MS #13-03-R2

Running Head: IAT AND ILLUSION OF CAUSALITY

Evidence for an illusion of causality when using the Implicit Association Test to measure learning

Miguel A. Vadillo[1], Jan De Houwer[2], Maarten De Schryver[2],
Nerea Ortega-Castro[3], & Helena Matute[3]

[1]*University College London, London, UK*
[2]*Universiteit Gent, Ghent, Belgium*
[3]*Universidad de Deusto, Bilbao, Spain*

In press. *Learning & Motivation.*

Mailing address:

Miguel A. Vadillo
Division of Psychology and Language Sciences
University College London
26 Bedford Way, London WC1H 0AH, United Kingdom

Tel: +44 20 7679 5364
e-mail: m.vadillo@ucl.ac.uk

Abstract

Our ability to detect causal relations and patterns of covariation is easily biased by a number of well-known factors. For example, people tend to overestimate the strength of the relation between a cue and an outcome if the outcome tends to occur very frequently. During the last years, several accounts have attempted to explain the outcome-density bias. On the one hand, dual-process performance accounts propose that biases are not due to the way associations are encoded, but to the higher-order cognitive processes involved in the retrieval and use of this information. In other words, the outcome-density bias is seen as a performance effect, not a learning effect. From this point of view, it is predicted that the outcome-density bias should be absent in any testing procedure that reduces the motivation or opportunity to engage in higher-order cognitive processes. Contrary to this prediction, but consistent with the most common single-process learning accounts, our results show that the outcome-density effect can be detected when the Implicit Association Test is used to measure the strength of cue-outcome associations.

Keywords: outcome-density effect; contingency learning; causal learning; implicit association test.

Evidence for an illusion of causality when using the Implicit Association Test to measure learning

One of the most remarkable features of human beings and other animals is our outstanding ability to adapt to the regularities in our environment. Quite surprisingly, however, our accurate sensitivity to statistical relations does not make us immune to blatant cognitive illusions, some of them with far-reaching consequences. For instance, many citizens in developed societies still recur to homeopathy and other complementary or alternative medicines (Barnes, Bloom, & Nahin, 2008), despite the fact that they are known to be ineffective (Shang et al., 2005) and also despite their huge economic costs (Nahin, Barnes, Stussman, & Bloom, 2009). The impact of pseudoscience and erroneous causal beliefs in the educative system is equally astonishing (Lilienfeld, Ammirati, & David, 2012).

During the last decades, cognitive psychologists have explored how erroneous beliefs arise as the result of confirmation biases, illusory correlations, overreliance on heuristics, and, most importantly, illusory perceptions of causality (Gilovich, 1991; Vyse, 1997). Current research on associative learning has contributed to our understanding of causal illusions by identifying factors that bias our ability to detect the covariation between a candidate cause and an effect. One of these factors is the probability with which the to-be-explained effect occurs. Imagine that you suffer very frequently from headaches and that you are looking for a remedy to ameliorate your condition. Even without any treatment, headaches tend to disappear very frequently in the interval of a few hours. This warranties that, whatever remedy you decide to take when you feel a headache, its consumption is very likely to be followed by a recovery, even if the remedy itself is absolutely non-effective. However, if this experience happens regularly, it is very tempting, almost

unavoidable, to conclude that the remedy must be effective, because it has been followed by the recovery so many times in the past. Taking this into account, it is hardly surprising that until the development of placebo-controlled, double blind tests, the history of medicine has been "the history of the placebo effect" (Shapiro & Shapiro, 1997). This example illustrates why the overall frequency of an effect can bias the perception of causality. By mere chance, frequent effects will usually happen after other factors, which will be seen as potential causes.

This effect has been studied extensively in the area of human contingency learning. In these experiments, participants are exposed to a series of trials in which a cue might be present or absent and an outcome might follow or not. Their task is to learn to predict the outcome based on the presence or absence of the cue in every trial. For example, a typical cover story used in these experiments invites participants to imagine that they are medical doctors who have to discover whether a patient is allergic to a given food. In each trial, they see whether the patient has taken that food and next, whether he or she suffered an allergic reaction. At the end of the experiment, participants are usually asked to rate the strength of the statistical or causal relationship between the candidate cause (the patient eating the food) and the outcome (her suffering an allergic reaction). Many experiments conducted with this or related tasks have found that even in situations in which there is no statistical relationship between the cue and the outcome, the participants still report that such a relationship exists if the outcome occurred in many trials (Allan & Jenkins, 1983; Allan, Siegel, & Tangen, 2005; Blanco, Matute, & Vadillo, in press; Musca, Vadillo, Blanco, & Matute, 2010). This outcome-density effect also takes place in situations in which participants are not judging the relationship between a neutral cue and an outcome, but the relationship between their own behavior and a consequence (Alloy & Abramson, 1979; Matute, 1995; Msetfi, Murphy, Simpson, & Kornbrot, 2005; Shanks, 1985, 1987).

Early explanations for the outcome-density bias were framed in associative terms (López, Cobos, Caño, & Shanks, 1998; Matute, 1996; Shanks, 1995). For a situation in which there is just one cue and one outcome, the standard associative explanation assumes that a node representing the cue and a node representing the context compete to become associated with the representation of the outcome. The strengths of the cue-outcome and the context-outcome associations are updated on a trial-by-trial basis by means of a simple error correction rule, such as the one proposed by Rescorla and Wagner (1972) in the area of classical conditioning. According to this learning rule, the cue and the context compete to become associated with the outcome, so that by the end of training their respective associations with the outcome will be proportional to their relative predictive validity. When adopting a competitive learning rule, in situations in which there is no statistical relation between the cue and the outcome, the context ends up accruing all the associative strength. However, it is often assumed that the salience of the context is lower than the salience of the cue. This implies that learning is slower for the context than for the cue or, in other words, that the context is a poor competitor during the first stages of learning. Therefore, early in training, the accidental pairings of the cue and the outcome can result in a spurious cue-outcome association that will only disappear afterwards, once the context has accrued enough associative strength. In sum, within association formation models of learning, the outcome-density effect is understood as a transient, preasymptotic bias arising from accidental cue-outcome pairings before the context becomes an effective competitor.

More recently, however, a dual-process model has been offered to account for the outcome-density effect. Allan et al. (2005) conducted a contingency learning experiment in which they manipulated the cue-outcome contingency and, orthogonally, the overall probability of the outcome. As in most contingency learning experiments, in each training trial participants first saw whether the cue was present and had to predict whether the

outcome would follow using a yes/no discrete response. After their response, they were told whether the outcome really appeared afterwards and proceeded to the next trial. At the end of training participants rated the strength of the cue-outcome relationship. The judgments collected at the end of training showed the expected outcome-density effect: for any given level of cue-outcome contingency, participants' judgments varied as a function of the probability of the outcome. Most interestingly, however, this effect was absent in a dependent measure computed from the yes/no responses that participants gave during training. Specifically, Allan et al. compared the proportion of "yes" responses in cue-present trials and the proportion of "yes" responses in cue-absent trials. If participants think that there is a positive relationship between the cue and the outcome, they should predict the outcome more often when the cue is present than when the cue is absent. However, Allan et al. found that this measure was not sensitive to the outcome-density effect. From their point of view, the fact that the predictive responses were unaffected shows that the outcome-density effect does not influence how people encode the relationship between the cue and the outcome. However, participants' judgments are not based solely on the encoded cue-outcome relation. Additional processes involved in the production of the judgment would be responsible for the outcome-density effect (see also Allan, Siegel, & Hannah, 2007). In other words, the outcome-density effect is assumed to be a performance phenomenon, not a learning phenomenon. Quite interestingly, Perales, Catena, Shanks, and González (2005) have proposed a similar explanation for a related bias that happens when it is the cue, instead of the outcome, that occurs very frequently, namely the cue-density effect (see also Matute, Yarritu, & Vadillo, 2011; Vadillo, Musca, Blanco, & Matute, 2011).

Although the theoretical framework proposed by Allan et al. (2005) and Perales et al. (2005) is certainly inspiring and sheds new light on these effects, it is based on limited

evidence. Dissociations between dependent variables can sometimes reflect the action of independent cognitive systems. But they can also be the product of a methodological artifact like, for instance, the limited reliability of one of the dependent measures (Shanks & St. John, 1994). In the case of the experiments conducted by Allan et al. and Perales et al., doubts can be raised about whether the dependent measures collected from trial-by-trial predictions are as reliable as the judgments provided at the end of training. For one thing, the outcome predictions were collected during training, that is, before the cue-outcome association had been properly encoded. Moreover, even if participants believe that there is a probabilistic relationship between the cue and the outcome, their predictive responses (yes/no) must be discrete. If these shortcomings of using trial-by-trial predictions as a dependent measure have a detrimental effect on their reliability, it is hardly surprising that they are more likely to remain uninfluenced by contextual manipulations, such as outcome-density.

Interestingly, research on the development of illusory correlations in stereotype formation can provide useful information in this debate. In these illusory correlation studies, participants are typically exposed to information about good and bad behaviors shown by two different social groups (Hamilton & Gifford, 1976). In every trial, they see information about a desirable or undesirable trait of a member of one group. Most importantly, there are more trials showing information about one of the groups than about the other. The proportion of desirable traits is also higher than the proportion of undesirable traits, although it is the same for both groups. The usual result is that even though both social groups show the same proportion of positive and negative traits, the group for which more information was shown (i.e., the majority group) is rated more positively than the group for which less information was shown (i.e., the minority group). In spite of obvious differences, this illusory correlation effect is to some extent similar to

the cue- and outcome-density effects explored in the contingency learning literature, assuming that social groups play the role of the cue and positive and negative traits play the role of the outcome. The higher proportion of one group over the other would mirror the cue-density manipulation, and the higher proportion of positive traits over negative ones would mirror the outcome-density manipulation. The illusory-correlation effect has traditionally been explained in terms of single-process learning accounts. For instance, Hamilton and Gifford (1976) argued that the illusory-correlation effect was mainly driven by the fact that uncommon events are more salient than frequent events. In the context of this paradigm, members of the minority group are, by definition, uncommon and negative behaviors are also uncommon. Therefore, their occasional coincidences are very salient, which boosts learning about the relationship between the minority group and the negative behaviors. More recent accounts differ in the details, but they also address the role of learning processes (e.g., Fiedler, 2000; Sherman, Kruschke, Sherman, Percy, Petrocelli, & Conrey, 2009).

In contrast to this dominant view, a recent study by Ratcliff and Nosek (2010) obtained support for a dual-process performance interpretation quite similar to the one proposed by Allan et al. (2005) and Perales et al. (2005) in the area of contingency learning. Unlike previous experiments on illusory correlation, in the study conducted by Ratliff and Nosek (2010) participants had to report their explicit attitudes towards the majority and the minority groups, and their attitudes also were measured by means of an Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998). The IAT is a reaction-time task in which participants categorize four sets of words or stimuli. In Ratliff and Nosek's study, participants categorized positive and negative words, on the one hand, and the names of people from the majority and the minority groups, respectively. During some phases of the IAT, participants had to use one key to respond to positive words and

to the names from the majority group and a different key to respond to negative words and to the names from the minority group. During some other phases, the response assignment was reversed, so that they had to use one key to respond to positive words and to the names of the minority group and the other key to respond to negative words and to the names of the majority group. If participants have a positive attitude towards the majority group, one would expect them to perform better in the first task than in the second task.

Crucially, Ratliff and Nosek (2010) did not find an illusory correlation effect in the IAT scores, even though the explicit measures did reflect a preference for the majority group. Although their theoretical interpretation of this finding is framed in different terms, it is surprisingly similar to the dual-process account provided by Allan et al. (2005, 2007) and Perales et al. (2005) for cue- and outcome-density biases in contingency learning. In line with the popular dual-process framework proposed by Gawronski and Bodenhausen (2006), Ratliff and Nosek concluded that the fact that only explicit evaluations showed the illusory correlation effect suggests that the bias is not due to the way associations are encoded in memory (which is assumed to be indexed by the IAT), but to deliberate reasoning processes that qualify the expression of these associations (which is assumed to influence only explicit measures). In other words, Ratliff and Nosek argued that the illusory correlation effect is not a learning phenomenon but a performance phenomenon.

Although the study of Ratliff and Nosek (2010) is focused on stereotype formation, their results are also relevant to our understanding of cue- and outcome-density biases as studied in contingency learning research. However, there are remarkable differences between the research paradigms used in both traditions that preclude any direct comparison. Moreover, as far as we know, the study of Ratliff and Nosek is the only study that has used the IAT in research on illusory correlations. Because it is possible that subtle aspects of their procedure were responsible for the null result, there is clearly a need for

more research on this topic. The purpose of the present experiment is to find out whether performance in the IAT is sensitive enough to measure the outcome-density effect that usually develops in the standard experimental paradigm used in contingency learning and causal judgment studies. On the basis of dual-process accounts such as those provided by Allan et al. (2005, 2007) and Perales et al. (2005), one would not expect to observe outcome-density effects in the IAT because the effect is assumed to occur not during association formation (i.e., learning) but when making explicit contingency judgments (i.e., performance). An association formation model, on the other hand, does attribute the outcome-density effect to learning processes and therefore predicts an effect on any measure of learning, regardless of whether the measure requires an elaborated, explicit judgment or not.

Method

Participants and Apparatus

Forty-eight psychology students at Ghent University participated in the experiment in exchange for €4.[1] The experimental task was programmed in Visual Basic 6.0 using Windows API functions QueryPerformanceCounter and QueryPerformanceFrequency to register reaction times.

Design

Most experiments on the outcome-density effect include a single cue that is paired with a single outcome. The participants are exposed to a series of trials in which either the cue and the outcome might both be present, or the cue is present but the outcome is not, or the cue is absent but the outcome is present, or, finally, both events are absent. By manipulating the frequency of each of these trial types, it is possible to expose different groups of participants to situations that involve the same overall statistical relationship between the cue and the outcome, but with different overall probabilities of the outcome.

However, this single-cue design is poorly suited to the requirements of the IAT. Given two categories, X and Y, and two attributes, 1 and 2, the IAT is assumed to provide a measure of the extent to which the X-1 and the Y-2 associations are stronger than the X-2 and Y-1 associations. Therefore, in order to make the design of our experiment compatible with the requirements of the IAT, the design of the learning phase was doubled, so that each participant received information about two cues and two outcomes (also see De Houwer & Vandorpe, 2010).

The design of the experiment is summarized in Figure 1. Two groups of participants were trained with two different cues, X and Y, in two different contexts, A and B. X was always presented in Context A, Y in Context B. For participants in Group 0.5, cue X was equally likely to be followed by outcomes O1 or O2. Therefore, the probability of any of the outcomes given X was 0.5. Most importantly, the probability of those outcomes was also 0.5 in the absence of X. This means that there was no statistical contingency between X and the outcomes or, in other words, that cue X did not provide any additional information beyond that conveyed by context A alone. The information provided in Context B about Y was exactly the same.

For participants in Group 0.9, cue X was more likely to be followed by O1. However, within context A the probability of O1 was equally high in the absence of X. Therefore, just as in the previous condition, O1 was not contingent upon cue X and, therefore, cue X did not provide any additional information beyond that conveyed by context A. The information given about cue Y in context B was the exact reverse: Cue Y was paired more often with O2 than with O1. However, within context B the probability of O2 was not higher in the presence than in the absence of Y.

Given the relatively large number of X-O1 and Y-O2 pairings in Group 0.9 relative to Group 0.5, we expected participants in the former group to give higher causal judgments

for those relationships by virtue of the outcome-density effect. However, it is unclear whether a parallel effect would be observed in an IAT measuring the strength of the X-O1 and Y-O2 associations. From the point of view of single-process learning models, there is no reason why the results observed in the IAT should diverge from those observed in explicit ratings. On the other hand, from the point of view of dual-process performance models, if the outcome-density bias is due to reasoning processes that take place when elaborating the judgments, then there is no reason why the bias should also be observed in a reaction time measure that does not require those processes, such as the IAT.

Procedure

To the best of our knowledge, the study conducted by De Houwer and Vandorpe (2010) is the only published experiment that has used the IAT to measure causal learning effects. Therefore, except for minor changes, the procedure used in the present experiment was kept as similar as possible to that of De Houwer and Vandorpe (2010).

Contingency learning task and causal judgments

All the instructions were given in Dutch. The names of fictitious drugs Dugetil and Aubinol, counterbalanced across participants, played the role of cues X and Y. The names "Test environment 1" and "Test environment 2", counterbalanced, were used as contexts A and B. The Dutch words huidirritatie (skin irritation) and misselijkheid (nausea), presented in red color, played the role of O1 and O2, respectively.

A trial started with the presentation of the message "Chemical substance used in this test" and the name of the cue underneath this message at the top of the screen. In trials in which only the context was present, the words "no substance" appeared instead of the cue. After 2,000 ms, the message "Result:" and the name of the allergic reaction, "skin irritation" or "nausea", was added at the bottom of the screen for 3,000 ms, with the cues still visible. The inter-trial interval was 3,000 ms.

Given that the present design uses two different cues, X and Y, and two different contexts, A and B, we expected that participants might have some difficulties learning the target contingencies. We therefore simplified the task by dividing the general sequence of 80 trials into four blocks of 20 trials each. Within each block, participants only saw information about one cue and one context. Thus, participants received 2 blocks of 20 trials each with Cue X in Context A, and 2 additional blocks of 20 trials with Cue Y in Context B. For half of the participants, the four blocks provided information about cues X and Y following the order X-Y-X-Y and for the other half the order was Y-X-Y-X. Blocks 1 and 3 always took place in Test environment 1 and Blocks 2 and 4 always took place in Test environment 2.

Immediately, after the contingency learning task, half of the participants performed the IAT and then provided causal judgments, while the other half provided causal judgments first and then completed the IAT. Participants were asked to judge to what extent cues X and Y caused the "skin irritation" and the extent to which they caused the "nausea". They could enter these causal judgements using a scroll whose left extreme was labelled as "it is definitely not the cause" and whose right extreme was "it is definitely the cause". The scroll used by the participants did not show any numerical scale. However, the spatial position of their response was translated to a 0-100 scale for statistical analyses. Participants made judgments for each cue-outcome combination on a separate screen, with judgments about "nausea" always being made before judgments about "skin irritation". Which cue was presented first (cue X or Y) was counterbalanced for each participant.

Implicit Association Test

The IAT is built upon the idea that it is easier to categorize two stimuli using the same response if they are associated in memory. In our experiment, the IAT consisted of seven phases. During Phase 1, in each trial participants were asked to press one key if the

word presented on the screen was related to outcome O1 and a different key if the word presented on the screen was related to O2. The Dutch words for to itch (jeuken), to scratch (krabben), skin rash (huiduitslag), and macula (huidvlekken) were used as stimuli referring to the outcome category skin irritation. The Dutch words for vomit (braaksel), throw up (overgeven), ill (ziek), and nauseous (misselijk) were used as stimuli referring to the outcome category nausea. Participants responded to those two categories by using keys A and P on an AZERTY key board. Each outcome stimulus was presented four times, resulting in 32 trials.

During Phase 2, participants were asked to use the same keys, A and P, to report whether the word presented on the screen was the name of cue X or the name of cue Y. Unlike the case in standard IAT procedures, there was only one stimulus per cue category. As in De Houwer and Vandorpe (2010), to solve this problem, we presented each name in four different fonts (lower case Arial Black, upper case Arial Black, lower case Fixedsys, and upper case Fixedsys), resulting in eight different concept stimuli. Each cue stimulus was presented four times, resulting in 32 trials.

During Phases 3 and 4, participants had to perform both tasks at the same time: They had to press one key when the word presented on the screen was either related to O1 or the name of cue X and a different key when the word presented on the screen was either related to O2 or was the name of cue Y. Although Phase 3 and Phase 4 were identical, participants were told that Phase 3 was a practice phase and Phase 4 was a testing phase. In each phase, each outcome and cue stimulus was presented twice, resulting again in 32 trials per phase. If there is a strong association between X and O1 and an association between Y and O2, then we would expect participants' responses to be relatively fast in Phases 3 and 4 because the same response is required for cue X and O1, on the one hand, and for Y and O2, on the other.

Phase 5 was identical to Phase 2, except that the key assignments were reversed: The key previously used to respond to cue X was now used to respond to Y and the key previously used to respond to Y was now used to respond to X. As in Phase 2, each cue stimulus was presented four times, resulting in 32 trials. Finally, during Phases 6 and 7 participants had to press one key when the word presented on the screen was either related to O1 or the name of cue Y and a different key when the word presented on the screen was either related to O2 or was the name of cue X. As in the case of Phases 3 and 4, Phase 6 was presented to participants as a practice phase and Phase 7 was presented as a testing phase. Each of them consisted of 32 trials. If there is a strong association between X and O1 and an association between Y and O2, then we would expect participants to find this task relatively difficult, as it requires them to use the same response for stimuli that are not associated (X and O2, on the one hand, and Y and O1, on the other). Therefore, their responses should be slower in Phases 6 and 7 than in Phases 3 and 4.

Before each phase, participants were informed about the assignment of the different categories to the left and the right key. In order to reduce variance and keep the experimental conditions as constant as possible across participants, neither the assignment of outcome/cue categories to the two response keys nor the order of phases in the IAT was counterbalanced. The order of trials in each phase was determined randomly for each participant. On each trial, a word was presented at the center of the screen until a valid response was registered. If the response was correct, the next word appeared after 400 ms. If the response was incorrect, a red cross was presented for 400 ms, also at the center of the screen, and the next word was presented 400 ms after the red cross disappeared.

Results

Several algorithms have been proposed in the literature to score performance in the IAT. Perhaps the two most common measures are the logarithmic score (initially used by

Greenwald et al., 1998) and the D600 algorithm proposed by Greenwald, Nosek, and Banaji (2003) because of its seemingly higher correlation with self-report measures. For the analysis of the present experiment, we computed both of these dependent variables. According to the guidelines of Greenwald et al. (2003), the D600 measure uses the reaction times of all trials in Phases 3, 4, 6, and 7. The reaction time of incorrect responses is replaced by the mean reaction time in its block plus a 600 ms penalty. The score for the training phases (3 and 6) is computed as the average of block 6 minus the average of block 3 divided by the pooled standard deviation of the reaction times in those two blocks. The score for the testing phases is computed similarly with data from blocks 7 and 4. The final D600 score is the average of the training phases' score and the testing phases' score. We also computed for each participant the logarithmic IAT score used by Greenwald et al. (1998), which takes into account only data from the testing blocks (4 and 7), removing the responses from the two first trials and also incorrect responses and any response with latency below 300 or above 3000 ms. All reaction times are then $\log_{10}$ transformed, and the difference between blocks 7 and 4 is taken as the final dependent variable. Both IAT scores were computed in such a way that positive scores indicated that X-O1 and Y-O2 associations were stronger than X-O2 and Y-O2 associations. The mean IAT scores in each group are shown in Table 1. As can be seen, both types of IAT scores suggest that the target associations are stronger in Group 0.9 than in Group 0.5. Two t-tests showed, that the difference between both groups was significant for the logarithmic IAT score, $t(46) = -2.72$, $p < .01$, $d = 0.80$, but did not reach conventional levels of significance for the D600, $t(46) = -1.61$, $p = .114$, $d = 0.47$.

The IAT does not provide a measure of the absolute strength of associations, but a measure of the relative strength of some associations over others. On the other hand, causal judgments require participants to focus on specific cue-outcome associations. Given this

divergence, the direct comparison of IAT scores with causal judgments is unwarranted. In order to improve the comparability between IAT scores and causal judgments, we converted the latter to compute a measure of the relative strength of associations X-O1 and Y-O2 over associations X-O2 and Y-O1, more similar to the way the IAT is assumed to measure the relative strength of associations. Specifically, we subtracted each participant's ratings for associations X-O2 and Y-O1 from his/her ratings for associations X-O1 and Y-O2 (i.e., [XO1 + YO2] – [XO2 + YO1]). The means of the resulting dependent variable for each group are shown in Table 1. A t-test confirmed that judgments were larger for Group 0.9 than for Group 0.5, $t(46) = -6.84$, $p < .001$, $d = 2.02$.

To compare these explicit ratings with the logarithmic IAT scores we standardized both measures and submitted them to a 2 (Type of measure: IAT_{LOG} vs. Judgments) x 2 (Group: 0.5 vs. 0.9) ANOVA. This analysis yielded a main effect of Group, $F(1, 46) = 37.80$, $p < .001$, indicating a cue-density effect, and a significant Type of measure x Group interaction, $F(1, 46) = 4.08$, $p = .049$, indicating that the cue-density effect was somewhat larger in judgments than in the IAT_{LOG} scores. Given that both measures had been standardized, the main effect of Type of measure was logically nonsignificant, $F < 1$. A similar ANOVA using the IAT_{D600} instead of IAT_{LOG} scores yielded identical results: a main effect of Group, $F(1, 46) = 23.65$, $p < .001$, and a Type of measure x Group interaction, $F(1, 46) = 9.1$, $p < .005$, but no effect of Group, $F < 1$.

Table 2 shows the correlations between the three dependent measures: IAT_{D600} scores, IAT_{LOG} scores and converted judgments. As can be seen, all three measures were significantly correlated with each other. Taken together with the previous analyses, these correlations suggest that all these variables are measuring the same construct, although they might differ in their sensitivity to the cue-density manipulation.

Discussion

The results of our experiment show that it is possible to detect an outcome density effect both with causal judgments and with the IAT. These results offer no support for recent attempts to explain the outcome density bias and other biases in causality detection from a dual-process framework. Based on a dissociation between the biases observed in causal judgments and the absence of biases in the trial-by-trial predictions that participants made in a preparation similar to ours, Allan et al. (2005) and Perales et al. (2005) argued that their results could be best accommodated by assuming that participants are able to learn the contingencies in an unbiased manner but that their judgments are also influenced by deliberate reasoning processes that give rise to the bias. Following a Signal Detection Theory approach, they suggested that trial-by-trial predictions reflect participants' sensitivity to the cue-outcome contingency (d'), while causal judgments were also influenced by the participants' response criterion (β) that varied with outcome and cue density. In sum, Allan et al. and Perales et al. assumed that a dual-process mechanism is necessary to explain why the outcome-density effect is observed in some measures but not in others. From their point of view, only dependent measures that require a deliberative reasoning process would yield an outcome-density effect. Therefore, it is unlikely that their theoretical framework can accommodate the present results. It is unclear how or why the reasoning processes responsible for the change in a response criterion could affect performance in a task such as the IAT that does not involve a judgment about cue-outcome relations.

Our results also conflict with those of Ratliff and Nosek (2010) in the area of stereotype formation. As explained in the Introduction, Ratliff and Nosek conducted an experiment on illusory correlation in which participants viewed information about a majority group and a minority group. Although both social groups showed exactly the same proportion of desirable and undesirable traits, the participants tended to perceive the

majority group more positively than the minority group. However, this preference for the majority only arose in the explicit measures. An IAT designed to measure their relative preference for the majority group failed to find any effect. In this case, it is difficult to question the sensitivity of the IAT measure used in this study. Ratliff and Nosek showed that when there were real differences between the majority and the minority groups, the IAT was able to detect the difference in the participants' attitudes towards these groups. Furthermore, to preclude any limitation in statistical power, their Experiment 2 was conducted with a large sample gathered over the Internet with almost 900 participants. The divergence between our results and those of Ratliff and Nosek suggests that despite their surface similarity, the cue- and outcome-density biases explored in causal learning and the illusory correlations studied in stereotype formation research might in fact reflect different processes. Alternatively, the minor differences between the stimuli used in their IAT and the ones used in the present experiment might account for the different results obtained in both studies.

Although we observed a significant outcome-density effect on the logarithmic IAT scores, only a trend in the same direction was observed on the D600 IAT scores. Admittedly, the D600 has become more popular in recent research with the IAT, but there is no obvious reason why one should favor one scoring algorithm over the other. The calculation of these two scores differs in several respects. A major difference between the two is that the practice mixed blocks (i.e., Phases 3 and 6) are taken into account in the D600 but not in the logarithmic IAT score. The logarithmic IAT also excludes reaction times from incorrect responses. However, our inspection of the data suggests that the divergence between the D600 and logarithmic IAT scores is mainly due to the fact that reaction times are log converted in the calculation of the logarithmic IAT but not in the D600 score.[2] In fact, we observed that the crucial effect of group in the analysis of the

D600 scores became significant if all reaction times were first log converted, $t(46) = -2.04$, $p = .047$, $d = 0.60$. Log transforming RT has the advantage that it reduces the impact of extremely long RTs and hence normalizes the typically skewed RT distribution. The way in which the logarithmic IAT score is calculated is also more akin to standard practice in the experimental reaction time research than the D600 score. The recommendation of Greenwald et al. (2003) to use the D600 score was based exclusively on its capacity to register individual differences. While the validation of the logarithmic IAT score used discriminant validity as a criterion (i.e., whether the IAT measured something different from explicit measures), the validation of the D600 score was based on positive correlations with self-report measures (i.e., convergent validity; see Greenwald et al., 2003). However, there is currently no research on whether a D600 measure is also better at detecting effects of experimental manipulations. In any case, the fact that we observed a clearly significant, large sized outcome-density effect in the logarithmic IAT scores, a trend of the same effect in the D600 score (which became significant when log-transforming RTs before calculating the D600 score), and a significant correlation between both scores provides a strong basis for concluding that the outcome-density effect in IAT performance is genuine.

Although our results strongly suggest that a single-process account is able to account for the outcome-density effect, it is important to note that a contribution of performance-related processes to the outcome-density effect cannot be excluded. As shown above, the effect of the outcome density manipulation was somewhat larger on the participants' judgments than on their IAT scores. A potential explanation for this asymmetry is that learning processes (measured with the IAT) are responsible for part of the outcome density effect and that, additionally, performance related processes further enhance this effect in judgments. An alternative and more parsimonious interpretation of this result is that both

the IAT and judgments were mainly measuring learning processes, but the latter were more sensitive than the former. In any case, the fact that the outcome density effect was found in the IAT scores shows that at least part of the effect must be attributed to processes different than those involved in the production of a judgment.

Previous failures to detect the outcome- and cue-density biases might have been due to the use of an insensitive dependent measure. As explained above, both Allan et al. (2005) and Perales et al. (2005) used the trial-by-trial predictive responses of participants to compute a dependent measure that reflected their online perception of contingency between the cue and the outcome. However, if participants believe that there is a probabilistic relationship between the cue and the outcome, these discrete yes/no responses might not be an optimal way to express that knowledge. Furthermore, these predictions are made during the training phase, including therefore many measures from trials in which participants are still learning what the relationship between the cue and the outcome is.

In a related experiment, we recently tried to find out whether the cue-density bias could be measured by using a predictive question (Vadillo et al., 2011). Unlike Allan et al. (2005) and Perales et al. (2005), however, participants had to answer this predictive question at the end of the experiment (i.e., once the training phase was over and they presumably had learned the relationship between cue and outcome) and by means of a 0-100 rating scale (i.e., as opposed to yes/no discrete responses). Contrary to their results, we observed that under these conditions the cue-density bias was even stronger in the predictive question than in the causal ones.

Together with the present results, these findings suggest that there is nothing special in causal judgments that make them more sensitive than other measures to cue- and outcome-density biases in contingency learning. Indeed, both the cue- and the outcome-density effect can be observed successfully with any measure of the participant's

perception of the cue-outcome association, as long as the measure is valid and sensitive enough. Such a conclusion fits well with an association formation view that attributes illusory correlations such as the outcome-density effect to processes during learning. Whether or not performance-related factors enhance this effect remains an open question that future research should address. However, in the absence of additional evidence, single-process learning accounts seem to offer the most parsimonious explanation for the outcome-density bias.

References

Allan, L. G., & Jenkins, H. M. (1983). The effect of representations of binary variables on judgment of influence. *Learning and Motivation, 14,* 381-405.

Allan, L. G., Siegel, S., & Hannah, S. (2007). The sad truth about depressive realism. *Quarterly Journal of Experimental Psychology, 60,* 482-495.

Allan, L. G., Siegel, S., & Tangen, J. M. (2005). A signal detection analysis of contingency data. *Learning & Behavior, 33,* 250-263.

Alloy, L. B., & Abramson, L. Y. (1979). Judgements of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of Experimental Psychology: General, 108,* 441-485.

Barnes, P. M., Bloom, B., & Nahin, R. L. (2008). Complementary and alternative medicine use among adults and children: United States, 2007. *National Health Statistics Reports, 12.*

Blanco, F., Matute, H., & Vadillo, M. A. (in press). Interactive effects of the probability of the cue and the probability of the outcome on the overestimation of null contingency. *Learning & Behavior.*

De Houwer, J., & Vandorpe, S. (2010). Using the Implicit Association Test as a measure of causal learning does not eliminate effects of rule learning. *Experimental Psychology, 57,* 61-67.

Fiedler, K. (2000). Illusory correlations : A simple associative algorithm provides a convergent account of seemingly divergent paradigms. *Review of General Psychology, 4,* 25-58.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132,* 692-731.

Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reason in everyday life.* New York, NY: Free Press.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. K. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74,* 1464-1480.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85,* 197-216.

Hamilton, D. L., & Gifford, R. K. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology, 12,* 392-407.

Lilienfeld, S. O., Ammirati, R., & David, M. (2012). Distinguishing science from pseudoscience in school psychology: Science and scientific thinking as safeguards against human error. *Journal of School Psychology, 50,* 7-36.

López, F. J., Cobos, P. L., Caño, A., & Shanks, D. R. (1998). The rational analysis of human causal and probability judgment. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 314-352). Oxford: Oxford University Press.

Matute, H. (1995). Human reactions to uncontrollable outcomes: Further evidence for superstitions rather than helplessness. *Quarterly Journal of Experimental Psychology, 48B,* 142-157.

Matute, H. (1996). Illusion of control: Detecting response-outcome independence in analytic but not in naturalistic conditions. *Psychological Science, 7,* 289-293.

Matute, H., Yarritu, I., & Vadillo, M. A. (2011). Illusions of causality at the heart of pseudoscience. *British Journal of Psychology, 102,* 392-405.

Msetfi, R. M., Murphy, R. A., Simpson, J., & Kornbrot, D. E. (2005). Depressive realism and outcome density bias in contingency judgments: The effect of the context and inter-trial interval. *Journal of Experimental Psychology: General, 134,* 10-22.

Musca, S. C., Vadillo, M. A., Blanco, F., & Matute, H. (2010). The role of cue information in the outcome-density effect: Evidence from neural network simulations and a causal learning experiment. *Connection Science, 20,* 177-192.

Nahin, R. L., Barnes, P. M., Stussman, B. J., & Bloom, P. (2009). Costs of complementary and alternative medicine (CAM) and frequency of visits to CAM practitioners: United States, 2007. *National Health Statistics Reports.*

Perales, J. C., Catena, A., Shanks, D. R., & González, J. A. (2005). Dissociation between judgments and outcome-expectancy measures in covariation learning: A signal detection theory approach. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1105-1120.

Ratliff, K. A., & Nosek, B. A. (2010). Creating distinct implicit and explicit attitudes with an illusory correlation paradigm. *Journal of Experimental Social Psychology, 46,* 721-728.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-CenturyCrofts.

Shang, A., Huwiler-Müntener, K., Nartey, L., Jüni, P., Dörig, S., Sterne, J. A. C., Pewsner, D., & Egger, M. (2005). Are the clinical effects of homeopathy placebo effects? Comparative study of placebo-controlled trials of homeopathy and allopathy. *The Lancet, 366,* 726-732.

Shanks, D. R. (1985). Continuous monitoring of human contingency judgment across trials. *Memory & Cognition, 13,* 158-167.

Shanks, D. R. (1987). Acquisition functions in contingency judgment. *Learning and Motivation, 18,* 147-166.

Shanks, D. R. (1995). Is human learning rational? *Quarterly Journal of Experimental Psychology, 48A,* 257-279.

Shanks, D. R., & St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences, 17,* 367-447.

Shapiro, A. K., & Shapiro, E. (1997). *The powerful placebo: From ancient priest to modern physician.* Baltimore: Johns Hopkins University Press.

Sherman, J. W., Kruschke, J. K., Sherman, S. J., Percy, E., Petrocelli, J. V., & Conrey, F. R. (2009). Attentional processes in stereotype formation: A common model for category accentuation and illusory correlation. *Journal of Personality and Social Psychology, 96,* 305-323.

Vadillo, M. A., Musca, S. C., Blanco, F., & Matute, H. (2011). Contrasting cue-density effects in causal and prediction judgments. *Psychonomic Bulletin & Review, 18,* 110-115.

Vyse, S. (1997). *Believing in magic: The psychology of superstition.* New York: Oxford University Press.

Footnotes

[1] The experiment was conducted with an initial sample of 32 participants, which yielded a pattern of results virtually identical to the one reported in the present paper. Given that the D600 score failed to yield conclusive results, we added a smaller sample of 16 participants to get clearer results. The *p*-values of all the effects that were already significant with the original sample became lower after adding the 16 additional participants. In the original sample, the difference between both groups in the D600 score was nonsignificant, $t(30) = -1.08$, $p = .289$, $d = 0.39$, but the logarithmic score yielded a significant difference, $t(30) = -2.28$, $p < .05$, $d = 0.83$.

[2] The other important difference between the logarithmic score and the D600 score is that the former is computed just with trials from blocks 4 and 7. However, computing the D600 score just with blocks 4 and 7 did not make it more sensitive to our manipulation, $t(46) = -1.417$, $p = .163$, $d = 0.42$.

Author Note

MAV, NOC, and HM were supported by Grant IT363-10 from Departamento de Educación, Universidades e Investigación of the Basque Government and Grant PSI2011-26965 from Ministerio de Ciencia e Innovación. JDH and MDS were supported by Methusalem Grant BOF09/01M00209 of Ghent University. NOC was supported by fellowship BFI09.102 from the Basque Government. Preparation of this manuscript was made possible by a Senior Visiting Postdoctoral Fellowship of the Research Foundation – Flanders (FWO) that was awarded to MAV for a visit to the laboratory of JDH. We would like to dedicate this paper to the memory of Lorraine Allan, whose work inspired us in this and so many other studies. Correspondence concerning this article should be addressed to Miguel A. Vadillo, Division of Psychology and Language Sciences, University College London, 26 Bedford Way, London WC1H 0AH, United Kingdom. E-mail: m.vadillo@ucl.ac.uk

Table 1

Descriptive statistics

	Group 0.5		Group 0.9	
Dependent Measure	Mean	SEM	Mean	SEM
IAT_{D600}	0.290	0.068	0.456	0.076
IAT_{LOG}	0.014	0.006	0.040	0.007
Judgments	3.333	8.521	115.666	14.033

Table 2

Correlations between dependent measures

	IAT_{D600}	IAT_{LOG}
IAT_{LOG}	.499 (.000)	-
Judgments	.305 (.035)	.299 (.039)

Note. Numbers outside brackets refer to the correlation coefficient *r* and numbers between brackets represent the corresponding *p* values.

Figure Captions

Figure 1. Design summary of the experiment. Numbers within each contingency table represent the number of trials of that type seen by participants in each group. Context A and Context B were two different Testing environments for chemical substances X and Y, counterbalanced. O1 and O2 were skin irritation and nausea, counterbalanced.

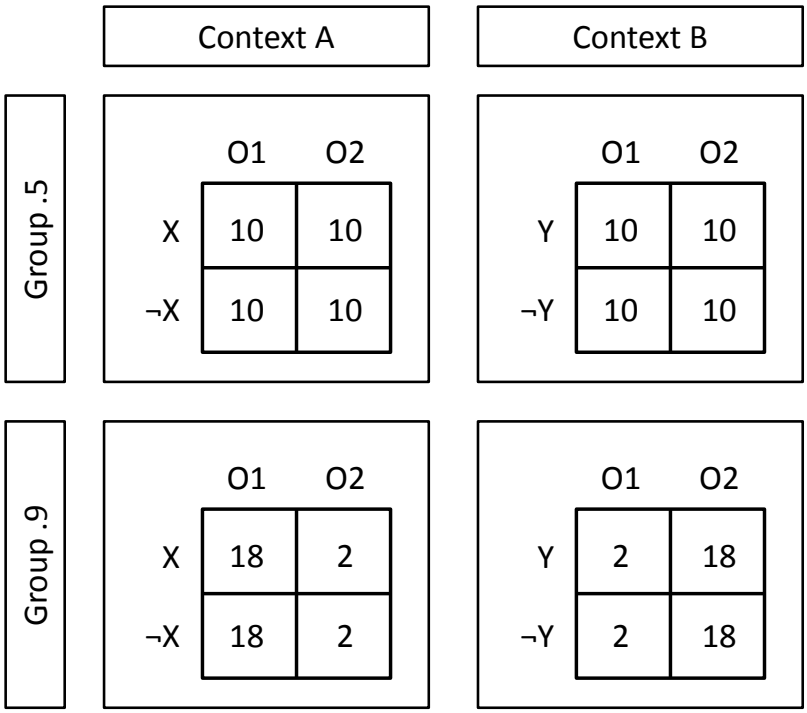
Context A
Context B
Group .5
O1
O2
X
10
10
¬X
10
10
Y
10
10
¬Y
10
10
Group .9
18
2
2
18

Figure #1