The Role of Causal Attributions in Observational Conditioning

Sarah Kasran[1]*, Sean Hughes[1], Jan De Houwer[1], and Tom Beckers[2]

[1] Department of Experimental Clinical and Health Psychology, Ghent University, Ghent, Belgium

[2] Leuven Brain Institute and Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium

In press. Collabra: Psychology.

Author Note

Correspondence concerning this article should be addressed to Sarah Kasran, Department of Experimental Clinical and Health Psychology, Ghent University, Henri Dunantlaan 2, 9000 Gent, Belgium. Email: Sarah.Kasran@UGent.be.

Abstract

Our behavior towards a stimulus can change as a result of observing a regularity between that stimulus and someone else's emotional reaction, a type of social learning referred to as observational conditioning. We explore the idea that causal attributions (i.e., the extent to which the observer attributes the model's reaction to the stimulus) play an important role in observational conditioning effects. In three experiments (total $N = 665$), participants watched videos in which one cookie was followed by a positive reaction and another cookie was followed by a negative reaction, after which their own evaluations of each cookie were measured via self-reports and an implicit association test (IAT). Critically, we manipulated whether the observed reactions were high or low in terms of distinctiveness (Experiments 1a and 1b) or consensus and consistency (Experiment 2). These three variables are known to influence stimulus attributions and were therefore predicted to moderate observational conditioning effects. In line with our predictions, high distinctiveness (Experiments 1a and 1b) and high consensus and consistency (Experiment 2) both resulted in larger observational conditioning effects, with one exception: high distinctiveness did not lead to larger changes in automatic evaluations (i.e., IAT effects). Taken together, our findings suggest that causal attributions play an important role in observational conditioning. We outline more elaborate analyses of the attributional processes that are involved and suggest potential future directions for research on observational conditioning.

Keywords: observational conditioning; social learning; evaluative learning; attribution; causal reasoning

The Role of Causal Attributions in Observational Conditioning

Like many species, humans can learn by observing the behavior of others. This is a powerful ability, one that allows us to quickly adapt to the environment without the need to personally come into contact with aversive stimuli. To illustrate, imagine that you are passing a newly opened restaurant, glance through the window, and see a woman taking a bite of her food. Upon tasting the food, her facial expression indicates clear disgust. After observing this sequence of events, you may expect the food in this restaurant to taste horrible and avoid visiting it yourself in the future.

However, now imagine that you actually know this woman and are aware that she is a very picky eater who dislikes many types of food. This information may affect the conclusions you draw: you may now be inclined to believe that the disgusted reaction was more revealing of the woman's general preferences than of the food being served. If so, you might rely less on what you saw when forming your own opinion of the restaurant and its offerings. Now imagine a different scenario where you have previously read several negative reviews about the restaurant online. In this case, not just this one woman but many people apparently dislike the restaurant's food. This information may lead you to a different conclusion: that the woman's disgusted reaction *was* due to the food, and that a negative opinion of the restaurant and its food is therefore justified.

The above examples illustrate two important points. The first is that observing another person's reaction to a stimulus can influence how much you yourself like it. As we will see, evidence indeed indicates that behavior towards a stimulus can change as a function of observing another person's emotional reaction to it (a type of social learning known as *observational conditioning*). The second point centers on the idea that certain types of information (e.g., information about the woman's general preferences or about other people's reactions to the food) are likely to influence your judgment about what caused the observed

reaction. This kind of judgment is usually referred to as a *causal attribution*: you attribute a given event (e.g., someone else's behavior) to a specific cause.

In this paper we examine whether causal attributions play an important role in observational conditioning. Specifically, we explored the idea that the size of observational conditioning effects would depend on the extent to which the observer is likely to attribute the observed reaction to the stimulus. We will begin by discussing research on observational conditioning and then turn to an influential theory of how people arrive at causal attributions. Next, we will apply this theory to observational conditioning to formulate specific predictions about variables that may strengthen or weaken observational conditioning effects.

Observational Conditioning

Mineka and colleagues introduced the term "observational conditioning" in a series of studies with rhesus monkeys who were found to exhibit fear of a toy snake after having observed another monkey (a "model") reacting fearfully to it (Cook et al., 1985; Mineka et al., 1984). Using classical conditioning terminology, they referred to the toy as a conditioned stimulus (CS), the model's emotional reaction as an unconditioned stimulus (US), and the resulting change in the observer's behavior towards the CS as an observational conditioning effect. Observational conditioning effects can thus be seen as a subtype of social learning effects that involves changes in behavior due to observing a model's reaction to a stimulus.

Numerous papers have reported evidence for observational conditioning in humans (for reviews see Askew & Field, 2008; Debiec & Olsson, 2017; Olsson & Phelps, 2007). Children readily learn to fear and avoid a new toy or an unknown animal based on seeing a model react fearfully in its presence (e.g., Broeren et al., 2011; Gerull & Rapee, 2002). Adults also show fear responses to a stimulus after they have seen others express pain or discomfort upon encountering it (e.g., Olsson et al., 2007; Szczepanik et al., 2020). In

contrast, observing positive or calm reactions to stimuli can decrease or prevent fear (e.g., Egliston & Rapee, 2007; Golkar & Olsson, 2016).

Importantly, observational conditioning is not limited to fear learning. Research indicates that people's likes or dislikes can also be influenced by observing the reactions of others. For instance, after observing a model react negatively while tasting one drink, children reported liking this drink less than a second drink that was followed by a neutral reaction (Baeyens et al., 1996, 2001). Likewise, both adults and children are more inclined to like a novel individual if a model shows positive nonverbal behavior while interacting with this individual than when the model shows negative nonverbal behavior (Castelli et al., 2008, 2012; Skinner et al., 2020; Skinner & Perry, 2020).

Cognitive theorizing on observational conditioning tends to center on two types of mental explanations. One possible explanation is that observational conditioning relies on associative processes (Askew & Field, 2008; Baeyens et al., 2001; Field, 2006; Mineka & Cook, 1993; Olsson & Phelps, 2007). According to this explanation, observing a stimulus (CS) together with a model's reaction (US) results in the formation of an association between the mental representations of those stimuli. When the observer later encounters the CS, the resulting activation of the CS representation spreads to the US representation, eliciting an emotional response similar to the one shown by the model. A second explanation that has been proposed is that observational conditioning effects are mediated by propositional and inferential processes (see Baeyens et al., 2001; Mineka & Cook, 1993; for a similar proposal concerning conditioning in general, see De Houwer, 2009; Mitchell et al., 2009). According to this idea, observers infer the stimulus' properties (e.g., how negative it is) based on the model's reaction. Critically, this inference is assumed to depend on various premises which need to be considered as true by the observer (e.g., that the model dislikes the stimulus, that the observer is exposed to the same stimulus, and that what applies to the model also applies

to the observer; Baeyens et al., 2001). Moreover, inferences rely on propositions, which – unlike simple associations – can encode the specific relation between events (e.g., whether one event *causes* or *prevents* another event). From these assumptions a number of predictions can be derived about the factors that may moderate observational conditioning, some of which diverge from predictions of associative accounts. For instance, observational conditioning effects should be influenced by beliefs about whether the observer is exposed to the same stimulus as the model, as well as by information specifying how the stimulus and the model's reaction are related (e.g., that the reaction of the model is genuine or faked). The fact that such findings have emerged (e.g., Baeyens et al., 2001; Kasran et al., 2022b) seems to challenge a purely associative explanation and highlights important moderators of observational conditioning effects.[1]

If inferences are indeed involved in observational conditioning, one type of inferences may be especially relevant: causal inferences. That is, observational conditioning may depend heavily on whether the observer believes that the model's behavior is actually caused by the stimulus (and not by something else, such as properties of the model or the situation). In this paper, we extend prior research by focusing on this specific type of inference. The question then becomes: how do we infer the cause(s) of other people's behavior? We will now turn to research on causal attributions, which aims to provide answers to this exact question.

Attribution Theories

[1] We should note at this point that many predictions that can be derived from a propositional account with regard to potential moderators of observational conditioning may also seem in line with (some) associative theories. For example, information specifying that the model's reactions were faked could be predicted to reduce observational conditioning effects if one assumes that it decreases the observer's attention to those reactions (reducing the opportunity for association formation) and/or that it influences the nature of the US representations. However, even though such predictions can be accommodated by associative theories, a propositional account more readily brings them to the fore because of its focus on the importance of beliefs. This is also the case for the current research: exploring implications of a propositional perspective led us to examine the impact of moderators that should affect beliefs about causality (i.e., attributions) – moderators that might tend to be overlooked from an associative perspective.

Broadly speaking, attribution theories are theories about how people arrive at causal explanations (i.e., attributions) for events, including (but not limited to) the behavior of other people (Kelley & Michela, 1980). Considerable knowledge has been gained with regard to the variables and conditions that influence which attributions people make.

In this paper, we focus primarily on a highly influential theory that was originally formulated by Kelley (Kelley, 1967, 1973). Central to this theory is the covariation principle: people attribute an event to a possible cause that is perceived to covary with it. Kelley suggested that for most events three candidate causes are considered: the person, the entity (which we will here refer to as the "stimulus"), and the time or moment (Kelley, 1973). For instance, if the event is a woman showing a disgusted reaction when tasting a dish in a restaurant, we could attribute this reaction to something about the woman showing the reaction (person), something about the dish (stimulus), something about this particular occasion (time), or a specific combination of any of these possible causes. So how does one decide between these candidate causes? Kelley's approach and the research it inspired focused mostly on three types of information that people may use to do so: consensus information (i.e., whether *other people* behave the same way to this stimulus), distinctiveness information (i.e., whether this person behaves the same way to *other stimuli*), and consistency information (i.e., whether this person behaves the same way to this stimulus *at other times*). Going back to our example, knowing that the woman is a picky eater (i.e., also dislikes many other types of food) can be seen as *low distinctiveness* information. In contrast, reviews indicating that other people also dislike the restaurant's food can be seen as *high consensus* information. Finally, if we had told you that the woman actually did enjoy the restaurant's food on other occasions, this could be seen as *low consistency* information.

Kelley predicted that specific patterns of these three variables – consensus, distinctiveness, and consistency – would result in specific attributions (Kelley, 1967). For

example, high consensus, high distinctiveness, and high consistency of a behavior would induce people to attribute that behavior to the stimulus, while low consensus, low distinctiveness, and high consistency of a behavior would induce them to attribute it to the person. These predictions were confirmed in empirical studies where participants were asked to judge the cause(s) of a given behavior (e.g., "John laughs at a cartoon") in the presence of sentences which reflected either a high or a low level of each variable (e.g., high consensus: "Almost everyone who sees the cartoon laughs at it"; low consensus: "Hardly anyone who sees the cartoon laughs at it"; example adapted from McArthur, 1972). The eight different configurations created by combining high versus low levels of the three variables were largely found to result in the expected attributions (e.g., Hewstone & Jaspars, 1987; McArthur, 1972). In addition, each variable was found to have an individual impact on attributions (McArthur, 1972; Ruble & Feldman, 1976), even when presented in an incomplete configuration or on its own (Orvis et al., 1975). Most importantly for the current purposes, participants were most likely to attribute a given behavior to the stimulus in the presence of high consensus, high distinctiveness, and/or high consistency information. In the current research, we therefore investigated the impact of these variables on observational conditioning.

As a caveat, we should note that certain aspects of Kelley's theory and the subsequent studies have been criticized. However, the three variables generally do have robust and well-established effects (especially for stimulus attributions), and some of the most important criticisms that have been voiced – such as the theory being unlikely to hold for voluntary actions (e.g., Malle, 2011; Zuckerman, 1978) or for highly scripted or normative behaviors (e.g., Hilton & Slugoski, 1986) – do not seem to apply to the type of situation that we are interested in. Nevertheless, we will discuss a number of other qualifications of this approach and their potential relevance to our research in the general discussion. In addition, while there

are other well-known attribution theories, some of them are not applicable to the present context because they mostly focus on person perception (e.g., the theory of correspondent inferences; Jones & Davis, 1965), whereas we are interested in *stimulus* attributions. As Kelley's approach provides clear predictions for stimulus attributions, it seems a good starting point for exploring whether attributions guide observational conditioning.

Finally, others have drawn attention to important parallels between research on conditioning (in both humans and animals) and the attribution literature (e.g., Alloy & Tabachnik, 1984; Eelen, 2018; Van Overwalle, 1996). As such, we are not the first to suggest that factors that influence attributions – more specifically, whether a US is attributed to a CS – will influence CS-US covariation judgments or the effect of CS-US pairings on the response to a CS. However, because historically much of the focus of the attribution literature was on explaining *people's behavior*, we believe that the knowledge gained from this literature may prove to be particularly relevant and applicable to *social* learning, which always involves learning from the behavior of a model.

The Current Research

With the above in mind, we set out to explore the role of causal attributions in observational conditioning. Specifically, we investigated whether consensus, distinctiveness, and consistency – three variables that have been shown to influence stimulus attributions – would also influence observational conditioning effects. Note that there is some evidence to suggest that other types of social learning may be sensitive to one or more of these variables. For example, even two-year-old children were more likely to copy a reinforced action that had been demonstrated by three different models than an action that had been demonstrated three times by one model (Haun et al., 2012), which could be seen as an effect of consensus. The children were also more likely to copy an action that had been demonstrated three times by the same model than an action that had been demonstrated once by another model, which

could be seen as a consistency effect (although the repeated demonstrations might just have biased the children's attention to the location where the former action could be performed). Evidence for an impact of consensus has also been found for gaze-induced effects on liking (i.e., the finding that a stimulus that is looked at by others is liked better than a stimulus that is looked away from). In one study, this effect was found only when seven faces were shown gazing at or away from stimuli, not when only one face was shown (Capozzi et al., 2015). In contrast, one observational conditioning study reported no evidence that calm responses shown by two models were more effective in preventing subsequent observational fear conditioning than calm responses shown by one model (Golkar & Olsson, 2016), although participants might of course not consider two models to reflect high consensus. In sum, although these studies suggest that causal attributions may play a role in observational conditioning, the available evidence is scarce and inconclusive. A more systematic investigation of several types of information (consensus, distinctiveness, consistency) in the context of observational conditioning is still lacking.

We tested our predictions in the context of evaluative responses acquired via an observational conditioning procedure, which we will refer to as observational evaluative conditioning or OEC (Baeyens et al., 1996, 2001). We employed a procedure that we have used in previous studies (Kasran et al., 2022b, 2022a), which consisted of showing participants videos of models tasting novel cookies (CSs). One cookie was followed by a positive nonverbal reaction, while another cookie was followed by a negative nonverbal reaction (we will refer to the first cookie as the CS_{pos} and the second cookie as the CS_{neg}). Afterwards, we measured participants' evaluative responses to the CSs in two ways. First, self-reported liking was assessed by asking them how much they would expect to like each cookie. Second, to avoid basing our conclusions only on self-reports, we also assessed participants' evaluations of the CSs using a variant of the Implicit Association Test (IAT;

Greenwald et al., 1998), a reaction time task that required participants to categorize the two cookies using the same set of responses that they had to use to categorize positively and negatively valenced words. The speed with which they could do so was taken as an index of how positively or negatively they evaluated each cookie. Given the nature of this task, evaluations measured within the IAT are usually considered to be more automatic (i.e., measured under conditions assumed to be suboptimal for cognitive processing, such as time pressure; see Moors & De Houwer, 2006). OEC effects were predicted to emerge on both measures. Specifically, we expected that participants would (a) rate the CS_{pos} more favorably than the CS_{neg} and (b) respond faster on IAT trials that required categorizing the CS_{pos} with the same key as positive words or the CS_{neg} with the same key as negative words, than on trials that required the opposite combination of responses.

Crucially, we investigated whether these OEC effects were affected by manipulations designed to influence stimulus attributions (i.e., the extent to which the modelled reactions would be attributed to the cookies). In Experiments 1a and 1b, we manipulated *distinctiveness* by providing information about the model's general attitudes towards other cookies. As stimulus attributions were expected to be stronger when distinctiveness was high, we predicted that participants who received high distinctiveness information would show larger OEC effects than participants who received low distinctiveness information. In Experiment 2, we employed a different manipulation for influencing stimulus attributions by providing *consensus* and *consistency* information. Stimulus attributions were expected to be stronger when participants received high consensus and high consistency information than when they received low consensus and low consistency information. Therefore, we predicted that OEC effects would be larger in the former condition than in the latter. In addition, we explored whether stimulus attributions and OEC effects could be reduced further by including an *explanation* for why consensus and consistency were low.

All experiments were conducted after obtaining advice and approval from the Ethical Committee of the Faculty of Psychology and Educational Sciences at Ghent University (application number 2018/53). We report how we determined our sample size, all data exclusions, all manipulations, and all measures. Prior to data collection, we pre-registered the hypotheses, planned sample size, procedural details, and planned analyses on the Open Science Framework (Experiment 1a: https://osf.io/cwksb/; Experiment 1b: https://osf.io/ux2rq/; Experiment 2: https://osf.io/v8q6a/). These pre-registrations were followed unless otherwise specified. With the exception of the videos, which have been replaced by anonymized versions because we did not have consent from the actors to publish them in their original form, all research materials are available on the OSF page (https://osf.io/ghd7f). Raw data, processed data, and all R code used for data processing and analysis are also available on the OSF page.

Experiment 1a

In Experiment 1a we investigated the impact of distinctiveness on OEC effects. Participants watched a first female model react positively to one cookie (CS_{pos}) and a second female model react negatively to another cookie (CS_{neg}). The cookies were referred to by fictional names (“Empeya” and “Plogo”). Prior to this observation phase, we manipulated the distinctiveness of both reactions by providing information about each model’s more general attitudes towards cookies. Participants in the *high distinctiveness* condition were told that the first model disliked most types of cookies while the second model liked most cookies. The models’ reactions during the observation phase were therefore high in distinctiveness (i.e., different from their alleged reactions to other stimuli of the same category). In contrast, participants in the *low distinctiveness* condition were told the exact opposite, namely that the first model liked most types of cookies while the second model disliked most cookies. The

reactions during the observation phase were therefore low in distinctiveness (i.e., similar to the models' reactions to comparison stimuli).

We predicted that participants in both conditions would evaluate the CS_{pos} more positively than the CS_{neg}, as reflected by their evaluative ratings as well as by their IAT performance (i.e., we expected to find OEC effects on both evaluative measures). We also predicted that these OEC effects (both in terms of the ratings and the IAT) would be smaller in the low distinctiveness condition than in the high distinctiveness condition.

Method

Participants and Design

In order to have 90% power to detect a medium-sized effect of distinctiveness in an ANOVA, a minimum sample size of $n = 171$ was required. Estimating 10% exclusions based on prespecified IAT performance criteria, we planned to recruit participants until we had complete and useable data for $n = 190$ (participants who provided incomplete data, who restarted and encountered different versions of the manipulation, or who reported technical issues were replaced during data collection). Participants were recruited via Prolific Academic (https://www.prolific.co/) using the following criteria: they were between ages 18 and 50, had indicated English as their first language, had successfully completed at least one Prolific study, had a study approval rate of at least 80%, and had not completed any previous studies from our lab. After excluding incomplete entries ($n = 7$), entries from participants who restarted ($n = 6$), one entry not linked to any Prolific ID ($n = 1$), and entries from participants who reported technical issues ($n = 2$), the sample consisted of 191 participants with complete and useable data (80 men, 111 women; $M_{age} = 29.99$, $SD_{age} = 8.85$; one more than planned because one participant provided useable data but timed out, which led to another participant being recruited automatically).

We employed a between-subjects design with two levels for distinctiveness: high distinctiveness and low distinctiveness. In addition, we wanted to counterbalance stimulus assignment (whether Empeya or Plogo served as the CS_{pos}), model assignment (whether Model A or Model B was shown reacting to the CS_{pos}), task order (whether participants first completed the ratings or the IAT), and IAT block order (whether participants first completed the learning-compatible or the learning-incompatible block of the IAT) across participants. However, because assignment to the different cells of the design was random, the sample was quite unbalanced for some of these factors (in subsequent experiments we changed the assignment strategy in order to address this issue).

Materials

Videos. Each video showed a female model taking a round cookie from a plate, taking a bite, and displaying a positive or negative reaction for approximately five to six seconds (example videos of a model who consented to video publication can be viewed at https://osf.io/zx4j8/). A label which was used to show the name of the cookie was placed next to the plate and clearly visible in the video. In preparation of the experiment, sixty videos of positive and negative reactions by three different models were rated in terms of valence and believability by separate samples of participants (one sample rated ten positive reactions from each model while another sample rated ten negative reactions from each model). Pre-rating materials and data are available at https://osf.io/cwdkr/. Based on these pre-ratings we selected four clearly valenced and sufficiently believable videos: one video of Model A showing a positive reaction, one video of Model B showing a positive reaction, one video of Model A showing a negative reaction, and one video of Model B showing a negative reaction. The two positive reactions did not differ from each other in terms of valence, $t(49) = 0.53$, $p = .60$, or believability, $t(49) = -0.80$, $p = .43$, and neither did the two negative videos (valence: $t(49) = 0.26$, $p = .80$; believability: $t(49) = 0.14$, $p = .89$). These four videos were

then edited to vary the name on the label (Empeya vs. Plogo) so that we could counterbalance stimulus assignment, resulting in eight videos in total. Finally, a pilot study confirmed that the selected videos led to clear OEC effects in terms of evaluative ratings and IAT performance (pilot materials and data are available at https://osf.io/bnygr/).

IAT. Five versions of each CS name (in lower- or uppercase and in regular, bold, or italic font) served as target stimuli in the IAT. The names of the two CSs ("Empeya" and "Plogo") served as labels for categorizing the target stimuli. As valenced stimuli, we used five positive (Tasty, Delicious, Nice, Enjoyable, and Wonderful) and five negative (Disgusting, Awful, Horrible, Unappetizing, and Repulsive) adjectives conceptually related to food. Based on the pilot study mentioned above, these stimuli were considered suitable for an IAT measuring evaluations of cookies. The phrases "I like" and "I dislike" were presented as labels for categorizing the valenced stimuli.

Procedure

The experiment was programmed in lab.js (Henninger et al., 2021) and hosted via a Ghent University server, enabling participants to complete the experiment via their browser. After providing informed consent and demographic information, participants were told that we were working with a company that was currently testing different recipes for new types of cookies, and that we had recorded videos of some people who were asked to try samples of these cookies and to show us how they felt about them. We also mentioned that we would mainly ask questions about two of the cookies, one called Empeya and another called Plogo (each name corresponding to a specific cookie recipe). Participants then read the distinctiveness information, watched the OEC videos, completed the evaluative measures, and answered a number of exploratory questions.

Distinctiveness Manipulation. Participants were informed that we had asked the models about some of their more general food-related attitudes. They were shown a picture of

one of the models and told the following: “The person below reported that in general, she is a very picky eater and dislikes almost all types of cookies.” On the next page, participants were shown a picture of the other model and told the following: “When asked about some of her more general food-related attitudes, this person reported that in general, she is not at all a picky eater and likes almost all types of cookies.” Which model was pictured on the first page and which on the second depended on the distinctiveness manipulation. In the high distinctiveness condition, the first page pictured the model who would be seen reacting positively to the CS_{pos} during the subsequent OEC phase (see below), while the second page pictured the model who would be seen reacting negatively to the CS_{neg}. In other words, the model who reacted positively was described as disliking most cookies whereas the model who reacted negatively was described as liking most cookies. In the low distinctiveness condition, the pictures were switched so that the model who would be seen reacting negatively was described as disliking most cookies while the model who would be seen reacting positively was described as liking most cookies. A small pilot study suggested that this manipulation of distinctiveness would likely influence attributions in the expected manner (pilot materials and data are available at https://osf.io/egp65/).

To ensure that participants read and processed this information, it was followed by a manipulation check that required participants to report what we had told them about each model. Only if participants correctly indicated that the first model generally disliked most cookies and that the second model generally liked most cookies could they proceed to the next phase; if not, they were required to re-read the information and complete the manipulation check until they answered both questions correctly.

OEC Phase. Participants were told that they would watch videos of people eating the Empeya and Plogo cookies. Each individual participant was shown only two of the eight videos: a video in which one model reacted positively to the CS_{pos} and a video in which the

other model reacted negatively to the CS_{neg}. Which model (Model A or Model B) and which cookie name (Empeya or Plogo) was included in the former video and which in the latter depended on stimulus assignment and model assignment. Similar to our previous studies with this type of task, both videos were shown three times in a random order with an inter-trial-interval of two seconds (Kasran et al., 2022a, 2022b).

Evaluative Ratings. Participants were asked to answer six questions (three per CS), which were presented on individual pages and in a random order. Specifically, they indicated on scales from -4 to +4 how pleasant or unpleasant they thought they would consider the CS to be (from *very unpleasant* to *very pleasant*), how much they thought they would like the CS (from *I would dislike it very much* to *I would like it very much*), and how good or bad they thought they would consider the CS to be (from *very bad* to *very good*). Zero was explicitly indicated as *neutral*.

IAT. On each IAT trial, a stimulus was presented on screen and participants used the D and K keys to categorize it as quickly as possible based on labels at the top left (D) and top right (K) of the screen. On some trials participants had to sort versions of the names "Empeya" and "Plogo" into these two categories and received error feedback (a red "X" presented for 200 ms) after incorrect responses. On other trials participants had to classify valenced adjectives in terms of whether they referred to something they liked or something they disliked. Because the instructions referred to subjective liking, no error feedback was presented on these trials. Note that the labels "I like" and "I dislike" and the omission of error feedback on these trials correspond to what is typically called a personalized IAT (Olson & Fazio, 2004), although we used adjectives rather than nouns so that the stimuli could be clearly food-related.

The IAT consisted of 180 trials in total, which were divided into seven blocks. In Block 1 (20 trials) participants sorted Empeya and Plogo into their respective categories. In

Block 2 (20 trials) participants had to sort the adjectives in terms of whether they referred to something they liked or disliked. These two blocks allowed participants to practice the initial response mappings. In Block 3 (20 trials) the two trial types were then combined: on some trials participants had to sort the CSs into their categories, whereas on other trials they had to sort the adjectives in terms of whether they referred to something liked or something disliked. Block 4 (40 trials) also combined these two trial types. Block 5 (20 trials) was again a practice block in which participants sorted the CS names, but with the opposite response mapping as before. Finally, in Block 6 (20 trials) and Block 7 (40 trials) the two trial types were once again combined but with the new response mapping for the CSs. During each block, trials were presented in a random order and the relevant labels remained on screen.

Block order was varied across participants. This meant that for some participants the initial response mappings were *compatible* with what they had observed during the OEC phase (i.e., they started by sorting the CS_{pos} with the same key as positive adjectives and the CS_{neg} with the same key as negative adjectives), whereas for others the initial response mappings were *incompatible* with what they had observed (i.e., they started by sorting the CS_{pos} with negative adjectives and the CS_{neg} with positive adjectives).

Exploratory Questions. Participants answered a number of exploratory questions. They were asked what type of reaction they had observed after each CS (contingency memory) and what they had been told about the models' general preferences (manipulation memory). Unless they reported not remembering the reactions (n = 1 for each reaction), they were also asked how distinctive (i.e., different from how the model usually reacted to cookies) they considered them to be on a scale from 1 to 9 (perceived distinctiveness). Importantly, to assess attributions, participants were asked to indicate for each reaction on scales from 1 to 9 to what extent they thought it had been due to something about the specific cookie (stimulus attribution), something about the specific person (person attribution), or

something about their specific combination (stimulus-person attribution), as well as to write down any other causes they had considered. In addition, we asked a number of open-ended questions about their thoughts during the videos, whether the distinctiveness information had influenced those thoughts, and how they had formed their opinions of the two CSs. These questions were included for purely exploratory reasons and will not be discussed further. We also asked to what extent their ratings and IAT performance had been based only on how they really felt (for the ratings) or only on being quick and accurate (for the IAT), on trying to behave in line with expectations (demand compliance), or on trying to resist behaving in line with expectations (reactance). Finally, we assessed hypothesis awareness: after two open-ended questions about the perceived goal of the experimenters, they were informed about our main hypothesis (i.e., that we expected the impact of the videos on their own opinions to depend on the distinctiveness information they had received) and indicated whether they had been aware of it or not. After reporting if they had encountered any technical issues, they were thanked and debriefed.

Results

Data Preparation

Based on prespecified criteria, we excluded participants who made more than 30% errors across the IAT (n = 2), who made more than 40% errors on any of the combined blocks (n = 8), or who completed more than 10% of trials faster than 300 ms (n = 1). The final sample consisted of 180 participants (76 men, 104 women; M_{age} = 29.85, SD_{age} = 8.63).

Evaluative ratings were averaged to create a mean rating for the CS_{pos} and a mean rating for the CS_{neg}. The CS_{neg} rating was then subtracted from the CS_{pos} rating to calculate a mean rating difference, such that a larger difference indicated a stronger preference for the CS_{pos} over the CS_{neg}. In addition, IAT reaction times were used to calculate scores in line with the D4-algorithm (Greenwald et al., 2003), in such a way that larger IAT scores

reflected a more positive evaluation of the CS_{pos} relative to the CS_{neg}. Note that we stated in our pre-registration that we would calculate IAT scores using the D1-algorithm. However, this algorithm does not include a specific treatment of trials on which errors were made, which is only appropriate if participants have to correct their error. As this was not the case in our task, using the D4-algorithm (which does include an error treatment) was more appropriate. We therefore report the results for the D4-score, but note if any of the main results diverged when the D1-score was used (which was not the case in Experiment 1a).

Data Analysis

Analytic Strategy. To test whether participants showed OEC effects in both conditions, we ran analyses of variance (ANOVAs) on the rating differences and IAT scores which included the counterbalanced variables as factors and thus reflected the structure of the design (which was important given that the sample was somewhat unbalanced). For both dependent variables and both conditions, we tested whether the intercept (representing the mean across the different cells) differed significantly from zero. Next, to test whether OEC effects were moderated by distinctiveness, we conducted ANOVAs which also included distinctiveness as a factor and tested for a main effect. We additionally checked whether the results were influenced (a) when participants who did not correctly remember the contingencies were excluded, (b) when participants who did not correctly remember the manipulation were excluded, or (c) when participants who reported demand compliance or reactance (i.e., who selected the midpoint or higher on those questions) were excluded. All other analyses were not pre-registered and were purely exploratory.

All hypothesis tests were conducted at the $\alpha = .05$ significance level. 95% confidence intervals are reported for Cohen's *d* and 90% confidence intervals are reported for η^2_p. We also report Bayes Factors (BF_{10}) which represent the probability of the alternative hypothesis compared to the null hypothesis given the observed data, using the calculations implemented

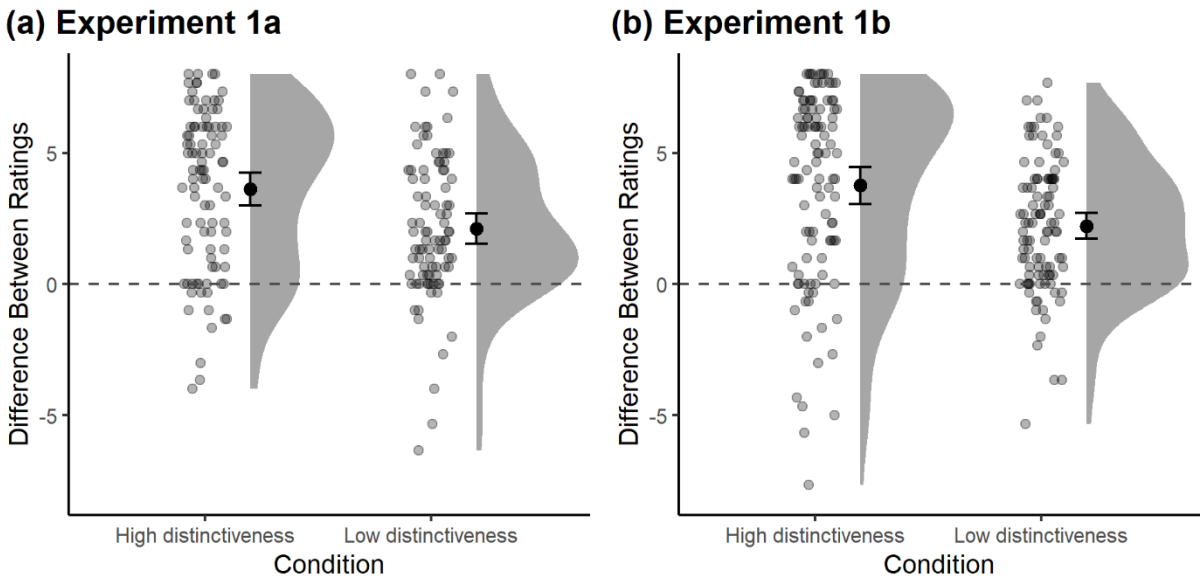
by the BayesFactor (Rouder et al., 2009, 2012) and bayestestR (Makowski et al., 2019) packages in R. All BFs reported for ANOVAs are "inclusion BFs" which reflect the evidence for including a term in the model based on model comparison between "matched" models (i.e., models that do not include any interactions with the term of interest but that do include the underlying main effects if the term of interest is itself an interaction term).

Main Analyses.

Evaluative Ratings. Figure 1a shows the means and distributions of the rating differences. In both conditions the difference between CS ratings was positive, meaning that the CS_{pos} was rated more positively than the CS_{neg}. In the high distinctiveness group, the intercept (3.76) was significantly different from zero, $F(1, 76) = 127.32$, $p < .001$. In the low distinctiveness group, the intercept was slightly smaller (2.38) but also significant, $F(1, 72) = 46.69$, $p < .001$. In other words, evaluative ratings reflected clear OEC effects in both groups. The OEC effect was significantly moderated by distinctiveness, $F(1, 148) = 8.16$, $p = .005$, $\eta^2_p = 0.05$, [0.01, 0.12], $BF_{10} = 83.53$, such that the difference between CS ratings was smaller in the low distinctiveness condition than in the high distinctiveness condition. Finally, the results were unchanged when participants with incorrect contingency memory were excluded ($n = 15$), when participants with incorrect manipulation memory were excluded ($n = 12$), or when participants who reported compliance or reactance were excluded ($n = 39$).

Figure 1

Means and Distributions of Differences Between CS Ratings in Experiments 1a and 1b



Note. Dots represent the data of individual participants (with random noise added along the horizontal axis to improve clarity). Black circles and error bars indicate means and 95% confidence intervals. Grey areas represent the distribution of the data. Zero (i.e., no difference between CS_{pos} and CS_{neg} ratings) is indicated by the dashed line. **(a)** Data obtained in Experiment 1a. **(b)** Data obtained in Experiment 1b.

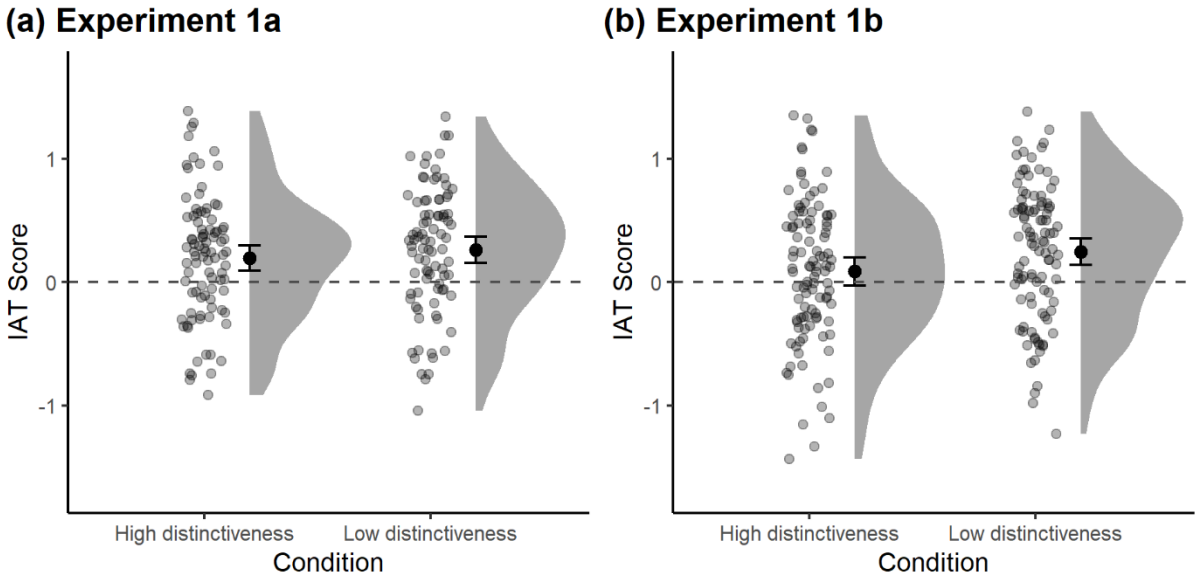
Because visual inspection of the data suggested that the impact of distinctiveness was different for the CS_{pos} and the CS_{neg} ratings, we conducted exploratory analyses on the individual ratings (rather than the difference between them). After changing the sign of the CS_{neg} ratings so that a higher score always reflected a larger learning effect, we subjected the ratings to a mixed ANOVA which included CS as a within-subjects factor. The CS × distinctiveness interaction was significant, $F(1, 148) = 5.29$, $p = .023$, indicating that the effect of distinctiveness indeed differed between CSs. In line with this, *t*-tests indicated that while CS_{neg} ratings were clearly moderated by distinctiveness, $t(177.32) = -4.19$, $p < .001$, CS_{pos} ratings were not, $t(177.91) = 1.11$, $p = .27$. Finally, it is worth noting that ratings of

both CSs differed significantly from zero in both groups, with the exception of the CS_{neg} ratings in the low distinctiveness group (M = -0.28, SD = 2.04), $t(87)$ = -1.27, p = .208.

IAT Scores. The (bootstrapped) split-half reliability of the IAT was .83. Figure 2a shows the means and distributions of IAT scores. The intercept in the high distinctiveness group (0.17) was significantly different from zero, $F(1, 76) = 12.14$, $p < .001$, and so was the intercept in the low distinctiveness group (0.24), $F(1, 72) = 21.47$, $p < .001$. In other words, both groups showed clear OEC effects in terms of IAT performance. Unlike the ratings, however, IAT scores were not moderated by distinctiveness, $F(1, 148) = 1.18$, $p = .279$, η^2_p = 0.01, [0.00, 0.05], BF_{10} = 0.48. Note that contrary to predictions, mean IAT scores were numerically *lower* in the high distinctiveness condition. Again, these results were unchanged when participants were excluded who did not remember the contingencies, who did not remember the distinctiveness information, or who reported demand compliance or reactance.

Figure 2

Means and Distributions of IAT Scores in Experiments 1a and 1b



Note. Dots represent the data of individual participants (with random noise added along the horizontal axis to improve clarity). Black points and error bars indicate means and 95%

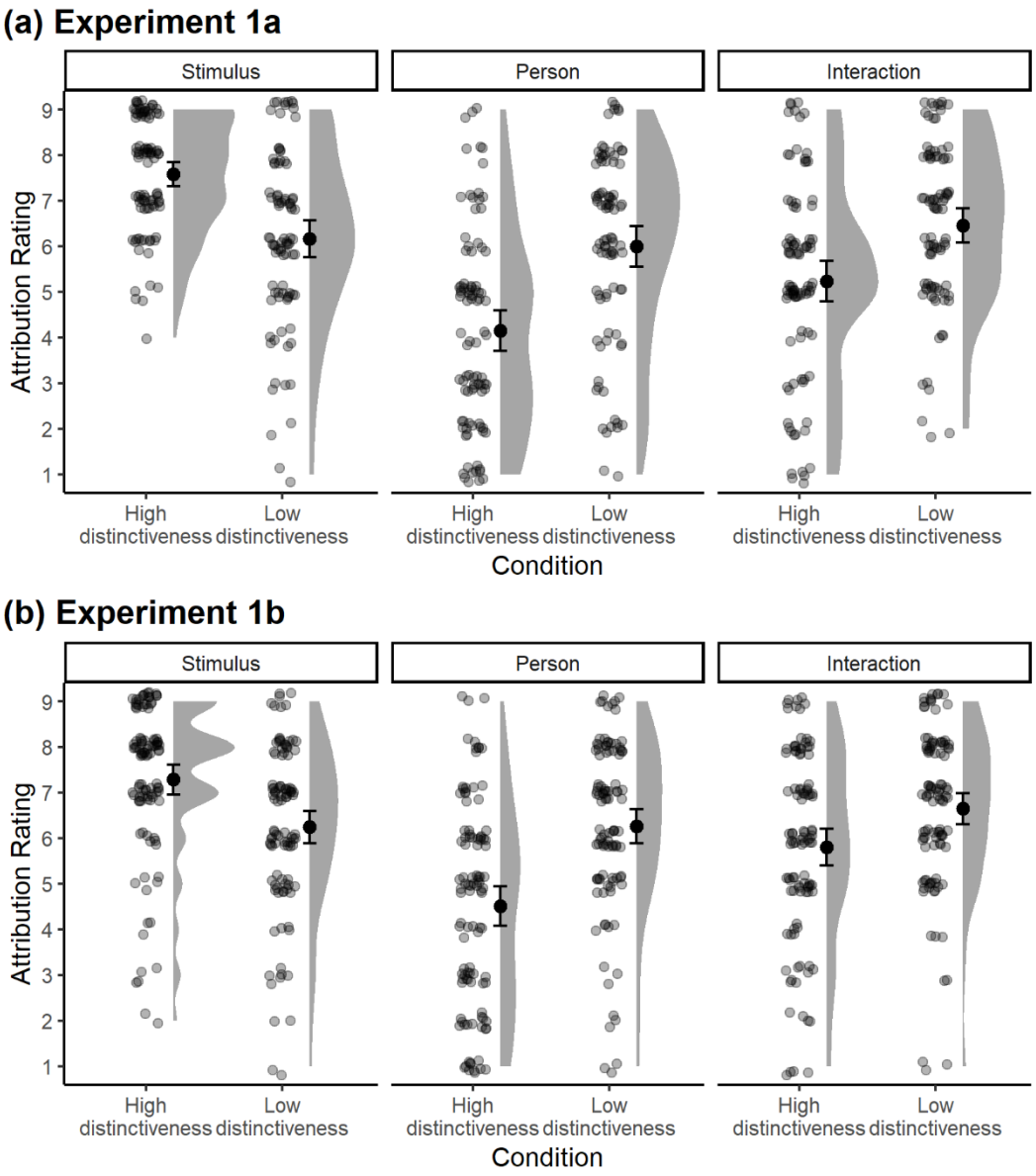
confidence intervals. Grey areas represent the distribution of the data. Zero (i.e., no difference between speed on learning-compatible and on learning-incompatible blocks) is indicated by the dashed line. **(a)** Data obtained in Experiment 1a. **(b)** Data obtained in Experiment 1b.

Exploratory Analyses. The perceived distinctiveness ratings indicated that our manipulation had been successful. The positive reaction was rated as significantly more distinctive in the high distinctiveness group (M = 7.10, SD = 1.71) than in the low distinctiveness group (M = 5.50, SD = 2.29), $t(161.07) = 5.28$, $p < .001$, $d = 0.79$, [0.48, 1.11], $BF_{10} > 10000$. The negative reaction was also rated as more distinctive in the high (M = 7.22, SD = 2.02) than in the low (M = 6.03, SD = 2.74) distinctiveness group, $t(159.74) = 3.28$, $p = .001$, $d = 0.49$, [0.19, 0.79], $BF_{10} = 23.15$. In addition, the distinctiveness rating of each reaction was correlated significantly with the evaluative rating of the corresponding CS (positive reaction – CS_{pos} rating: $r = .25$, $t(177) = 3.38$, $p < .001$; negative reaction – CS_{neg} rating: $r = -.33$, $t(177) = -4.68$, $p < .001$). Mean distinctiveness ratings were not correlated with IAT scores, $r = .09$, $t(176) = 1.24$, $p = .217$.

Figures 3a and 4a show the pattern of attributions for the positive reaction and for the negative reaction, respectively. In line with expectations, stimulus attributions were rated significantly higher in the high distinctiveness group than in the low distinctiveness group, both for the positive reaction (M = 7.59, SD = 1.26, versus M = 6.17, SD = 1.91), $t(149.79) = 5.85$, $p < .001$, $d = 0.88$, [0.56, 1.20], $BF_{10} > 10000$, and the negative reaction (M = 8.01, SD = 1.46, versus M = 5.20, SD = 2.47), $t(140.03) = 9.21$, $p < .001$, $d = 1.39$, [1.03, 1.74], $BF_{10} > 10000$. Whereas stimulus attributions for the negative reaction were correlated with CS_{neg} ratings, $r = -.40$, $t(178) = -5.91$, $p < .001$, stimulus attributions for the positive reaction were not significantly correlated with CS_{pos} ratings, $r = .14$, $t(178) = 1.86$, $p = .065$. Mean stimulus attributions were uncorrelated with IAT scores, $r = -.01$, $t(178) = -0.19$, $p = .848$.

Figure 3

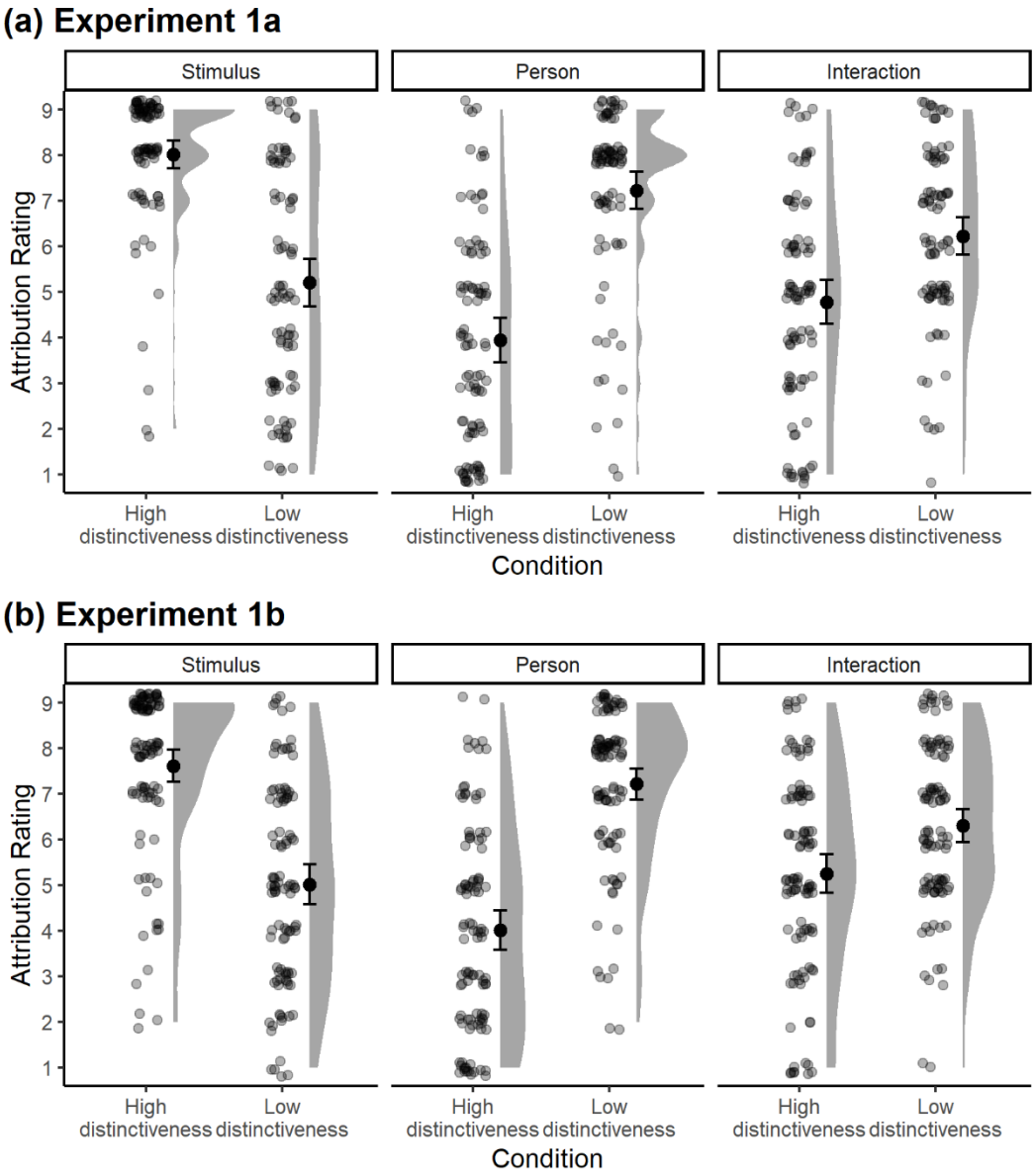
Means and Distributions of Attributions for the Positive Reaction in Experiments 1a and 1b



Note. The left panels show the results for stimulus attributions, the middle panels show the results for person attributions, and the right panels show the results for stimulus-person attributions. Dots represent the data of individual participants (with random noise added along the horizontal and vertical axes to improve clarity). Black circles and error bars indicate means and 95% confidence intervals. Grey areas represent the distribution of the data. **(a)** Data obtained in Experiment 1a. **(b)** Data obtained in Experiment 1b.

Figure 4

Means and Distributions of Attributions for the Negative Reaction in Experiments 1a and 1b



Note. The left panels show the results for stimulus attributions, the middle panels show the results for person attributions, and the right panels show the results for stimulus-person attributions. Dots represent the data of individual participants (with random noise added along the horizontal and vertical axes to improve clarity). Black circles and error bars indicate means and 95% confidence intervals. Grey areas represent the distribution of the data. **(a)** Data obtained in Experiment 1a. **(b)** Data obtained in Experiment 1b.

Finally, we should mention that after being informed what our main hypothesis was, many participants (70% in the high distinctiveness group and 73% in the low distinctiveness group) reported that they had been aware of this hypothesis.

Discussion

In Experiment 1a we tested whether OEC effects would be moderated by the distinctiveness of the modelled reactions. Our results provided partial support for this prediction: relative to the low distinctiveness group, the high distinctiveness group perceived the reactions to be more distinctive, attributed them more to the CSs, and showed larger OEC effects in terms of evaluative ratings. Interestingly, this seemed mostly due to participants' ratings of the CS_{neg} rather than their ratings of the CS_{pos}, a finding that we will return to in the general discussion. The IAT results were not entirely in line with predictions: although IAT scores revealed clear OEC effects, they did not differ significantly between groups and were numerically slightly smaller in the high distinctiveness condition.

In sum, based on Experiment 1a we have initial evidence that OEC effects indexed by self-reports are sensitive to distinctiveness information, but we cannot say the same for OEC effects indexed by automatic evaluations. This limits the conclusions we can draw based on this experiment, especially if we take into account the high degree of hypothesis awareness (although note that we may have overestimated this since we simply asked participants if they were aware of the hypothesis after explaining it). If we assume that IAT performance is more difficult to control than self-reports, the high level of hypothesis awareness leaves open the possibility that the pattern on the self-reports was simply the result of demand compliance.

However, the lack of an effect of distinctiveness on the IAT could also have had other reasons. For example, the evaluative ratings and the IAT in Experiment 1a differed not only in terms of their automaticity conditions (such as time pressure), but also in terms of what was measured. Specifically, the self-report questions required participants to rate how much

they thought they would like each cookie (i.e., assessed anticipated liking), whereas the labels used in the IAT ("I like" and "I dislike") referred to "actual" liking. Before drawing strong conclusions based on the findings of Experiment 1a, we wanted to see if the same results would be obtained if we used an IAT that referred to *anticipated* liking, especially since a pilot study suggested that the characteristics of an "anticipated liking" IAT differed slightly from those of the "actual liking" IAT used in Experiment 1a (materials and data for the pilot are available at https://osf.io/k48rg/). We therefore repeated Experiment 1a but with an anticipated liking IAT (Experiment 1b). This allowed us to (a) see if we could replicate the initial self-report findings of Experiment 1a and (b) test the impact of distinctiveness on automatic evaluations measured with a different IAT.

Experiment 1b

Experiment 1b was largely identical to Experiment 1a, with the important exception that we used an anticipated liking IAT rather than an actual liking IAT. Although we altered some of the exploratory questions to increase their clarity and left out some open-ended questions, no other fundamental changes were made to the procedure. We again expected to find clear OEC effects and predicted that these effects would be moderated by distinctiveness, both in terms of evaluative ratings and in terms of IAT performance.

Method

Participants and Design

In order to have 90% power to detect a main effect of distinctiveness based on the effect size found in Experiment 1a ($\eta^2_p = 0.052$), a sample of 194 participants was required. Taking into account 10% exclusions based on IAT performance, we planned to recruit participants until we had complete and useable data for $n = 216$ (during data collection, we replaced participants who provided incomplete data, who restarted and encountered different versions, who reported technical issues, or who indicated that we should exclude their data).

Participants were recruited via Prolific based on the same criteria as for Experiment 1a. After excluding incomplete entries (n = 10), entries from participants who restarted (n = 6), entries from participants who reported technical issues (n = 3), and entries from participants who said to exclude their data (n = 3), the sample consisted of 216 participants with complete and useable data (100 men, 112 women, 4 non-binary people; M_{age} = 30.23, SD_{age} = 8.40).

The design was identical to that of Experiment 1a. Because fully random assignment had resulted in a quite unbalanced sample for Experiment 1a, we adopted a different sampling strategy that ensured an adequate level of counterbalancing. Specifically, we first recruited 170 participants using random assignment, then recruited additional participants per cell until there were at least five participants in each cell, and finally used random assignment to cells with only five or six participants until we reached the planned sample size. The resulting sample was still somewhat unbalanced but much less so than in Experiment 1a.

Materials

The videos were the same as those used in Experiment 1a. However, as we now used an anticipated liking IAT (see below) rather than an actual liking IAT, the IAT stimuli were changed. The target stimuli now consisted of four versions of each CS name (in lower- or uppercase and regular or italic font). As in Experiment 1a, the two names (“Empeya” and “Plogo”) served as labels for categorizing these stimuli. The valenced stimuli were four positive (Promising, Tempting, Enticing, Probably good) and four negative (Off-putting, Unappealing, Repulsive, Probably bad) adjectives referring to expectations about food. The labels “I think I would like” and “I think I would dislike” were used to sort these stimuli.

Procedure

The procedure was largely identical to that of Experiment 1a. The main difference was the IAT, which was designed to measure anticipated liking of the CSs rather than actual liking. Instead of asking participants to classify the valenced adjectives based on whether

they referred to something they liked or something they disliked (as in Experiment 1a), we instructed them to categorize the adjectives based on whether they represented synonyms of "I think I would like" or "I think I would dislike". Since these instructions clearly indicated a normatively correct response (unlike in Experiment 1a), error feedback was provided on these trials as well. In addition, because there were only four stimuli in each category rather than five, Blocks 1, 2, 3, 5, and 6 consisted of 16 trials (rather than 20) and Blocks 4 and 7 consisted of 32 trials (rather than 40), leading to 144 trials in total.

We also made some changes to the exploratory questions. The open-ended hypothesis awareness question was simplified and asked immediately after the main experiment. We also asked participants to report their awareness separately for our first hypothesis (i.e., that we would find OEC effects) and our second hypothesis (i.e., that OEC effects would be influenced by the distinctiveness information). The demand compliance and reactance questions were simplified as well (by asking participants to indicate "Yes", "No", or "I don't know" when asked whether these had influenced their responses, rather than asking them to provide ratings). Finally, we also asked whether they believed their data should be excluded (e.g., because of responding randomly or being distracted), emphasizing that they would be compensated for their time regardless of their answer. Responses to this question were used as an exclusion criterion (see "Participants and Design").

Results

Data Preparation

We excluded participants who made more than 30% errors across the IAT (n = 2), who made more than 40% errors on any of the combined blocks (n = 7), or who completed more than 10% of trials faster than 300 ms (n = 1). The final sample consisted of 206 participants (96 men, 107 women, 3 non-binary people; M_{age} = 30.24, SD_{age} = 8.38). Rating differences and IAT scores were calculated in the same way as for Experiment 1a. We report

the results for the D4-score rather than for the D1-score as originally pre-registered, but we note if any of the main results diverged when the latter score was used instead.

Data Analysis

The analytic strategy was identical to that of Experiment 1a, with the exception that demand compliance and reactance exclusions were based on participants responding "Yes" or "I don't know" rather than them selecting the midpoint or higher on a rating scale.

Main Analyses.

Evaluative Ratings. Figure 1b shows the means and distributions of the rating differences. In both conditions the CS_{pos} was rated higher than the CS_{neg}. In the high distinctiveness group, the intercept (3.90) was significantly different from zero, $F(1, 88) = 124.16$, $p < .001$; in the low distinctiveness group, the intercept (2.14) was slightly smaller but still highly significant, $F(1, 86) = 69.06$, $p < .001$. In other words, both groups showed clear OEC effects. There was also an impact of distinctiveness, $F(1, 174) = 16.28$, $p < .001$, $\eta^2_p = 0.09$, [0.03, 0.16], $BF_{10} = 105.37$, such that the OEC effect was smaller in the low distinctiveness condition than in the high distinctiveness condition. We should note that this main effect was qualified by a significant interaction with task order, $F(1, 174) = 5.67$, $p = .018$, such that the effect of distinctiveness was only significant if the ratings were completed before the IAT, $F(1, 82) = 26.51$, $p < .001$, not if they were completed after the IAT, $F(1, 92) = 1.17$, $p = .283$. Finally, these results were unchanged when participants with incorrect contingency memory were excluded ($n = 14$) or when participants with incorrect manipulation memory were excluded ($n = 8$). When participants who reported possible demand compliance or reactance were excluded ($n = 73$), the interaction between distinctiveness and task order was non-significant ($p = .116$).

We again conducted exploratory analyses on the individual ratings. There was a significant CS × distinctiveness interaction, $F(1, 174) = 30.45$, $p < .001$, indicating that the

effect of distinctiveness differed between CSs. Similar to Experiment 1a, CS_{neg} ratings were clearly moderated by distinctiveness, $t(202.42) = -5.32$, $p < .001$, while CS_{pos} ratings were not, $t(188.70) = 0.05$, $p = .960$. Also replicating Experiment 1a, CS_{neg} ratings in the low distinctiveness group ($M = -0.11$, $SD = 1.95$) did not differ from zero, $t(101) = -0.57$, $p = .567$.

IAT Scores. The (bootstrapped) split-half reliability of the IAT was .82. Figure 2b shows the means and distributions of IAT scores. In the high distinctiveness group, the intercept (0.11) was not significantly different from zero, $F(1, 88) = 3.93$, $p = .051$ (note that it was significant for the D1-score, $p = .020$, and when participants who reported demand compliance or reactance were excluded, $p = .050$). The intercept in the low distinctiveness group (0.25) on the other hand was highly significant, $F(1, 86) = 26.01$, $p < .001$. There was a significant main effect of distinctiveness, $F(1, 174) = 3.96$, $p = .048$, $\eta^2_p = 0.02$, [0.0001, 0.0701], $BF_{10} = 1.23$, in the opposite direction than predicted: IAT scores were smaller in the high distinctiveness condition than in the low distinctiveness condition. However, this effect was not robust: it was non-significant for the D1-score ($p = .081$) as well as when participants with incorrect memory were excluded ($p = .348$), when participants with incorrect manipulation memory were excluded ($p = .080$), or when participants who reported possible demand compliance or reactance were excluded ($p = .262$).

Exploratory Analyses. The positive reaction was rated as significantly more distinctive in the high distinctiveness group ($M = 6.94$, $SD = 2.04$) than in the low distinctiveness group ($M = 5.60$, $SD = 2.19$), $t(202.39) = 4.55$, $p < .001$, $d = 0.64$, [0.35, 0.92], $BF_{10} = 1825.17$. The negative reaction was also rated as more distinctive in the high ($M = 7.12$, $SD = 2.15$) than in the low ($M = 5.83$, $SD = 2.64$) distinctiveness group, $t(194.20) = 3.82$, $p < .001$, $d = 0.53$, [0.25, 0.81], $BF_{10} = 122.06$. In addition, the distinctiveness of each reaction was correlated with the rating of the corresponding CS (positive reaction – CS_{pos}

rating: r = .21, t(204) = 3.09, p = .002; negative reaction – CS_{neg} rating: r = -.18, t(204) = -2.64, p = .009). However, mean distinctiveness ratings were uncorrelated with IAT scores, r = .02, t(204) = 0.24, p = .814.

Figures 3b and 4b show the pattern of attributions in the two groups. As expected, stimulus attributions were rated significantly higher in the high distinctiveness group than in the low distinctiveness group, both for the positive reaction (M = 7.29, SD = 1.68, versus M = 6.25, SD = 1.79), t(202.47) = 4.31, p < .001, d = 0.60, [0.31, 0.89], BF_{10} = 714.14, and for the negative reaction (M = 7.62, SD = 1.82, versus M = 5.02, SD = 2.20), t(195.88) = 9.22, p < .001, d = 1.29, [0.96, 1.61], BF_{10} > 10000. Stimulus attributions for each reaction were also correlated with the evaluative rating of the corresponding CS (positive reaction – CS_{pos} rating: r = .27, t(204) = 3.93, p < .001; negative reaction – CS_{neg} rating: r = -.43, t(204) = -6.74, p < .001). Mean stimulus attributions were uncorrelated with IAT scores, r = -.05, t(204) = -0.66, p = .508.

Finally, many participants (64% in the high distinctiveness group and 65% in the low distinctiveness group) reported having been aware we expected to find OEC effects. A smaller but still substantial percentage in each group (respectively 50% and 40%) reported having been aware we expected distinctiveness to have an impact.

Discussion

In Experiment 1b we replicated the self-report results of Experiment 1a: relative to participants in the low distinctiveness condition, participants in the high distinctiveness condition found the reactions more distinctive, attributed them more strongly to the stimuli, and showed larger OEC effects (which was again primarily due to the CS_{neg} ratings). However, IAT scores did not exhibit the same pattern: the high distinctiveness group showed very weak OEC effects (if any), while the low distinctiveness group showed clear OEC effects. In sum, although the results of Experiments 1a-1b are in line with the idea that

attributions play a role in OEC effects indexed via self-reports, they do not provide any evidence that this is also the case for OEC effects as indexed via more automatic evaluations.

So far we tried to influence attributions solely by manipulating the distinctiveness of the observed reactions. However, distinctiveness is only one variable that is considered to be important by attribution theories. In order to try and generalize our conclusions, we set out to use a different (preferably stronger) manipulation of attributions in Experiment 2.

Experiment 2

For Experiment 2 we developed a different manipulation of stimulus attributions. Based on the attribution literature, we considered the possibility of manipulating all three variables (consensus, consistency, and distinctiveness) at once, because configurations can have cumulative effects on attributions (McArthur, 1972) and direct participants to a "logical" attribution (Hewstone & Jaspars, 1987). However, a pilot study indicated that while low consensus, consistency, and distinctiveness (LLL) information led to slightly weaker stimulus attributions than high consensus, consistency, and distinctiveness (HHH) information, the difference was actually less pronounced than in Experiments 1a-1b (pilot materials and data are available at https://osf.io/3wb8g/). As including the distinctiveness information somewhat complicated matters (because we had to show two models in order to manipulate this variable for both reactions), we conducted a second pilot study to see if providing only consensus and consistency information for reactions shown by a single model would have a clearer impact on stimulus attributions (pilot materials and data are available at https://osf.io/cfxp7/). Because low consensus and low consistency (LL) information indeed reduced stimulus attributions relative to high consensus and high consistency (HH) information, we decided to compare participants who received HH information to participants who received LL information in Experiment 2. We again tested the impact of this manipulation on OEC effects assessed via evaluative ratings and an anticipated liking IAT.

The pilot also indicated that there was a large amount of variance in the LL group, which suggested that the LL information mainly created uncertainty because it did not direct participants toward a clear alternative cause to which the reactions could be attributed. Therefore, we wanted to explore whether it would be possible to further reduce stimulus attributions in an LL condition by providing a possible explanation for why the observed reactions were low in terms of consensus and consistency. A third condition was therefore included which provided LL information as well as an explanation (LL-E condition).

Most importantly, we predicted that participants would show OEC effects and that these effects would be smaller in the LL and LL-E conditions than in the HH condition. We also formulated two more exploratory predictions: OEC effects were predicted to be smaller in the LL-E group than in the LL group, and the amount of variance in OEC effects was also predicted to be smaller in the former group relative to the latter. Although these latter predictions focused mostly on self-reports (because these were clearly correlated with stimulus attributions in Experiments 1a-1b), we also tested them for the IAT scores.

Method

Participants and Design

In order to have 90% power to detect medium-sized differences between conditions with two-sample *t*-tests, a minimum sample size of 258 participants (86 per condition) was required. Because we estimated around 5% exclusions based on IAT performance and we wanted to fully counterbalance the design (see below), we planned to recruit participants until we had complete and useable data for $n = 288$ (after replacing participants who provided incomplete data, who reported technical issues, who stated their data should not be used, or who failed more than one attention check). Participants were recruited via Prolific using the same criteria as before. After excluding incomplete entries ($n = 10$), entries from a participant who restarted and completed the ratings twice ($n = 2$), entries from participants who indicated

their data should be excluded (n = 1), and entries from participants who failed more than one attention check (n = 3), the sample consisted of 289 participants with complete and useable data (112 men, 172 women, 5 non-binary people; M_{age} = 31.11, SD_{age} = 7.94; one more than planned because one participant timed out but provided useable data).

We used a between-subjects design with three levels for condition: HH information, LL information, and LL-E information. In addition, we counterbalanced stimulus assignment, task order, and IAT block order across participants. Unlike in Experiments 1a-1b, we recruited participants separately for each cell, resulting in almost perfect counterbalancing.

Materials

Because we now needed videos of only one model and no longer needed to ensure that valence and believability were matched between two models, we used the pre-ratings collected in preparation of Experiment 1a to select the most suitable positive reaction and the most suitable negative reaction of one model (Model A from Experiments 1a-1b). After selecting two videos that were clearly valenced and sufficiently believable, we again edited these videos to vary the name on the label, resulting in four videos in total.

The IAT stimuli were identical to those used in Experiment 1b.

Procedure

Participants were again told that we were working with a company that was testing new cookie recipes and that we had asked some people to try samples of two of those cookies (Empeya and Plogo). We also mentioned that in order to get as much information as possible, we had asked multiple people to taste the cookies and had also asked each person to taste each cookie several times. Participants then watched the OEC videos, read the consensus and consistency information, completed the evaluative measures, and answered several exploratory questions. Note that unlike in Experiments 1a-1b, we included the manipulation of attributions *after* the OEC phase. This was done because otherwise any obtained

differences could have been due to participants in the LL group simply paying less attention to the videos after already inferring from the LL information that the reactions would not be informative (which was not an issue in Experiments 1a-1b because participants could not know whether distinctiveness was high or low until they had actually watched the videos). We also included several attention checks throughout the experiment so that we could identify and exclude participants who did not read the questions.

OEC Phase. Participants were informed that during the taste tests we had asked people to show us how they felt and recorded videos, and that they would see two of those videos. Each participant was then shown two of the four videos: one in which the model reacted positively to the CS_{pos} and one in which the same model reacted negatively to the CS_{neg}. Which cookie name (Empeya or Plogo) was shown in the former video and which in the latter depended on stimulus assignment. Similar to Experiments 1a-1b, both videos were shown three times in a random presentation order with an inter-trial-interval of two seconds.

Consensus-Consistency Manipulation. Participants were given additional information supposedly obtained from the other taste tests that we had conducted. Those in the HH condition were told that "*Almost everybody else's reactions* to the Empeya and Plogo cookies were *similar* to the reactions you saw in the videos" (high consensus) and that "On all of the *other occasions* that the person from the videos tasted the Empeya and Plogo cookies, she *always* reacted to them like she did in the videos you saw" (high consistency). In contrast, participants in the LL condition were told that "*Almost everybody else's reactions* to the Empeya and Plogo cookies were *different* from the reactions you saw in the videos" (low consensus) and that "On all of the *other occasions* that the person from the videos tasted the Empeya and Plogo cookies, she *never* reacted to them like she did in the videos you saw" (low consistency). Finally, participants in the LL-E condition also received the low consensus and low consistency information, followed by a potential explanation: "You might be

wondering why the reactions you saw in the videos differed from how most other people reacted to the cookies and from how this person reacted to the cookies on other occasions. It is important to mention that the person in the videos reported that she had eaten an *extremely spicy dish* shortly before these videos were recorded. She told us that this seemed to *influence her sense of taste* during the taste test."

The information was followed by a manipulation check that required participants to report what had happened when other people tasted the cookies (i.e., consensus) and what had happened when the model tasted the cookies on other occasions (i.e., consistency). Participants could only proceed after answering both questions correctly.

Evaluative Measures. The evaluative ratings and IAT were identical to those used in Experiment 1b, except that a first attention check – which simply asked to indicate a specific value on the rating scale – was intermixed with the ratings.

Exploratory Questions. Participants answered a number of exploratory questions (including two further attention checks, one asking to indicate a specific value and one asking to fill in a specific word). They were first asked to type in what they believed our hypotheses to be (hypothesis awareness). Next, we assessed attributions for the stimulus, the person, and the circumstances (i.e., "something about the specific circumstances in which the video was filmed"). They were also asked to what extent they thought that the observed reactions could inform them about the taste of the cookies (informativeness). We again assessed contingency and manipulation memory; for the LL-E group, this also included an open-ended question about the potential explanation that we had mentioned. Finally, we explained the main hypotheses and asked participants whether they had been aware of each during the study, whether their ratings had been based on demand compliance or reactance, whether their data should be excluded, and whether they had encountered any issues.

Results

Data Preparation

We excluded participants who made more than 30% errors across the IAT (n = 1), who made more than 40% errors on any of the combined blocks (n = 7), or who completed more than 10% of trials faster than 300 ms (n = 2). The final sample consisted of 279 participants (108 men, 166 women, 5 non-binary people; M_{age} = 31.04, SD_{age} = 8.01). Evaluative ratings and IAT scores were treated as in Experiments 1a-1b (in line with our pre-registration for Experiment 2, we conducted the IAT analyses only with the D4-score).

Data Analysis

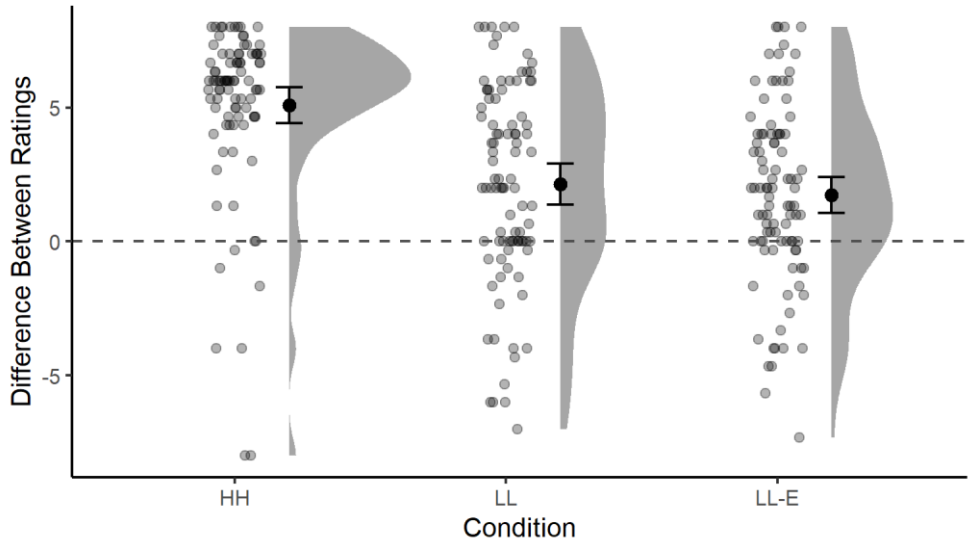
Analytic Strategy. Unlike for Experiments 1a-1b, the design was almost perfectly balanced and it was not necessary to include the counterbalanced factors in the analyses (although we did check afterwards if including them affected the interpretation of the main results, which was not the case). To test for OEC effects, one-sample *t*-tests were conducted for each group to see if the mean rating differences and IAT scores were significantly different from zero. To test whether the groups differed from each other in the expected manner, we conducted pairwise *t*-tests (using Holm-Bonferroni correction for multiple comparisons) to compare the three conditions to one another. We also conducted an *F*-test of equality of variances to test whether the variance was smaller in the LL-E condition than in the LL condition. In line with our pre-registration, we explored whether stimulus attributions and informativeness ratings were influenced by the manipulation in the expected way. Finally, we checked whether the main results were changed if participants with incorrect contingency memory were excluded, if participants with incorrect manipulation memory were excluded, or if participants who reported potential demand compliance or reactance were excluded. All other analyses were not pre-registered and were purely exploratory.

Main Analyses.

Evaluative Ratings. Figure 5 shows the means and distributions of the rating differences. The difference between CS ratings was significantly larger than zero in the HH condition (M = 5.08, SD = 3.18), $t(90) = 15.26$, $p < .001$, d = 1.60, [1.29, 1.91], $BF_{10} >$ 10000, in the LL condition (M = 2.14, SD = 3.70), $t(92) = 5.58$, $p < .001$, d = 0.58, [0.36, 0.80], $BF_{10} >$ 10000, and in the LL-E condition (M = 1.73, SD = 3.35), $t(94) = 5.05$, $p < .001$, d = 0.52, [0.30, 0.73], BF_{10} = 6759.77. In other words, all three groups showed clear OEC effects. Pairwise comparisons indicated that OEC effects were significantly larger in the HH group than in the LL group, (corrected) $p < .001$, d = 0.85, [0.54, 1.17], $BF_{10} >$ 10000, as well as larger than in the LL-E group, $p < .001$, d = 1.03, [0.70, 1.35], $BF_{10} >$ 10000. However, the size of OEC effects did not differ between the LL and the LL-E groups, $p = .420$, d = 0.12, [-0.17, 0.40], BF_{10} = 0.21. Finally, while the variance in the LL-E group (11.19) was slightly smaller than the variance in the LL group (13.70), the difference was not significant, $F(92, 94) = 1.22$, $p = .332$. The results were unchanged when participants with incorrect contingency memory were excluded (n = 35), when participants with incorrect manipulation memory – including those in the LL-E group who could not report the provided explanation – were excluded (n = 14), or when participants who reported potential demand compliance or reactance were excluded (n = 44).

Figure 5

Means and Distributions of Differences Between CS Ratings in Experiment 2



Note. Dots represent the data of individual participants (with random noise added along the horizontal axis to improve clarity). Black circles and error bars indicate means and 95% confidence intervals. Grey areas represent the distribution of the data. Zero (i.e., no difference between CS_{pos} and CS_{neg} ratings) is indicated by the dashed line. HH: high consensus and high consistency. LL: low consensus and low consistency. LL-E: low consensus, low consistency, and availability of an explanation.

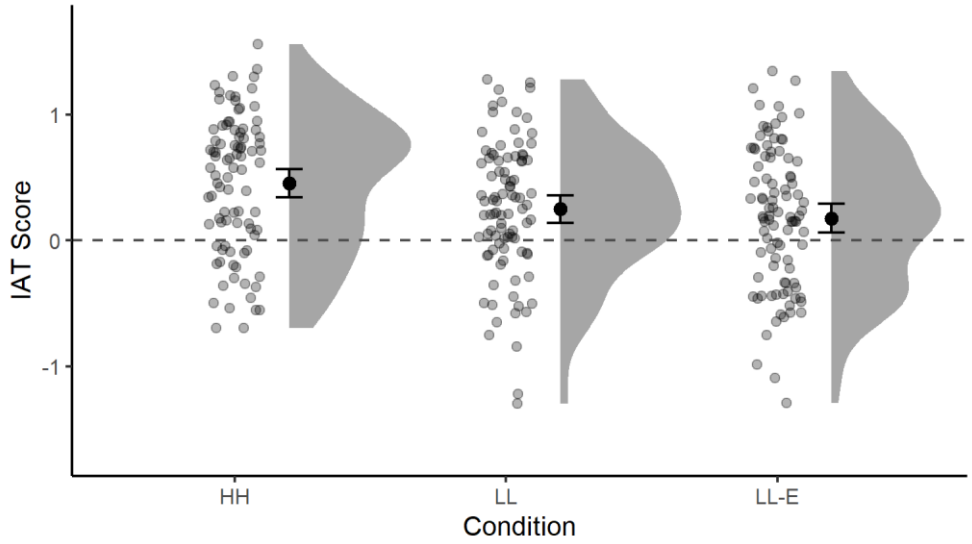
We again conducted exploratory analyses on the individual CS ratings. Both the ratings of the CS_{pos} and the ratings of the CS_{neg} differed significantly from zero in all three conditions, with the exception of the CS_{neg} ratings in the LL-E group (M = -0.21, SD = 2.07), $t(94) = -1.01$, $p = .316$. Otherwise, the results for the individual ratings were similar to the results reported for the difference scores.

IAT Scores. The (bootstrapped) split-half reliability of the IAT was .81. Figure 6 shows the means and distributions of IAT scores in the three conditions. IAT scores were significantly larger than zero in the HH condition (M = 0.45, SD = 0.55), $t(90) = 7.94$, $p <$

.001, d = 0.83, [0.59, 1.07], BF_{10} > 10000, the LL condition (M = 0.25, SD = 0.53), t(92) = 4.52, p < .001, d = 0.47, [0.25, 0.68], BF_{10} = 917.60, and the LL-E condition (M = 0.18, SD = 0.56), t(94) = 3.08, p = .003, d = 0.32, [0.11, 0.52], BF_{10} = 9.15. Similar to the ratings, OEC effects were larger in the HH condition than in the LL condition, p = .024, d = 0.38, [0.08, 0.67], BF_{10} = 3.35, as well as larger than in the LL-E condition, p = .002, d = 0.50, [0.20, 0.80], BF_{10} = 33.10, while the LL and LL-E conditions did not differ, p = .360, d = 0.13, [-0.15, 0.42], BF_{10} = 0.23. The variance was also similar in the LL (0.28) and the LL-E (0.31) conditions, F(92, 94) = 0.91, p = .634. The results were unchanged when participants with incorrect contingency memory were included, although the evidence for the comparisons between the HH and the two other groups was more convincing (HH vs. LL: p = .002, d = 0.52, BF_{10} = 26.24; HH vs. LL-E: p < .001, d = 0.67, BF_{10} = 575.91). The results were also unchanged when excluding participants who did not remember the manipulation or who reported potential compliance or reactance.

Figure 6

Means and Distributions of IAT Scores in Experiment 2



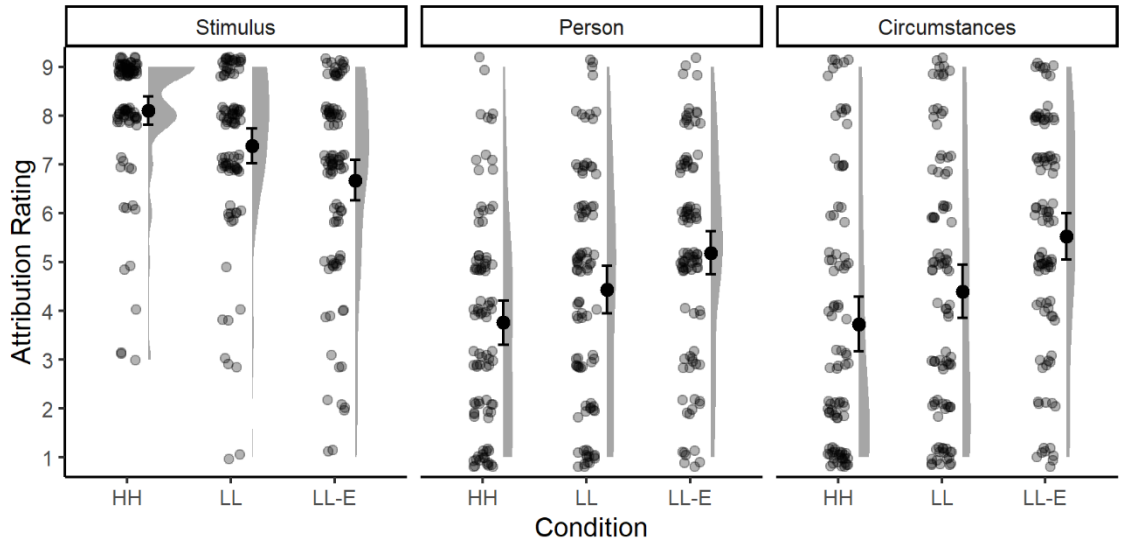
Note. Dots represent the data of individual participants (with random noise added along the horizontal axis to improve readability). Points and error bars indicate means and 95% confidence intervals. Grey areas represent the distribution of the data. Zero (i.e., no difference between speed on learning-compatible and on learning-incompatible blocks) is indicated by the dashed line. HH: high consensus and high consistency. LL: low consensus and low consistency. LL-E: low consensus, low consistency, and availability of an explanation.

Exploratory Analyses. Figures 7 and 8 show the pattern of attributions (for the positive and the negative reaction, respectively) in the three groups, which seemed largely in line with expectations. Stimulus attributions for the positive reaction were strongest in the HH group (M = 8.10, SD = 1.40), slightly weaker in the LL group (M = 7.38, SD = 1.74), and weakest in the LL-E group (M = 6.67, SD = 2.04). Pairwise comparisons confirmed that all three groups differed significantly from one another (HH vs. LL: p = .011, d = 0.46, [0.16, 0.75], BF_{10} = 13.14; HH vs. LL-E: p < .001, d = 0.81, [0.50, 1.12], BF_{10} > 10000; LL vs. LL-E: p = .011, d = 0.37, [0.08, 0.66], BF_{10} = 3.11). The pattern for the negative reaction was similar, with the strongest stimulus attributions in the HH group (M = 7.54, SD = 2.16), weaker stimulus attributions in the LL group (M = 6.81, SD = 2.21), and the weakest stimulus attributions in the LL-E group (M = 5.85, SD = 2.54), and significant differences between all three groups (HH vs. LL: p = .033, d = 0.33, [0.04, 0.63], BF_{10} = 1.73; HH vs. LL-E: p < .001, d = 0.71, [0.41, 1.02], BF_{10} = 5977.85; LL vs. LL-E: p = .010, d = 0.40, [0.11, 0.69], BF_{10} = 5.07). We also checked whether the variance of stimulus attributions was larger in the LL than in the LL-E condition, which was not the case for either the positive reaction (variances of respectively 3.02 and 4.16), F(92, 94) = 0.73, p = .125, or the negative reaction (variances of respectively 4.88 and 6.47), F(92, 94) = 0.75, p = .175. Finally, stimulus attributions for the positive reaction were correlated with the CS_{pos} ratings, r = .37, t(277) =

6.65, $p < .001$, stimulus attributions for the negative reaction were correlated with the CS_{neg} ratings, $r = -.38$, $t(277) = -6.82$, $p < .001$, but mean stimulus attributions were uncorrelated with IAT scores, $r = .07$, $t(277) = 1.23$, $p = .221$.

Figure 7

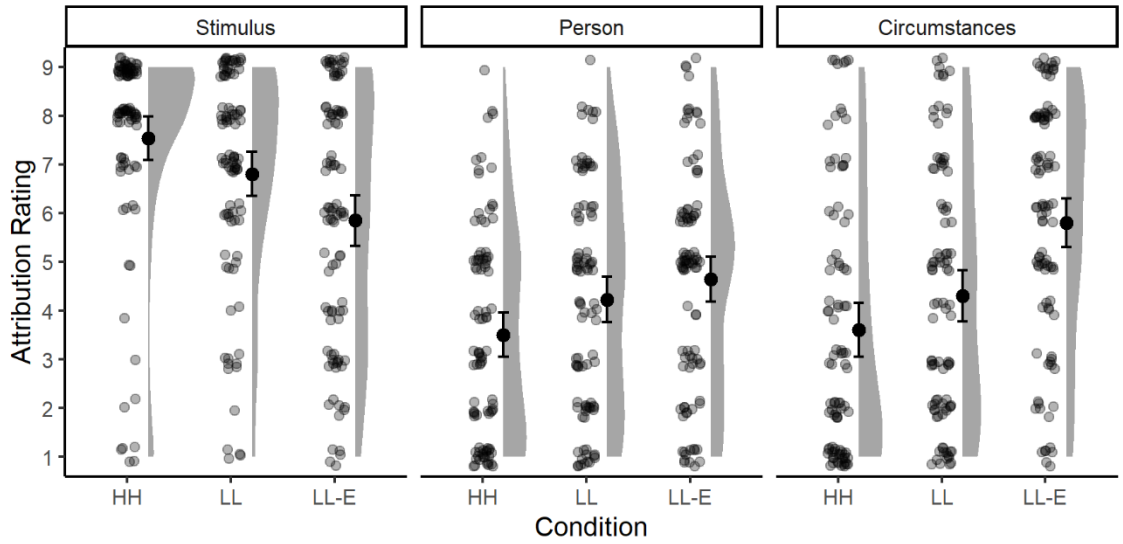
Means and Distributions of Attributions for the Positive Reaction in Experiment 2



Note. The left panels show the results for stimulus attributions, the middle panels show the results for person attributions, and the right panels show the results for circumstances attributions. Dots represent the data of individual participants (with random noise added along the horizontal and vertical axes to improve clarity). Black circles and error bars indicate means and 95% confidence intervals. Grey areas represent the distribution of the data. HH: high consensus and high consistency. LL: low consensus and low consistency. LL-E: low consensus, low consistency, and availability of an explanation.

Figure 8

Means and Distributions of Attributions for the Negative Reaction in Experiment 2



Note. The left panels show the results for stimulus attributions, the middle panels show the results for person attributions, and the right panels show the results for circumstances attributions. Dots represent the data of individual participants (with random noise added along the horizontal and vertical axes to improve clarity). Black circles and error bars indicate means and 95% confidence intervals. Grey areas represent the distribution of the data. HH: high consensus and high consistency. LL: low consensus and low consistency. LL-E: low consensus, low consistency, and availability of an explanation.

The reactions were also rated as significantly more informative in the HH group (M = 7.77, SD = 1.34) than in the LL group (M = 6.17, SD = 2.57), $p < .001$, d = 0.78, [0.46, 1.08], $BF_{10} > 10000$, and than in the LL-E group (M = 5.47, SD = 2.30), $p < .001$, d = 1.21, [0.87, 1.55], $BF_{10} > 10000$. The LL and LL-E groups differed slightly from one another, but the evidence for this difference was weak, p = .026, d = 0.29, [-0.003, 0.575], BF_{10} = 0.95. The variance of informativeness ratings was not significantly larger in the LL group (6.62) than in the LL-E group (5.27), $F(92, 94)$ = 1.26, p = .273. Interestingly, informativeness ratings were

correlated significantly with evaluative ratings, $r = .54$, $t(277) = 10.64$, $p < .001$, as well as with IAT scores, $r = .19$, $t(277) = 3.27$, $p = .001$.

Finally, many participants (68% of the HH group, 53% of the LL group, and 61% of the LL-E group) reported having been aware that we expected to find OEC effects. A smaller but still substantial percentage of participants in each group (respectively 29%, 37%, and 48%) reported having been aware that we expected consensus and consistency (and the availability of an explanation) to have an impact.

Discussion

Experiment 2 again tested the idea that attributions play a role in OEC but relied on a different manipulation. Our main prediction was confirmed: participants who received low consensus and consistency information showed significantly smaller OEC effects than participants who received high consensus and consistency information. Unlike in Experiments 1a-1b, the manipulation influenced not only evaluative ratings but also IAT performance. We did not obtain evidence to support our secondary prediction that including an explanation (i.e., the LL-E condition) would reduce both the size and variance of OEC effects relative to the LL condition. Given participants' responses to the attribution and informativeness questions, this is not very surprising: stimulus attributions and informativeness ratings were significantly but only slightly reduced in the LL-E group relative to the LL group, while the amount of variance did not differ. In other words, it seems that although including the explanation may have helped to move participants slightly further away from stimulus attributions, this impact was too weak to significantly reduce OEC effects. It is worth mentioning, however, that participants in the LL-E group did not rate the CS_{neg} significantly below zero, whereas participants in both other groups did.

General Discussion

In this paper we set out to explore the role that causal attributions play in observational conditioning. Drawing on the attribution literature, we expected that distinctiveness information (Experiments 1a and 1b) as well as consensus and consistency information (Experiment 2) would influence how much a model's emotional reaction towards a stimulus (a cookie) would be causally attributed to that stimulus. Accordingly, we tested whether these types of information influenced how much an observed reaction changed the observer's own evaluations of the stimulus; in other words, whether distinctiveness, consensus, and consistency moderated observational evaluative conditioning (OEC) effects.

Overall, we obtained clear evidence of OEC effects, as reflected by ratings and two different IATs. Moreover, distinctiveness had the expected impact on OEC effects as indexed by evaluative ratings, which seemed to be mostly due to the ratings of the stimulus that was followed by a *negative* reaction. However, OEC effects as indexed by automatic evaluations were not moderated by distinctiveness in the expected manner. In contrast, a combination of consensus and consistency information did moderate both self-reports and automatic evaluations. While including a potential explanation alongside the low consensus and low consistency information served to further reduce stimulus attributions, this was not reflected in the OEC effects. Finally, stimulus attributions were generally affected by our manipulations in the expected manner and mostly correlated with the size of OEC effects measured via self-reports. Taken together, our findings seem in line with the idea that causal attributions play an important role in observational conditioning.

In what follows, we consider the theoretical implications of our findings for observational conditioning research and take a closer look at some of our findings from the perspective of the broader attribution literature. Finally, we consider the limitations of our work as well as promising avenues for future investigation.

Theoretical Considerations

As mentioned in the introduction, our research was inspired by the idea that observational conditioning effects (at least in humans) are mediated by inferential processes (e.g., Baeyens et al., 2001; Kasran et al., 2022b; see also Mitchell et al., 2009). We focused on a specific type of inference, namely causal attributions, and found that manipulations designed to target people's causal attributions (i.e., to what extent they considered the stimulus to be the cause of the model's reaction) influenced observational conditioning. Overall, our findings seem more in line with an inferential than with a (purely) associative explanation of observational conditioning. For instance, it is unclear how one could explain the impact of the consensus and consistency information on OEC effects from a purely associative perspective, especially given that the information was provided only *after* the OEC phase, at which point the relevant associations would presumably have already been formed.

We should note, however, that our findings cannot distinguish between a *purely* propositional account (which assumes that all types of information are encoded in propositions and that all types of evaluations emerge from propositional processes) and a dual-process account (which assigns a role to both associative and propositional processes; e.g., McConnell & Rydell, 2014; Gawronski & Bodenhausen, 2018). In fact, some of our findings seem to fit better with the latter. Because many dual-process theories assume (a) that pairing-based information (such as the co-occurrence of a stimulus and a model's reaction) is encoded via associative processes, (b) that evaluations often emerge from some combination of propositional and associative processes, and (c) that automatic evaluations are more likely than self-reported evaluations to reflect an impact of associations, such theories can easily account for the fact that self-reported evaluations were moderated by distinctiveness information while automatic evaluations were not. However, a purely propositional perspective can explain such patterns as well (albeit in a post hoc manner) by assuming that

participants also form a proposition based on the mere pairings (e.g., De Houwer 2018) and that such a proposition is easier to retrieve automatically than a proposition based on combining pairing-based information with additional information (as some episodic memory models propose; Stahl & Aust, 2018). In sum, our findings can be accommodated by both propositional and dual-process perspectives (depending on the assumptions that are made). Still, they are informative because they add further weight to the idea that theories of observational conditioning should assign an important role to inferential processes.

While most of our main findings were in line with predictions, we also obtained some results that were more unexpected. Specifically, evaluations of the stimulus that was followed by a positive reaction were not always sensitive to our manipulations, the distinctiveness manipulation had only a weak or even no impact on OEC effects, and stimulus attributions were quite strong even in the low distinctiveness (Experiments 1a-1b) and low consensus and consistency (Experiment 2) conditions. However, it is important to realize that we focused on only three variables that are known to influence attributions. When we consider certain qualifications and criticisms of this approach that have been voiced in the attribution literature, some of these more unexpected findings appear to fit quite well with what is known about causal attributions.

First, participants' evaluations of the stimulus that was followed by a positive reaction seemed to be less affected by distinctiveness than evaluations of the stimulus that was followed by a negative reaction. This asymmetry may have been due to the type of stimuli that we used (i.e., cookies). Specifically, people have considerable prior knowledge about cookies in general, which could have a large impact given that consensus, distinctiveness, and consistency do not provide all of the covariation information that may be considered relevant for attributions (e.g., Cheng & Novick, 1990; Försterling, 1989; Pruitt & Insko, 1980). For example, they do not specify how other people usually react to other stimuli of the same type.

These unspecified relations are not simply ignored when making attributions; instead, people tap into their knowledge about the real world to fill in the gaps (Novick et al., 1992) and to determine how to use the information that is explicitly provided (Hilton & Slugoski, 1986). In our case, participants may have assumed that "most people react positively to most cookies" based on what they know about real cookies. If so, they may have been predisposed to attribute a positive reaction to the cookie and made limited use of the information provided to them. In contrast, a negative reaction may have been seen as a counter-normative event which participants were already be more inclined to attribute to the model (see Hilton, 2017), which they then readily did if the information (e.g., low distinctiveness) was in line with this.

Second, distinctiveness had a surprisingly weak or even no impact on OEC effects. One explanation may be that this information was provided prior to the observation phase and was perhaps retrieved only to a limited extent later on, while the consensus and consistency information was encountered immediately before the evaluative measures. However, there may also be a theoretical reason for the relatively weak impact of distinctiveness. An important distinction that has been made in the attribution literature is that between *causal explanations* (i.e., concluding that an event was *caused* by a person or stimulus) and *dispositional attributions* (i.e., attributing a general *characteristic or property* to a person or stimulus; Hilton et al., 1995). In the context of observational conditioning the latter kind of attribution may actually be more relevant: we would expect a negative evaluation of a cookie to depend on the inference that the cookie is bad (i.e., the attribution of negative properties to the cookie). Importantly, research has shown that how people use covariation information may differ for causal explanations and dispositional attributions: while causal stimulus attributions tend to rely more on "contrast" information such as distinctiveness (i.e., what happens when the stimulus is absent), dispositional stimulus attributions tend to rely more on

"generalization" information such as consensus (i.e., what happens in other instances when the stimulus is present; Hilton et al., 1995; Van Overwalle, 1997).

In our studies the distinctiveness information may therefore have been less informative than the consensus and consistency information. Specifically, *low consensus and consistency* information implies a lack of generalization (across persons and time) which may prevent one from inferring the properties of the stimuli, while *low distinctiveness* information merely implies that there is an alternative cause (i.e., the person), which does not necessarily mean that the observed reactions do not align with the actual properties of the stimuli. Similarly, *high consensus and consistency* information implies a strong degree of generalization which allows one to confidently infer the properties of the stimuli, while *high distinctiveness* information merely implies that the observed reactions diverged from how the models usually react, which could lead to several inferences that would not necessarily result in stronger observational conditioning (e.g., some participants mentioned in their open-ended responses that the cookies may not have tasted like typical cookies or may have contained a specific ingredient). Finally, note that if dispositional attributions are indeed more relevant for observational conditioning, the "informativeness" question included in Experiment 2 may actually have been more appropriate than the "stimulus attribution" question (in this regard, it is interesting that only responses to the former were correlated with the size of IAT effects).

Of course, the above considerations do not explain why the IAT effect was numerically larger in the low distinctiveness group than in the high distinctiveness group (although this trend was weak and only significant in Experiment 1b). We do not have a readily available explanation for this unexpected pattern, which may simply be spurious. One possibility is that exposing participants to two different models (which, while necessary to manipulate distinctiveness, was not central to our purposes) introduced a source of noise into

their IAT performance.[2] If so, the pattern observed on the IAT in Experiments 1a-1b may simply have been an artifact of the set-up that we used in those experiments.

Finally, it was interesting that stimulus attributions, even when reduced by our manipulations, remained quite strong (i.e., often still above the midpoint of the scale). This difficulty to draw participants away from stimulus attributions may be related to the more general tendency for people to believe that behaviors that fall into the category of *emotions* – as opposed to behaviors that fall into the categories of voluntary actions or accomplishments, for example – are due to stimuli (e.g., Hilton, 2017; McArthur, 1972).

In sum, what we know about attributions in general can help explain some of the subtleties in our findings. This strengthens our conclusion that attributions guide observational conditioning and paves the way for a more sophisticated analysis of the attributional processes involved in this and other types of (social) learning.

Limitations and Future Directions

The current research has a number of limitations, some of which may inform future studies on this topic.

First, similar to most observational conditioning research but unlike typical attribution research – which usually requires participants to make attributions for a single behavior at a time – the OEC phase showed two reactions, a positive reaction to one CS and a negative reaction to another CS. Because we wanted to influence attributions for both reactions, it was challenging to adapt typical consensus, distinctiveness, and consistency manipulations to the current context. For example, saying that a model likes most cookies would constitute low

[2] For instance, each cookie was related not only to a reaction with a specific valence (i.e., positive or negative) but also to a specific model (i.e., the picky or the non-picky model). Therefore, during the IAT the CS could have induced memory retrieval of the valenced reaction as well as of the model who showed the reaction. In the low distinctiveness condition, for example, the CS_{neg} may have made participants think about the picky model who was said to dislike many things, and who may have been negatively valenced herself as a result. This could have exacerbated participants' evaluative responses to the CS_{neg} relative to the high distinctiveness condition, where the CS_{neg} was related to the non-picky model.

distinctiveness information for the positive reaction but high distinctiveness information for the negative reaction, which meant that we had to show two different models in order to manipulate the distinctiveness of both reactions. Future research could look at situations in which attributions for only one stimulus are targeted and use information that is as close as possible to manipulations that have proven to be effective in attribution research. Relatedly, it is worth noting that we presented the information aimed at manipulating attributions in a different manner than the information about the reactions (i.e., via language versus via videos).[3] Given that videos are likely rather salient and memorable relative to a few sentences of text, it is possible that this choice influenced the results that we obtained (e.g., by making it particularly difficult for participants to ignore or question the observed reactions to the cookies). Therefore, it would be interesting if future research tried to eliminate such differences in presentation format.

Second, although this does not affect our main conclusions in terms of OEC effects, our assessment of attributions was not very fine-tuned. We assessed attributions via rating scales which may have been somewhat ambiguous to participants. In addition, we did not make participants rate all possible causes and combinations thereof (although we did provide the opportunity to supplement their ratings with open-ended responses). In Experiment 2 we also used the phrase "the specific circumstances", which could be interpreted as either a time attribution or an attribution to the specific combination of person, stimulus, and time (Försterling, 1989; see Hewstone & Jaspars, 1987, for an argument that "the particular occasion" may be more appropriate). Finally, the questions were probably more targeted at causal explanations than at dispositional attributions, while the latter may be more relevant to

[3] When we pilot tested the distinctiveness manipulation, we did explore whether we could manipulate this factor by showing additional videos of the models reacting positively or negatively to several other cookies. However, the video-based manipulation did not influence the expected attributions as clearly as the text-based manipulation and created a number of issues (e.g., participants trying to remember all cookie names or comparing the strength of the target reactions to the additional reactions). We therefore opted to use the text-based manipulation in our main experiments.

observational conditioning. Therefore, future research could use more refined measures of attributions and phrase the questions in such a way that participants are asked to report dispositional attributions (see for example Hilton et al., 1995; Van Overwalle, 1997, 2006).

Third, with the exception of the IAT in Experiment 2, our conclusions are based mostly on self-reports. If participants were aware of how we expected them to respond, demand compliance could have played a role in the patterns that we observed (i.e., the predicted impact of distinctiveness, consensus, and consistency on evaluative ratings). Still, the mere fact that many participants reported having been aware of this hypothesis does not necessarily mean that they actually complied with the perceived experimenter demand. In addition, we used a very liberal criterion for hypothesis awareness; far fewer participants mentioned the distinctiveness or the consensus-consistency hypotheses in their open-ended answers. Nevertheless, it remains a possibility that participants intentionally manipulated their responses. Moreover, while evaluations measured with an IAT might be considered more difficult to control than self-reports, research has shown that participants can increase or decrease their IAT scores when instructed to do so, usually by intentionally slowing down on either the congruent or the incongruent blocks (Fiedler & Bluemke, 2005; Röhner et al., 2013). It is therefore possible that some participants in the current study manipulated their IAT performance, in which case our inclusion of the IAT did not provide any safeguard against the impact of demand compliance. In sum, it would be good to replicate the current results in contexts where the hypothesis is less obvious to participants (e.g., because a cover story successfully hides the study's purpose) as well as in contexts that allow for the inclusion of a measure of actual behavior (e.g., letting participants choose one of the two types of cookies to taste themselves).

Fourth, we focused on the formation of evaluations. Future research could extend these ideas to cases in which existing evaluations are changed via observational conditioning,

as well as to other behaviors (e.g., fear). Yet another question is to what extent these ideas could be applied to observational conditioning in children or even non-human species. Although attributions can of course only be assessed with verbal organisms, this does not preclude the possibility that certain determinants of attributions (see below) could moderate observational conditioning in nonverbal organisms (especially given that some of the determinants of attributions seem to overlap with known moderators of classical conditioning; Alloy & Tabachnik, 1984; Eelen, 2018). Perhaps an organism's capacity for observational conditioning might even be related to the extent to which they show sensitivity to manipulations of attributions.

Finally, we focused on only three variables that are known to affect attributions (consensus, consistency, and distinctiveness). However, many other predictions can be derived from applying attribution theories to observational conditioning. For example, Kelley (1973) drew attention to how a given cause may be "discounted" when another plausible cause is already present and producing the effect; conversely, a cause may be "augmented" when the effect occurs in the presence of another, inhibitory cause. Although dispositional attributions may be slightly less sensitive to discounting and augmentation than causal explanations (Van Overwalle, 2006), it would still be interesting to test for observational conditioning when another cause seems to be producing or inhibiting the model's behavior. We also mentioned how attributions may rely on covariation information left unspecified by the three variables. The influence of such information could be investigated by explicitly manipulating it (e.g., Cheng & Novick, 1990; Pruitt & Insko, 1980) or by measuring the assumptions that participants have about it based on prior knowledge (e.g., Novick et al., 1992). It is important to realize that these assumptions will depend heavily on the type of stimulus: while positive expectations are likely attached to cookies, participants may have

completely different assumptions about other stimuli. As a consequence, attributions are probably more easily manipulated for some stimuli than for others.

Taking a step back, the points above suggest that when we focus solely on the information contained within the pairing of a stimulus and a model's reaction, we risk missing much of the larger picture. That is, a wealth of other information (background knowledge, prior experiences, assumptions, and so on) can shape how we symbolically respond to that pairing. This echoes a recent proposal that many types of evaluative learning (including OEC) are actually symbolic instances of learning, rather than non-symbolic instances of learning driven purely by the co-occurrence of two stimuli in space and time (De Houwer & Hughes, 2016, 2020). In other words, if we focus disproportionately on the observed pairings and not enough on how our history of learning is brought to bear on our response to those pairings, we effectively treat observational conditioning as a non-symbolic phenomenon, which seems to go against much of what we know about (human) behavior.

Conclusion

Prior research has suggested that observational conditioning (i.e., a change in behavior due to observing a model's emotional reaction to a stimulus) relies at least partially on inferential processes. Here, we focused on a specific type of inferences, namely causal attributions, and applied knowledge from the attribution literature to derive predictions about factors that may strengthen or weaken observational conditioning effects. Both distinctiveness information and a combination of consensus and consistency information were found to influence the extent to which the model's reaction was attributed to the stimulus, as well as the magnitude of the observational conditioning effect (although distinctiveness had the expected impact only on self-reports, not on automatic evaluations). These results are largely in line with the idea that causal attributions play an important role in observational conditioning, further strengthening the case for an inferential explanation of

this phenomenon and illustrating the potential value of attribution theories for future work on observational conditioning and social learning more generally.

References

Alloy, L. B., & Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review*, *91*(1), 112–149. http://dx.doi.org/10.1037/0033-295X.91.1.112

Askew, C., & Field, A. P. (2008). The vicarious learning pathway to fear 40 years on. *Clinical Psychology Review*, *28*(7), 1249–1265. https://doi.org/10.1016/j.cpr.2008.05.003

Baeyens, F., Eelen, P., Crombez, G., & De Houwer, J. (2001). On the role of beliefs in observational flavor conditioning. *Current Psychology*, *20*(2), 183–203. https://doi.org/10.1007/s12144-001-1026-z

Baeyens, F., Vansteenwegen, D., De Houwer, J., & Crombez, G. (1996). Observational conditioning of food valence in humans. *Appetite*, *27*(3), 235–250. https://doi.org/10.1006/appe.1996.0049

Broeren, S., Lester, K. J., Muris, P., & Field, A. P. (2011). They are afraid of the animal, so therefore I am too: Influence of peer modeling on fear beliefs and approach–avoidance behaviors towards animals in typically developing children. *Behaviour Research and Therapy*, *49*(1), 50–57. https://doi.org/10.1016/j.brat.2010.11.001

Capozzi, F., Bayliss, A. P., Elena, M. R., & Becchio, C. (2015). One is not enough: Group size modulates social gaze-induced object desirability effects. *Psychonomic Bulletin & Review*, *22*(3), 850–855. https://doi.org/10.3758/s13423-014-0717-z

Castelli, L., Carraro, L., Pavan, G., Murelli, E., & Carraro, A. (2012). The power of the unsaid: The influence of nonverbal cues on implicit attitudes. *Journal of Applied Social Psychology*, *42*(6), 1376–1393. https://doi.org/10.1111/j.1559-1816.2012.00903.x

Castelli, L., De Dea, C., & Nesdale, D. (2008). Learning social attitudes: Children's sensitivity to the nonverbal behaviors of adult models during interracial interactions. *Personality and Social Psychology Bulletin*, *34*(11), 1504–1513. https://doi.org/10.1177/0146167208322769

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology*, *58*(4), 545–567. http://dx.doi.org/10.1037/0022-3514.58.4.545

Cook, M., Mineka, S., Wolkenstein, B., & Laitsch, K. (1985). Observational conditioning of snake fear in unrelated rhesus monkeys. *Journal of Abnormal Psychology*, *94*(4), 591–610. http://dx.doi.org/10.1037/0021-843X.94.4.591

De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learning & Behavior*, *37*(1), 1–20. https://doi.org/10.3758/LB.37.1.1

De Houwer, J. (2018). Propositional models of evaluative conditioning. *Social Psychological Bulletin*, *13*(3), e28046. https://doi.org/10.5964/spb.v13i3.28046

De Houwer, J., & Hughes, S. (2016). Evaluative conditioning as a symbolic phenomenon: On the relation between evaluative conditioning, evaluative conditioning via instructions, and persuasion. *Social Cognition*, *34*(5), 480–494. http://dx.doi.org/10.1521/soco.2016.34.5.480

De Houwer, J., & Hughes, S. (2020). Learning to like or dislike: Revealing similarities and differences between evaluative learning effects. *Current Directions in Psychological Science*, 0963721420924752. https://doi.org/10.1177/0963721420924752

Debiec, J., & Olsson, A. (2017). Social fear learning: From animal models to human function. *Trends in Cognitive Sciences*, *21*(7), 546–555. https://doi.org/10.1016/j.tics.2017.04.010

Eelen, P. (2018). Classical conditioning: Classical yet modern. *Psychologica Belgica*, *58*(1), 196–211. https://doi.org/10.5334/pb.451

Egliston, K.-A., & Rapee, R. M. (2007). Inhibition of fear acquisition in toddlers following positive modelling by their mothers. *Behaviour Research and Therapy*, *45*(8), 1871–1882. https://doi.org/10.1016/j.brat.2007.02.007

Fiedler, K., & Bluemke, M. (2005). Faking the IAT: Aided and unaided response control on the Implicit Association Tests. *Basic and Applied Social Psychology*, *27*(4), 307–316. https://doi.org/10.1207/s15324834basp2704_3

Field, A. P. (2006). Is conditioning a useful framework for understanding the development and treatment of phobias? *Clinical Psychology Review*, *26*(7), 857–875. https://doi.org/10.1016/j.cpr.2005.05.010

Försterling, F. (1989). Models of covariation and attribution: How do they relate to the analogy of analysis of variance? *Journal of Personality and Social Psychology*, *57*(4), 615–625. http://dx.doi.org/10.1037/0022-3514.57.4.615

Gawronski, B., & Bodenhausen, G. V. (2018). Evaluative conditioning from the perspective of the associative-propositional evaluation model. *Social Psychological Bulletin*, *13*(3), e28024. https://doi.org/10.5964/spb.v13i3.28024

Gerull, F. C., & Rapee, R. M. (2002). Mother knows best: Effects of maternal modelling on the acquisition of fear and avoidance behaviour in toddlers. *Behaviour Research and Therapy*, *40*(3), 279–287. https://doi.org/10.1016/S0005-7967(01)00013-4

Golkar, A., & Olsson, A. (2016). Immunization against social fear learning. *Journal of Experimental Psychology: General*, *145*(6), 665–671. https://doi.org/10.1037/xge0000173

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality*

and Social Psychology, *74*(6), 1464–1480. http://dx.doi.org/10.1037/0022-3514.74.6.1464

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*(2), 197–216. http://dx.doi.org/10.1037/0022-3514.85.2.197

Haun, D. B. M., Rekers, Y., & Tomasello, M. (2012). Majority-biased transmission in chimpanzees and human children, but not orangutans. *Current Biology*, *22*(8), 727–731. https://doi.org/10.1016/j.cub.2012.03.006

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2021). lab.js: A free, open, online study builder. *Behavior Research Methods*. https://doi.org/10.3758/s13428-019-01283-5

Hewstone, M., & Jaspars, J. (1987). Covariation and causal attribution: A Logical Model of the intuitive analysis of variance. *Journal of Personality and Social Psychology*, *53*(4), 663–672. http://dx.doi.org/10.1037/0022-3514.53.4.663

Hilton, D. J. (2017). Social attribution and explanation. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 645–674). Oxford University Press.

Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, *93*(1), 75–88. http://dx.doi.org/10.1037/0033-295X.93.1.75

Hilton, D. J., Smith, R. H., & Kin, S. H. (1995). Processes of causal explanation and dispositional attribution. *Journal of Personality and Social Psychology*, *68*(3), 377–387. http://dx.doi.org/10.1037/0022-3514.68.3.377

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: The attribution process in person perception. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 2, pp. 219–266). Academic Press.

Kasran, S., Hughes, S., & De Houwer, J. (2022a). Learning via instructions about observations: Exploring similarities and differences with learning via actual observations. *Royal Society Open Science*, *9*(3), 220059. https://doi.org/10.1098/rsos.220059

Kasran, S., Hughes, S., & De Houwer, J. (2022b). Observational evaluative conditioning is sensitive to relational information. *Quarterly Journal of Experimental Psychology*, *75*(11), 2043-2063. https://doi.org/10.1177/17470218221080471

Kelley, H. H. (1967). Attribution theory in social psychology. *Nebraska Symposium on Motivation*, *15*, 192–238.

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, *28*(2), 107–128. http://dx.doi.org/10.1037/h0034225

Kelley, H. H., & Michela, J. L. (1980). Attribution theory and research. *Annual Review of Psychology*, *31*(1), 457–501. https://doi.org/10.1146/annurev.ps.31.020180.002325

Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software*, *4*(40), 1541. https://doi.org/10.21105/joss.01541

Malle, B. F. (2011). Attribution theories: How people make sense of behavior. In D. Chadee (Ed.), *Theories in social psychology* (pp. 72–95). Wiley Blackwell.

McArthur, L. A. (1972). The how and what of why: Some determinants and consequences of causal attribution. *Journal of Personality and Social Psychology*, *22*(2), 171–193. https://doi.org/10.1037/h0032602

McConnell, A. R., & Rydell, R. J. (2014). The Systems of Evaluation Model: A dual-systems approach to attitudes. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual process theories of the social mind* (pp. 204–217). Guilford.

Mineka, S., & Cook, M. (1993). Mechanisms involved in the observational conditioning of fear. *Journal of Experimental Psychology: General*, *122*(1), 23–38. http://dx.doi.org/10.1037/0096-3445.122.1.23

Mineka, S., Davidson, M., Cook, M., & Keir, R. (1984). Observational conditioning of snake fear in rhesus monkeys. *Journal of Abnormal Psychology*, *93*(4), 355–372. http://dx.doi.org/10.1037/0021-843X.93.4.355

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences*, *32*(2), 183–198; discussion 198-246. http://dx.doi.org/10.1017/S0140525X09000855

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin*, *132*(2), 297–326. https://doi.org/10.1037/0033-2909.132.2.297

Novick, L. R., Fratianne, A., & Cheng, P. W. (1992). Knowledge-based assumptions in causal attribution. *Social Cognition*, *10*(3), 299–333. http://dx.doi.org/10.1521/soco.1992.10.3.299

Olson, M. A., & Fazio, R. H. (2004). Reducing the influence of extrapersonal associations on the Implicit Association Test: Personalizing the IAT. *Journal of Personality and Social Psychology*, *86*(5), 653–667. https://doi.org/10.1037/0022-3514.86.5.653

Olsson, A., Nearing, K. I., & Phelps, E. A. (2007). Learning fears by observing others: The neural systems of social fear transmission. *Social Cognitive and Affective Neuroscience*, *2*(1), 3–11. https://doi.org/10.1093/scan/nsm005

Olsson, A., & Phelps, E. A. (2007). Social learning of fear. *Nature Neuroscience*, *10*(9), 1095–1102. http://dx.doi.org/10.1038/nn1968

Orvis, B. R., Cunningham, J. D., & Kelley, H. H. (1975). A closer examination of causal inference: The roles of consensus, distinctiveness, and consistency information. *Journal of Personality and Social Psychology*, *32*(4), 605–616. http://dx.doi.org/10.1037/0022-3514.32.4.605

Pruitt, D. J., & Insko, C. A. (1980). Extension of the Kelley attribution model: The role of comparison-object consensus, target-object consensus, distinctiveness, and consistency. *Journal of Personality and Social Psychology*, *39*(1), 39–58. http://dx.doi.org/10.1037/0022-3514.39.1.39

Röhner, J., Schröder-Abé, M., & Schütz, A. (2013). What do fakers actually do to fake the IAT? An investigation of faking strategies under different faking conditions. *Journal of Research in Personality*, *47*(4), 330–338. https://doi.org/10.1016/j.jrp.2013.02.009

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374. https://doi.org/10.1016/j.jmp.2012.08.001

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Ruble, D. N., & Feldman, N. S. (1976). Order of consensus, distinctiveness, and consistency information and causal attributions. *Journal of Personality and Social Psychology*, *34*(5), 930–937. http://dx.doi.org/10.1037/0022-3514.34.5.930

Skinner, A. L., Olson, K. R., & Meltzoff, A. N. (2020). Acquiring group bias: Observing other people's nonverbal signals can create social group biases. *Journal of Personality and Social Psychology*, *119*(4), 824–838. http://dx.doi.org/10.1037/pspi0000218

Skinner, A. L., & Perry, S. (2020). Are attitudes contagious? Exposure to biased nonverbal signals can create novel social attitudes. *Personality and Social Psychology Bulletin, 46*(4), 514–524. https://doi.org/10.1177/0146167219862616

Stahl, C., & Aust, F. (2018). Evaluative conditioning as memory-based judgment. *Social Psychological Bulletin, 13*(3), e28589. https://doi.org/10.5964/spb.v13i3.28589

Szczepanik, M., Kaźmierowska, A. M., Michałowski, J. M., Wypych, M., Olsson, A., & Knapska, E. (2020). Observational learning of fear in real time procedure. *Scientific Reports, 10*(1), 16960. https://doi.org/10.1038/s41598-020-74113-w

Van Overwalle, F. (1996). The relationship between the Rescorla-Wagner associative model and the probabilistic joint model of causality. *Psychologica Belgica, 36*(3), 171–192. https://doi.org/10.5334/pb.898

Van Overwalle, F. (1997). Dispositional attributions require the joint application of the methods of difference and agreement. *Personality and Social Psychology Bulletin, 23*(9), 974–981. https://doi.org/10.1177/0146167297239007

Van Overwalle, F. (2006). Discounting and augmentation of dispositional and causal attributions. *Psychologica Belgica, 46*(3), 211–234. https://doi.org/10.5334/pb-46-3-211

Zuckerman, M. (1978). Actions and occurrences in Kelley's cube. *Journal of Personality and Social Psychology, 36*(6), 647–656. http://dx.doi.org/10.1037/0022-3514.36.6.647

Contributions

SK: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, visualization, writing – original draft, writing – review & editing. SH: conceptualization, funding acquisition, methodology, supervision, writing – review & editing. JDH: conceptualization, funding acquisition, methodology, resources, supervision, writing – review & editing. TB: conceptualization, methodology, supervision, writing – review & editing. All authors approve the publication of this version of the article.

Acknowledgements

We would like to thank Eva Sengeleng for helping to record and edit the videos.

Funding Information

This research was conducted with the support of PhD fellowship FWO18/ASP/119 from the Research Foundation Flanders (FWO) to SK and grant BOF16/MET_V/002 from Ghent University to JDH. TB is supported by KU Leuven grant C16/19/002.

Competing Interests

We have no competing interests.

Data Accessibility Statement

All research materials, data, and R code used for data processing and analysis are available on the Open Science Framework (https://osf.io/ghd7f), although the modelling videos have been replaced by anonymized versions because the actors did not consent to publishing them in their original form.