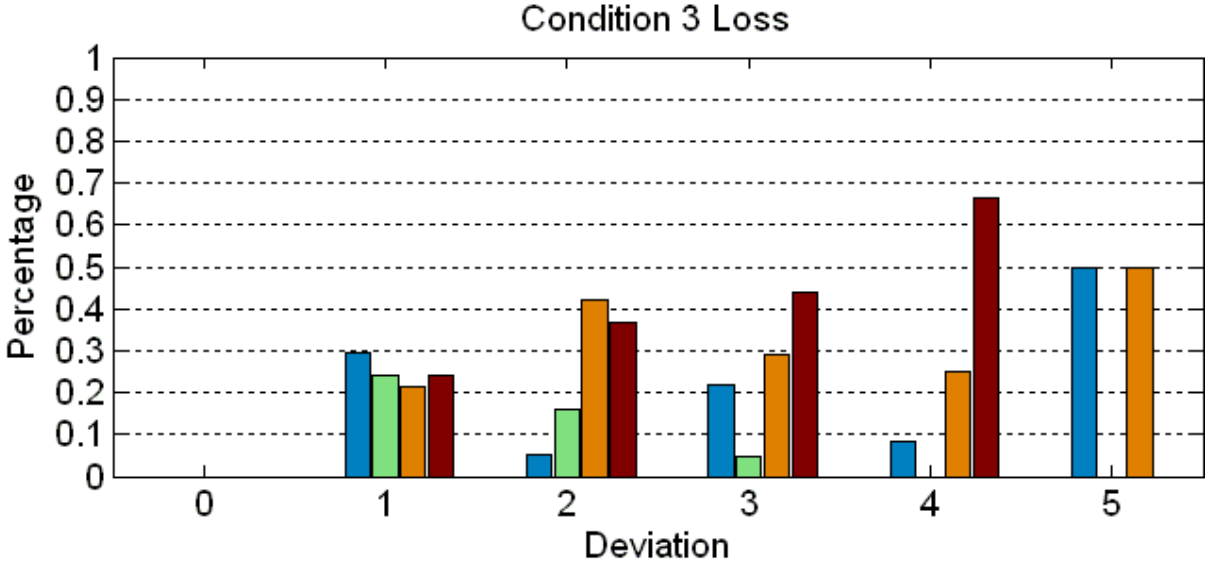


Figure 20

9.1.1.2 Win

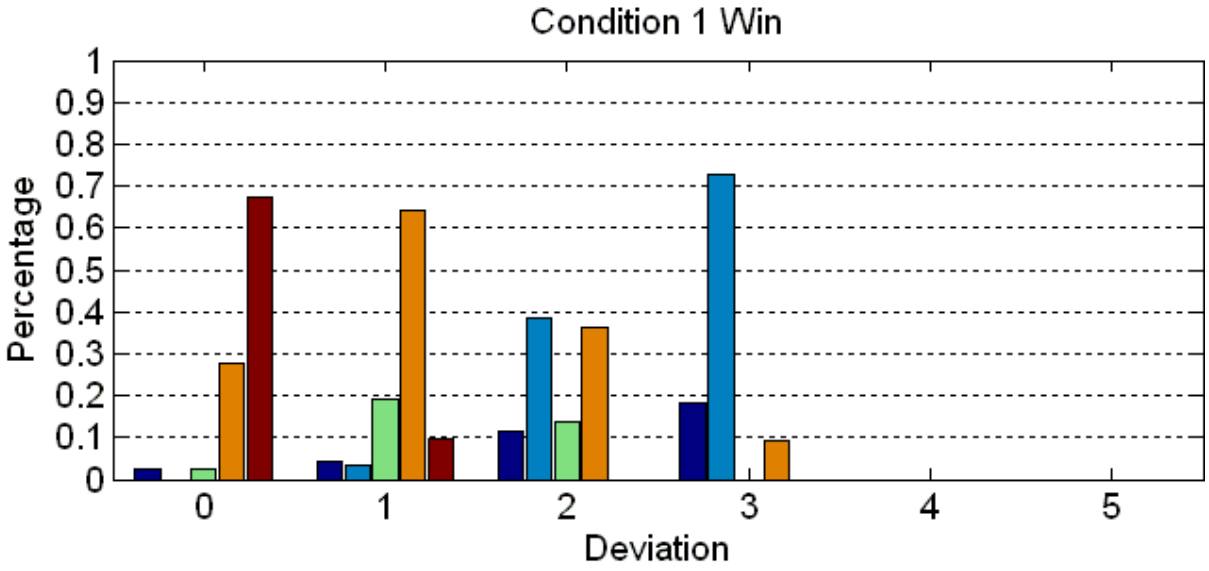


Figure 21

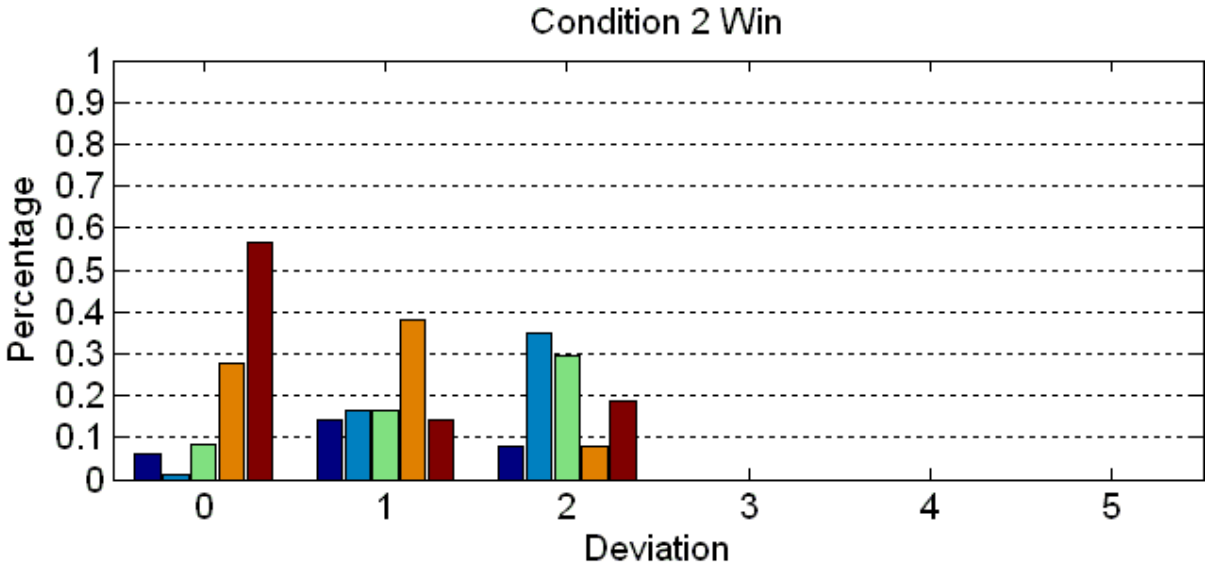


Figure 22

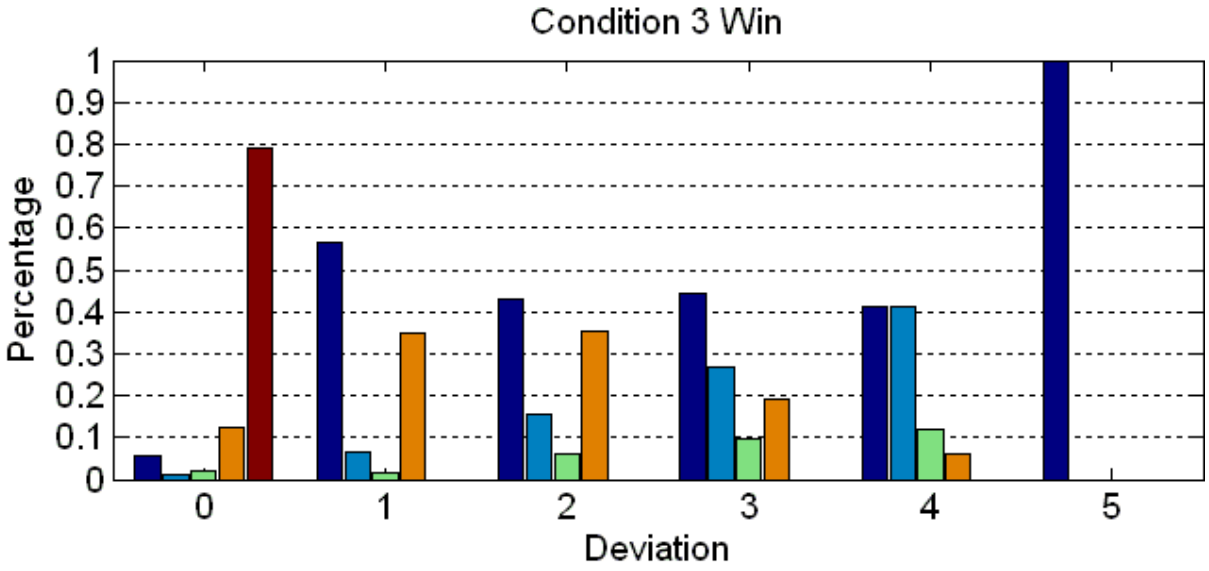


Figure 23

9.1.2 Experiment 2

9.1.2.1 Loss

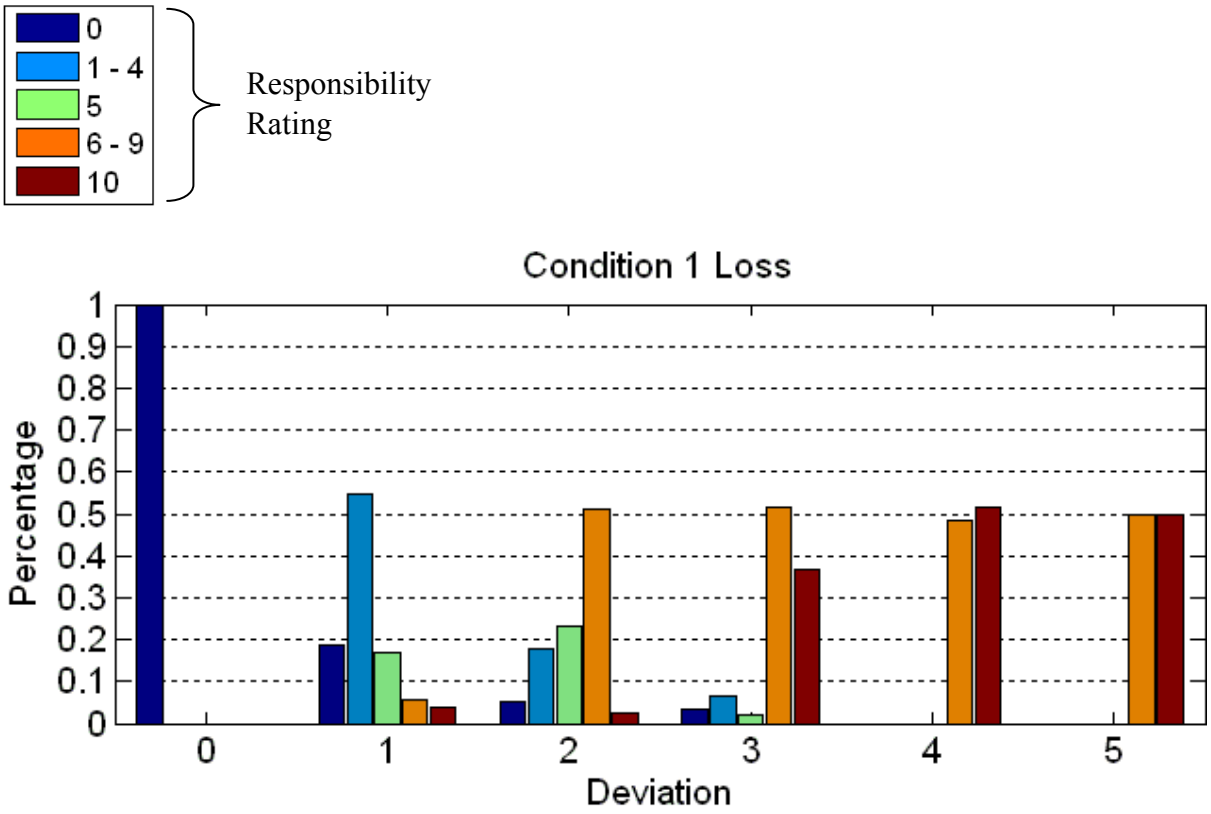


Figure 24

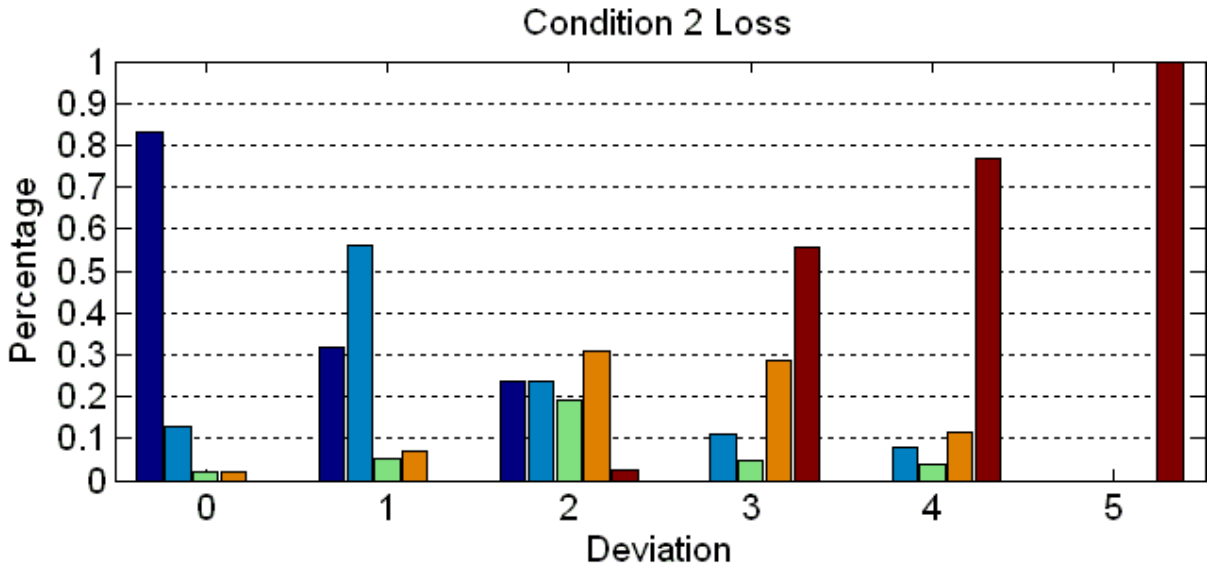


Figure 25

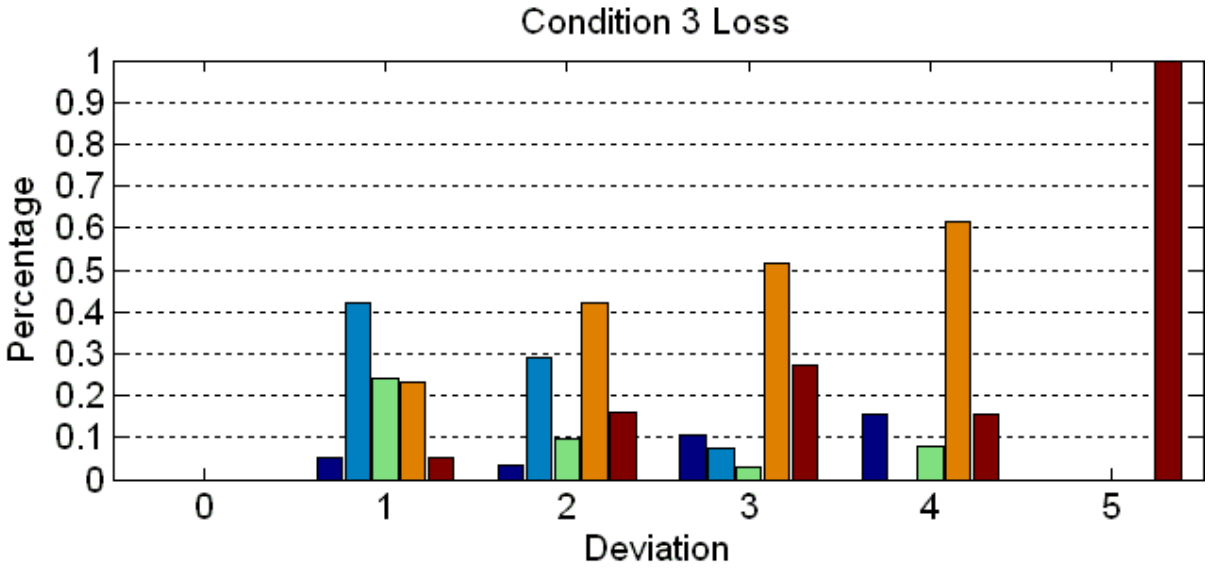


Figure 26

9.1.2.2 Win

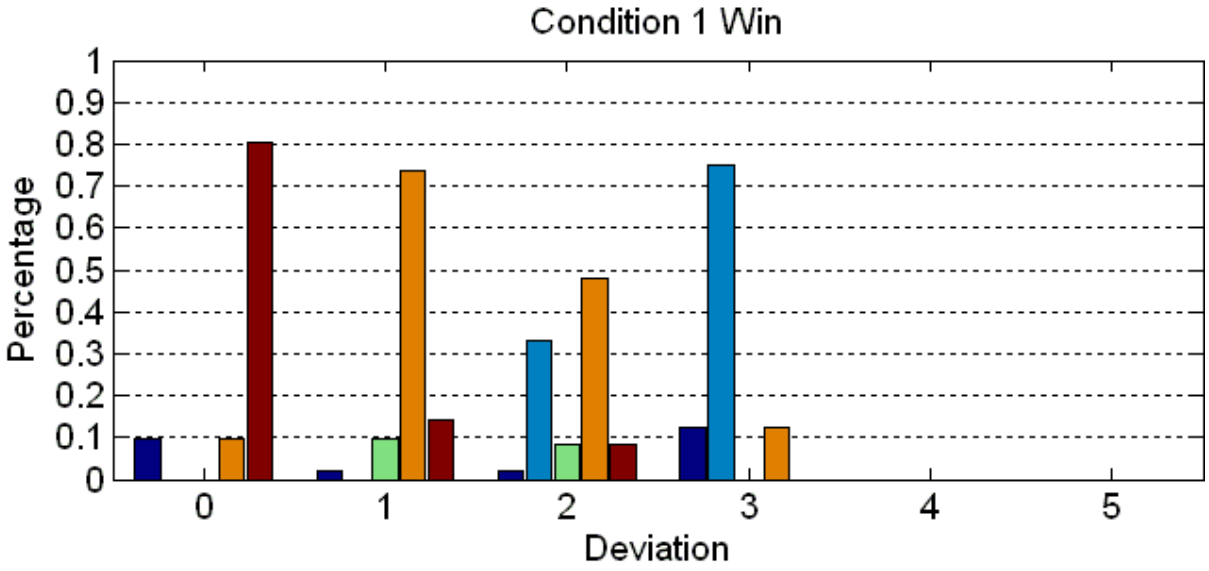


Figure 27

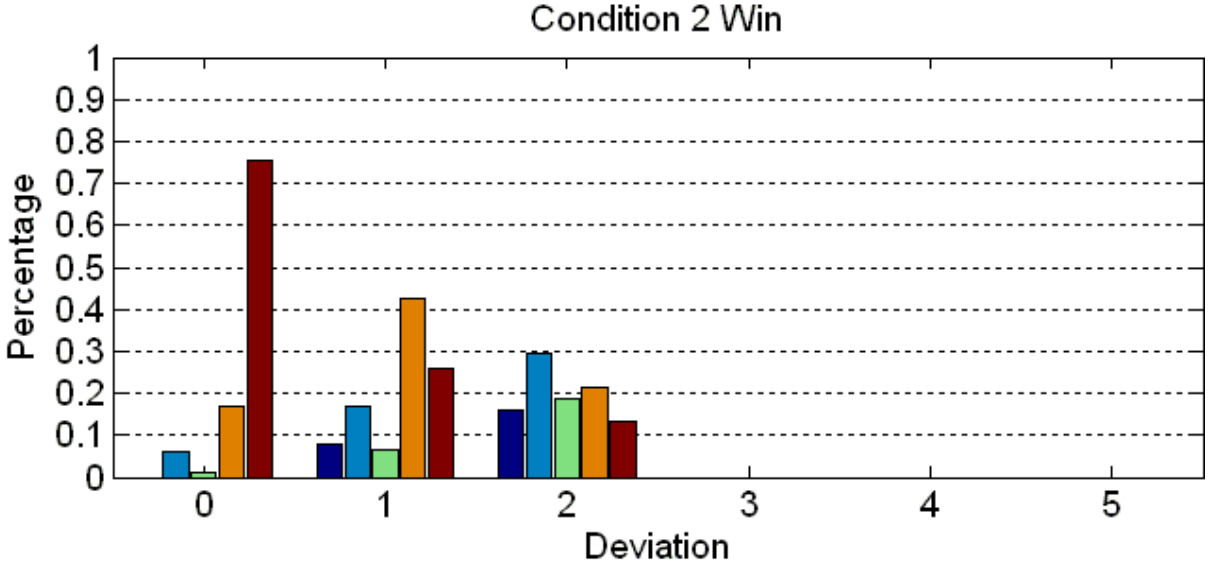


Figure 28

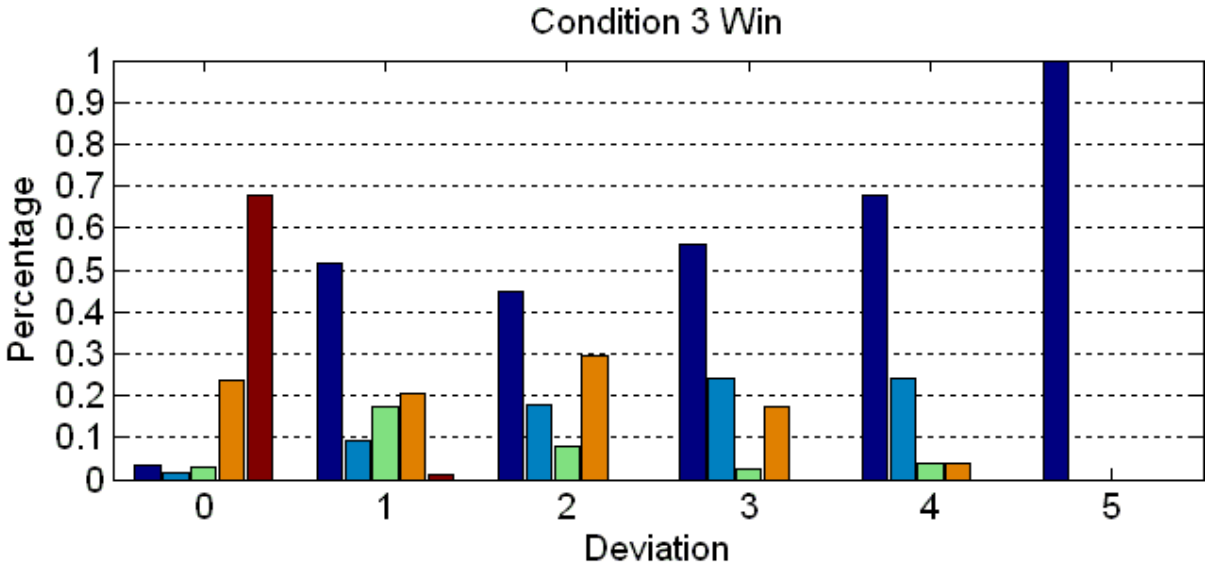


Figure 29

9.2 Correlations between Model Predictions and Empirical Responsibility Ratings

Table 11 and Table 12 show the correlations of the three models with the empirical responsibility ratings for the three experimental conditions.

9.2.1 Experiment 1

Table 11

	Condition 1	**Condition 2**	**Condition 3**
Matching	.51	.46	.0
BC	.47	.36	.43
CH	.56	.55	.54

9.2.2 Experiment 2

Table 12

	Condition 1	**Condition 2**	**Condition 3**
Matching	.50	.45	.14
BC	.50	.31	.35
CH	.59	.52	.47

9.3 Descriptive Statistics of the Computer Players' Deviation

Table 13 and Table 14 show the mean deviation of the three computer players for losses and wins, respectively. These values indicate how well the computer players performed.

9.3.1.1 Loss

Table 13

	N	Minimum	Maximum	Mean	Std. Deviation
Computer 1	166	0	3	1,42	,948
Computer 2	166	0	4	2,29	1,221
Computer 3	166	0	4	2,57	1,233
Valid N (listwise)	166				

9.3.1.2 Win

Table 14

	N	Minimum	Maximum	Mean	Std. Deviation
Computer 1	183	0	3	1,13	,764
Computer 2	183	0	4	1,37	1,150
Computer 3	183	0	4	,82	1,198
Valid N (listwise)	183				

9.4 Difficulty of the Different Diagrams

Table 15 and Table 16 show the mean, mode and range of the participants' deviation in the 10 rounds of the game for experiment 1 and experiment 2, respectively. The values indicate the difficulty of particular diagrams. The higher the mean, the more difficult was the diagram.

9.4.1 Experiment 1

Table 15

Round	1	2	3	4	5	6	7	8	9	10
Mean	1.97	1.72	1.97	0.75	3	0.78	1.66	0.47	0.75	1.97
Mode	1	1	2	1	3	0	1	0	0	1
Range	0-5	0-5	0-5	0-3	0-16	0-3	0-8	0-2	0-3	0-4

9.4.2 Experiment 2

Table 16

Round	1	2	3	4	5	6	7	8	9	10
Mean	1.86	2.65	2.11	0.97	2.76	0.95	1.27	0.57	0.97	1.46
Mode	1	3	2	0	1	0	0	0	1	2
Range	0-4	0-10	0-5	0-6	0-9	0-4	0-4	0-3	0-3	0-3

9.5 Output of SISA – Significance Test of the Difference between Correlations

For an example, the difference of the correlation coefficient in the second condition of experiment 1 between the basic counterfactual model (r = .36) and the modified model (r = .55) is calculated.

For this comparison, SISA shows the following output:

independent (two sample) difference

Confidence Intervals for the difference

C.I.	lower		d(r)		upper
80%	-0.310343	<	-0.19	<	-0.160661
90%	-0.330524	<	-0.19	<	-0.138675
95%	-0.347795	<	-0.19	<	-0.119485
99%	-0.380836	<	-0.19	<	-0.081771

r1-r2= -0.19; z= -3.89774 (**p=0.00005**; 1-p=1)
(multiply p-value with 2 for double sided testing

9.6 Preliminary Version of a Weighted CH Model

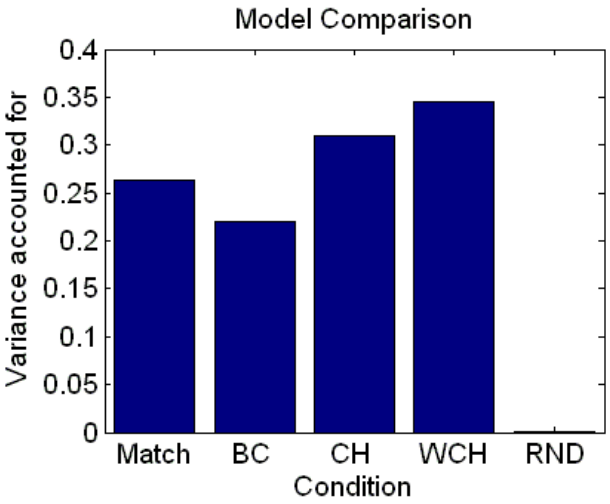


Figure 30

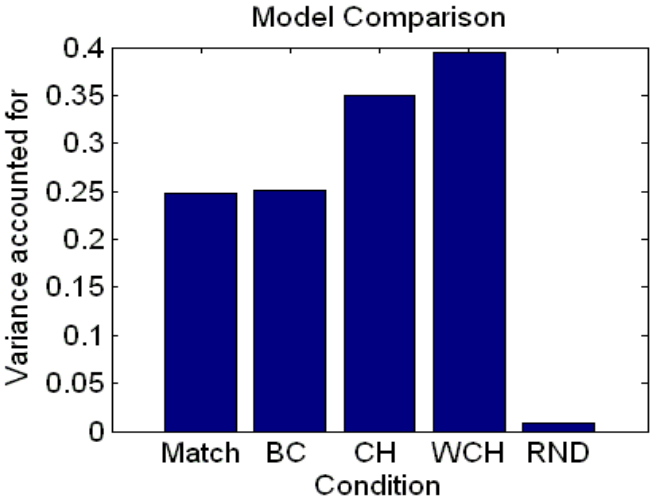


Figure 31

Figure 30 and Figure 31 show the model fits for condition 1 of the three models already discussed as well as the weighted Chockler and Halpern model (WCH) in experiment 1 and 2, respectively. In both experiments, the WCH model fits the data marginally better than the CH model. Since the WCH model predictions only differ from the CH model for the first condition, only the model fits for this condition are shown. Table 17 shows how the predictions of the CH and WCH model differ for an example pattern of deviation.

Table 17

		CH	WCH
Player	**Dev**	**Loss**	**Loss**
John	3	1/2	3/(5+3) = 0.375
Kathy	5	1/2	5/(2+5) = 0.714
Mark	3	1/2	3/(5+3) = 0.375
Toby	2	1/2	2/(5+2) = 0.286

The CH model only takes into account how many players need to be changed to establish counterfactual dependence. The WCH model weights the change by the deviation that the player, whose answer needed to be changed, had. For John, for example, the answer of Kathy

needs to be changed to make the result counterfactually dependent on him. That is why Kathy's deviation of 5 appears in the denominator of the equation. For Kathy, on the other hand, it would have been sufficient if Toby had given a different answer, to make the result counterfactually dependent on her. In general terms the formula for the WCH model can be described as follows:

$$\text{Resp}(C_i) = \frac{Deviation(C_i)}{Deviation(Y) + Deviation(C_i)}$$, where Y denotes the player with the minimal deviation sufficient to establish counterfactual dependence for C_i.