Cognition 115 (2010) 166–171

Contents lists available at ScienceDirect

Cognition

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

Spreading the blame: The allocation of responsibility amongst multiple agents

Tobias Gerstenberg, David A. Lagnado *

Department of Cognitive, Perceptual and Brain Sciences, University College London, UK

ARTICLE INFO

Article history:
Received 3 March 2009
Revised 27 November 2009
Accepted 15 December 2009

Keywords:
Collective responsibility
Attribution
Counterfactual reasoning
Causal models
Experimental games

ABSTRACT

How do people assign responsibility to individuals in a group context? Participants played a repeated trial experimental game with three computer players, in which they counted triangles presented in complex diagrams. Three between-subject conditions differed in how the group outcome was computed from the individual players' answers. After each round, participants assigned responsibility for the outcome to each player. The results showed that participants' assignments varied between conditions, and were sensitive to the function that translated individual contributions into the group outcome. The predictions of different cognitive models of attribution were tested, and the Structural Model (Chockler & Halpern, 2004) predicted the data best.

1. Introduction

Who do you blame when your soccer team loses in the final minutes of the game? Your goalkeeper for conceding a soft goal? Your strikers for missing several good opportunities? The whole team for playing below par? Attributing responsibility is a commonplace activity that has attracted widespread research in psychology (Alicke, 2000; Hilton, McClure, & Slugoski, 2005; Lagnado & Channon, 2008; Shaver, 1985), philosophy (Hart & Honoré, 1959/1985) and law (Moore, 2009). The typical focus is on how people attribute blame to individual agents; however, in many situations it is a group of individuals that collectively determines the outcome, and responsibility must be distributed amongst the group. Team sports provide a paradigm example, but issues of group responsibility arise in many areas, including business, medicine, and law. The allocation of credit or blame in such contexts can be problematic, because it is often hard to isolate the separate contributions that each individual made. Consider the traditional sport tug-of-war. Individual power, stamina and technique, as well as coordination within the team, are important determinants of success. How much responsibility should each player bear for the team's win or loss? Should players be held responsible according to their individual contribution? Or perhaps according to whether their contribution made a critical difference to the team's result?

This difficulty in allocating responsibility is compounded by the fact that causes can combine in various different ways to bring about an outcome. Thus, there are several different functions that can translate the actions of each individual member into the group outcome (Steiner, 1972). The nature of this combination function can depend on the rules of the game, the relevant physical or social laws, or practical aspects of the situation (Waldmann, 2007). Three common functions are *addition*, *conjunction* or *disjunction*. In the additive case, each cause contributes something to the final outcome. Tug-of-war is a prototypical example, where each member contributes to the team's overall success. In the conjunctive case, all causes need to surpass a certain threshold. The final outcome is determined by the weakest member of the team.

* Corresponding author. Address: Department of Cognitive, Perceptual and Brain Sciences, University College London, Gower Street, London WC1E 6BT, UK.

E-mail address: d.lagnado@ucl.ac.uk (D.A. Lagnado).

0010-0277/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2009.12.011

For example, a climbing team is only as fast as its slowest member. In the disjunctive case, it only takes one cause to bring about the outcome. The team is as good as its best member. One example is a team quiz, where a correct answer from just one member is sufficient for the team to win the point. How sensitive are people's responsibility judgments to these different combination functions?

Despite the importance of these questions for attribution research, they have received little attention in the psychological literature. This paper introduces a novel experimental set-up to examine how people distribute credit or blame amongst team members, and whether they are sensitive to the different ways that members can combine to produce an outcome. It also evaluates how well these judgments are captured by three competing models of responsibility attribution.

2. Models of attribution

People's judgments of credit or blame are presumed to be based on prior causal attributions, but modulated by various factors such as intention, foresight, mitigating circumstances or potential justifications (Lagnado & Channon, 2008; Shaver, 1985). This paper focuses on the causal attribution stage and investigates to what extent the allocation of responsibility is influenced by people's knowledge of the causal function that translates individual actions into a group outcome. We test three models of responsibility allocation. All models involve two steps: (1) determine which of the agents in the group are causes of the collective outcome and (2) distribute responsibility amongst the identified causes.

2.1. The Matching Model

The Matching Model sees each agent within the group as a cause of the collective outcome. It predicts that people assign responsibility in direct proportion to the individual contribution of each agent. Applied to the tug-of-war example, each player's pulling power might serve as a proxy for responsibility allocation. However, this strategy becomes problematic when the individual contributions are hard to estimate. Furthermore, there is a strong intuition that a player should only be held responsible if his action had the potential to make a difference to the team's result. If the team would have won irrespective of what the player did, we are hesitant to attribute any responsibility to him. Despite these limitations, the Matching Model serves as a useful benchmark against which to compare other models.

2.2. The Counterfactual Model

The Counterfactual Model incorporates the intuition that the potential of making a difference is a precondition for being held responsible. In the first step, it employs the counterfactual theory of causality (Lewis, 1973) to decide which of the agents caused the collective outcome. On this theory two conditions must be met to qualify an event *A* as the cause of another event *B*: *A* and *B* must both have occurred, and if *A* had not occurred then *B* would not have occurred. In the second step, the model assigns full responsibility to each agent identified as a cause. Several problems with counterfactual theories have been pointed out, the major one being that of causal overdetermination (Collins, Hall, & Paul, 2004). Consider a variation of the tug-of-war example where team A, consisting of four players, beats team B, consisting of only three players. Suppose that team A would have won even if only three of their players had engaged in the game. In this situation the Counterfactual Model would assign a responsibility of 0 for the win to each player in team A. None of the players would be identified as a cause because each player's individual action did not make a critical difference to the team's outcome.

2.3. The Structural Model

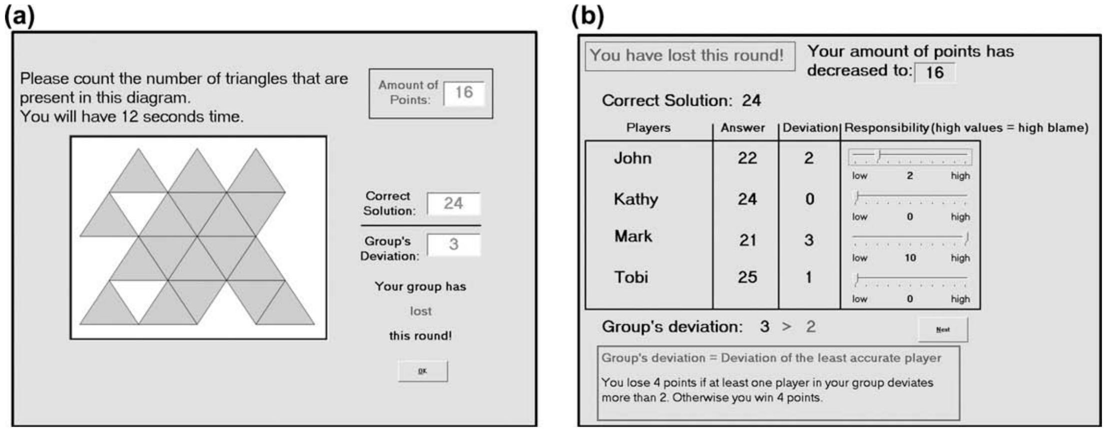
Chockler and Halpern (2004) have developed a model of responsibility attribution that accommodates cases of overdetermination. Their model is cast in the framework of causal models to capture the counterfactual dependencies between sets of events (Halpern & Pearl, 2005). In the first step, their theory offers a relaxed criterion of counterfactual dependence. *A* is a cause of *B* if and only if there is a *possible situation* under which *B* counterfactually depends on *A*. In the second step, the degree of responsibility of an individual cause a_1 (from a set of causes a_i) for an effect *b* is determined by the equation: $\text{Resp}(a_1) = 1/(N + 1)$. *N* denotes the *minimal number of changes* that must be made to the original situation to obtain a modified situation where *b* counterfactually depends on a_1. Applied to the tug-of-war example, this means that each of the four players in team A receives a responsibility of 1/2 for their win. Only if one player had dropped out (i.e. one change from the actual situation) would each remaining player's action have been critical for the outcome of the contest.

This paper aims to test these three models of responsibility attribution. Although the relation between causal and counterfactual judgments has been extensively investigated (Kahneman & Miller, 1986; Roese, 1997), the Structural Model has not yet been subjected to empirical test. We aim to discover how participants attribute responsibility to individual persons for outcomes they have brought about collectively and whether differences in the underlying causal structure influence participants' responsibility ratings.

3. Experiment

To investigate how people attribute responsibility in group contexts we developed the Triangle Game. This is an interactive computer game. The participant's task was to count triangles presented in complex diagrams for a brief period of time. Participants were instructed that they were not playing the game individually but in a group together with three computer players. Whether a particular round in the game was won or lost depended on the accuracy of each player in the group.

Each round of the game consisted of two consecutive steps. In the first step, participants were shown the



Please count the number of triangles that are present in this diagram.
You will have 12 seconds time.

Amount of Points: 16

Correct Solution: 24

Group's Deviation: 3

Your group has lost this round!

OK

You have lost this round!

Your amount of points has decreased to: 16

Correct Solution: 24

Players	Answer	Deviation	Responsibility (high values = high blame)
John	22	2	low 2 high
Kathy	24	0	low 0 high
Mark	21	3	low 10 high
Tobi	25	1	low 0 high

Group's deviation: 3 > 2

Next

Group's deviation = Deviation of the least accurate player

You lose 4 points if at least one player in your group deviates more than 2. Otherwise you win 4 points.

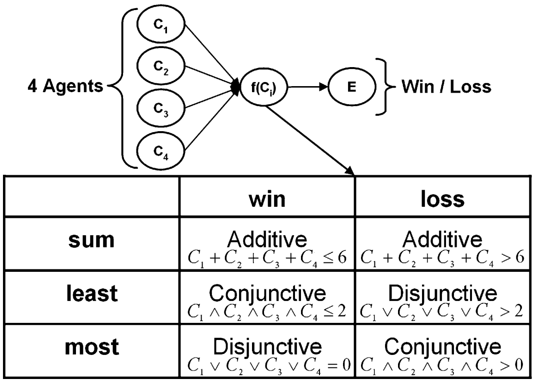
Fig. 1. (a) First step of each round: triangle count. Note: not only the small triangles count but also larger triangles that are made up of smaller triangles. (b) Second step of each round: responsibility attribution. The first three rows are the computer players and the fourth row is the participant.

diagram for 12 s (see Fig. 1a). They had to count the triangles and type in their answer. They then saw the correct solution and were informed about whether their group won or lost the round. The team's points tally changed accordingly. Participants were instructed that the aim of the game was to win as many points as possible.

In the second step, participants viewed a table, which contained the answers of each player in their group (see Fig. 1b). Furthermore, participants could see how much each player's answer deviated from the correct solution. Participants used sliders to indicate each player's responsibility for the group's loss or win in that particular round. The scale ranged from 0 (not responsible at all) to 10 (very responsible).

The experiment had three experimental conditions, which differed only in terms of how the players' contributions were combined to determine the team's outcome. In the *sum* condition, the group's deviation equaled the *sum* of each player's individual deviation from the correct solution. If this sum exceeded the comparison value of 6 the group lost the round. In the *least* condition, the group's deviation equaled the deviation of the *least accurate* player in that round. If this player's deviation from the correct solution exceeded 2, the group lost. In the *most* condition, the group's deviation equaled the deviation of the *most accurate* player. The group won the round only if this player gave the correct solution. Fig. 2 gives an overview of the different forms that the integration function $f(C_i)$ could take depending both on the experimental condition and on whether a round was won or lost.

These three conditions map onto the three main combination functions. In the *sum* condition, the integration function was additive for both wins and losses. Thus, each player's contribution was added to determine the group's result. In the *least* condition, the integration function was conjunctive for winning and disjunctive for losing. In order for the group to win a round the answers of all players needed to deviate 2 or less from the correct solution. In the case of losing it was sufficient if a single player's answer deviated more than 2. The nature of the integration function in the *most* condition was opposite to that in the *least* condition: the group won if at least one player gave the correct solution, independent of how accurate the other players in the team were. It lost if no player was able to get it right.



	win	loss
sum	Additive $C_1 + C_2 + C_3 + C_4 \leq 6$	Additive $C_1 + C_2 + C_3 + C_4 > 6$
least	Conjunctive $C_1 \wedge C_2 \wedge C_3 \wedge C_4 \leq 2$	Disjunctive $C_1 \vee C_2 \vee C_3 \vee C_4 > 2$
most	Disjunctive $C_1 \vee C_2 \vee C_3 \vee C_4 = 0$	Conjunctive $C_1 \wedge C_2 \wedge C_3 \wedge C_4 > 0$

Fig. 2. Above: general causal structure of the game. Below: different forms of the integration function for the *sum*, *least* and *most* condition. For each condition, the integration function is labeled and formally specified for both wins and losses.

There are many causal structures in the real world that map onto the different integration functions embedded in these three experimental conditions. This experiment explores how well the three cognitive models capture peoples' distributions of responsibility under these different circumstances. The complex patterns across the three conditions present a substantial challenge, especially given that the models contain no adjustable parameters for wins vs. losses, or for different combination functions.

3.1. Method

3.1.1. Participants and materials

Sixty-nine participants were recruited via E-Mail and played the Triangle Game on individual computers. They participated for the chance of winning one of three prizes in a £150 draw.

3.1.2. Procedure

In a between-subjects design, participants were assigned randomly to one of the three experimental conditions. The instructions in each condition were identical except for how the deviation of the group's answer was calculated from the individual player's answers. Participants were able to remind themselves of the rules throughout the game. Each round of the game consisted of two steps: first, the triangle count and then the responsibility attribution as described above. The game finished after 10 rounds were played.

3.2. Results

3.2.1. Mean responsibility ratings

Fig. 3a shows the mean responsibility rating assigned to a player depending on their deviation for rounds that the group lost. Fig. 3b shows these ratings for rounds that the group won. To establish whether participants gave different patterns of responsibility ratings in the three conditions we conducted separate ANOVAs for losses and wins, with Condition and Deviation as factors. Comparisons were only made for values of Deviation in which responsibility ratings were available for all three conditions, i.e. for a deviation of 1–4 for losses and 0–2 for wins.

For losses there was no effect of Condition, $F(2,652) = 0.14$, $p = .875$, $\eta^2 = .042$, but an effect of Deviation, $F(3,652) = 10.99$, $p < .01$, $\eta^2 = .846$, and an interaction between Condition and Deviation, $F(6,652) = 13.54$, $p < .001$, $\eta^2 = .111$. For wins there was no effect of Condition, $F(2,752) = 2.58$, $p = .190$, $\eta^2 = .562$, but a marginal effect of Deviation, $F(2,752) = 6.54$, $p = .055$, $\eta^2 = .765$, and an interaction between Condition and Deviation, $F(4,752) = 25.03$, $p < .001$, $\eta^2 = .117$.

Fig. 3a shows that the effect of Deviation for losses is due to the general trend of attributing more responsibility for increased deviation values, which held in all three conditions. The significant interaction for both wins and losses shows that the relationship between responsibility ratings and a player's deviation were qualitatively different between the three conditions. These analyses establish that the experimental variation had an influence on the general trend of responsibility attributions. However, they cannot reveal how people distribute responsibility given specific configurations of players' answers. We thus tested the predictions of the proposed cognitive models against participants' responsibility ratings.

3.2.2. Model predictions

The Matching Model predicts that participants match the degree of deviation to the assigned responsibility rating. For losses there is a direct match, e.g. if a player's answer deviated by 4, he gets a responsibility of 4. For wins, the responsibility is determined by subtracting the degree of deviation from 10, e.g. if a player deviated 3, he gets a responsibility of 7. The Counterfactual Model predicts that a player receives full responsibility if the result of the group was counterfactually dependent on her answer. If she could not have changed the outcome of the group by having changed her answer, she gets a responsibility rating of 0. This implies that in cases of causal overdetermination, where no individual player could have changed the team's result by having given a different answer, all players receive a responsibility rating of 0. The Structural Model assigns responsibility based on the minimal number of changes that need to be made to make a player's answer critical to the group's result. The formula used to determine how much responsibility each player should bear, is $\text{Resp}(C_1) = 10 * 1/(N+1)$, where N denotes the minimal number of changes. The factor 10 in the formula is used to fit the values to the attribution scale of the Triangle Game.

Table 1 shows each model's predictions for an example in which the team lost in the *least* condition. Since the situation depicts a loss, the Matching Model's predictions simply equal the deviation values of each player. Because the loss in this situation is overdetermined due to John and Toby both deviating more than 2, the Counterfactual Model identifies none of the players as the cause of the loss and hence assigns no responsibility. The Structural Model, however, identifies both John and Toby as causes and distributes the responsibility between them. Toby receives a responsibility of 5 because one change would have been needed, e.g., a change in John's deviation from 4 to 2, to

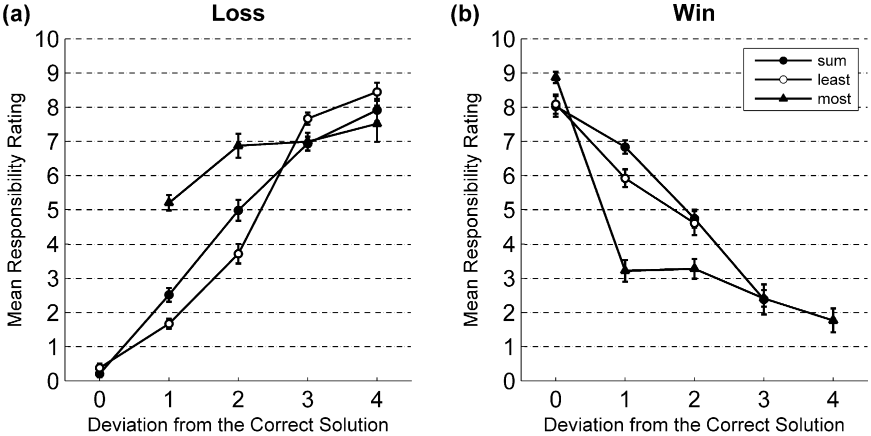


Fig. 3. Mean responsibility ratings with SEMs for (a) losses and (b) wins.

Table 1
Responsibility ratings predicted by the three different models for a loss in the *least* condition.

Player	Deviation	Matching Model	Counterfactual Model	Structural Model
John	4	4	0	5
Kathy	1	1	0	0
Mark	2	2	0	0
Toby	3	3	0	5

make Toby's answer critical for the result of the group. It is worth mentioning that in its present state, the model does not consider *how much* the answer of a player needs to be changed. Thus, the notion of minimal change is applied on the coarse-grained level of individual players rather than on the fine-grained level of deviation points.

To measure model performance, we correlated each model's predictions with the responsibility ratings of each participant (10 rounds × 4 ratings = 40 data points per participant). Fig. 4 shows box plots of the median correlations over participants. A Friedman test established that the correlations differed significantly between the three conditions ($\chi^2(2) = 65.42$, $p < .001$). Wilcoxon tests with Bonferroni corrected alpha value at .016 revealed that the correlations of the Structural Model (*Median* = 0.61) were significantly higher than both the Matching Model (*Median* = 0.42, $T = 616$) and the Counterfactual Model (*Median* = 0.46, $T = 1$). The Matching Model and the Counterfactual Model did not differ significantly ($T = 1039$). 52 of 69 participants were best fit by the Structural Model and the remaining 17 participants were best fit by the Matching Model.

4. Discussion

This paper investigated how people distribute responsibility amongst members of a group. The results showed that people's judgments were sensitive to the causal function that translates individual actions into a group outcome. Moreover, the Structural Model (Chockler & Halpern, 2004) predicted judgments better than a simple counterfactual or Matching Model. The Structural Model assumes that people use a modified notion of counterfactual dependence to make attributions; in particular, that people are guided both by the contribution that an individual member *actually* makes to the group outcome, and by the contribution that they *would have* made had the situation been somewhat different (e.g., if the score of other players had been changed). This model avoids problems of overdetermination, and has the flexibility to accommodate different causal functions without introducing adjustable parameters.

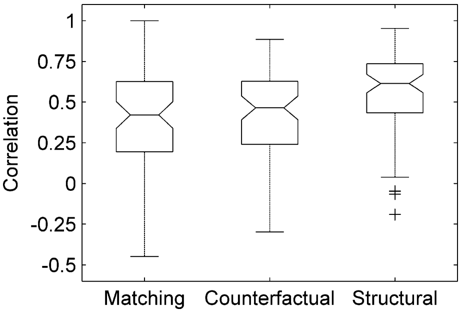


Fig. 4. Model comparison. Median correlations of the three models with the empirical data. Notches indicate 95% confidence intervals of the median.

These findings raise various issues for future research. First, although the Structural Model predicted the data best, there is considerable latitude for developing and testing other models. Indeed the pattern of results suggests that a weighted version of this model might account for the data even better. For example, the graded rather than step-like nature of the responsibility ratings in the *least* and *most* conditions imply that participants were weighting individual players according to their absolute deviations as well as their difference-making contributions. Future studies will explore this possibility. Second, it is unclear to what extent participants engaged in explicit counterfactual reasoning (e.g., via mental simulation of different possible outcomes, cf. Kahneman & Miller, 1986) rather than implicit processing. This question could be addressed via various process tracing methods such as assessing the time course of participants' attributions, introducing a working memory load condition or evaluating verbal protocols (Ericsson & Simon, 1993). Third, in the current paradigm participants are explicitly told the appropriate combination function. But how readily could they learn these functions from merely playing the game, and how would this affect their distribution of attributions? We conjecture that in the absence of explicit instruction people's attributions will initially be determined by individual deviation values (i.e., the Matching Model), but with increased knowledge of the combination function, counterfactual thinking will play a larger role. Finally, empirical work has shown that individual-level attributions are strongly influenced by mentalistic variables such as intention and foresight (Alicke, 2000; Heider, 1958; Hilton et al., 2005; Lagnado & Channon, 2008). In the current study no explicit information was given about such variables. Presumably participants assumed that each player shared the same aims and beliefs. Future work will seek to manipulate these variables, and thus examine the influence of intention and foresight on group attributions.

The distribution of responsibility in groups represents an important but under-explored question. It has implications for understanding human judgment in a wide variety of contexts, including team sports, business, medicine and law. This paper is a first step towards identifying and modeling the principles that underpin people's attributions in situations of collective responsibility.

Acknowledgments

We would like to thank Joe Halpern and Nick Chater for very helpful comments and Maarten Speekenbrink for assistance with the statistical analyses. This research work was supported by the ESRC Centre for Economic Learning and Social Evolution.

References

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

Chockler, H., & Halpern, J. Y. (2004). Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research, 22*, 93–115.

Collins, J. D., Hall, E. J., & Paul, L. A. (Eds.). (2004). *Causation and counterfactuals*. Cambridge, MA: MIT Press.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (rev. ed.). Cambridge, MA: MIT Press.

Halpern, J., & Pearl, J. (2005). Causes and explanations: A structural-model approach. *British Journal of Philosophy of Science, 56*, 843–887.

Hart, H. L. A., & Honoré, A. M. (1959/1985). *Causation in the law* (2nd ed.). Oxford: Oxford University Press.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Hilton, D. J., McClure, J., & Slugoski, B. (2005). Counterfactuals, conditionals and causality: A social psychological perspective. In D. R. Mandel, D. J. Hilton, & P. Catellani (Eds.), *The psychology of counterfactual thinking* (pp. 44–60). London: Routledge.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93*, 136–153.

Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The influence of intentionality and foreseeability. *Cognition, 108*, 754–770.

Lewis, D. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.

Moore, M. S. (2009). *Causation and responsibility: An essay in law, morals, and metaphysics*. Oxford: Oxford University Press.

Roese, N. J. (1997). Counterfactual thinking. *Psychological Bulletin, 121*, 133–148.

Shaver, K. G. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness*. New York: Springer-Verlag.

Steiner, I. D. (1972). *Group process and productivity*. New York: Academic Press.

Waldmann, M. R. (2007). Combining versus analyzing multiple causes: How domain assumptions and task context affect integration rules. *Cognitive Science, 31*, 233–256.