The Shared Features Principle: If two objects share a feature, people assume those objects also share other features

Sean Hughes, Jan De Houwer, Simone Mattavelli, & Ian Hussey

In this paper we introduce the *shared features principle* which refers to the idea that, when two stimuli share one feature, people often assume that they share others features as well. This principle can be recognized in several known psychological phenomena, most of which were until now never considered to be related in this way. To illustrate the generative power of the principle, we report eight pre-registered studies ($N = 1614$) in which participants completed an acquisition phase containing three stimuli: a neutral target, a positive source, and a negative source. Our results indicate that behavioral intentions, automatic evaluations, and self-reported ratings of a target object were influenced by the source object with which the target shared a feature. This was even the case when participants were told that there was no relation between source and target objects. Taken together, the shared features principle appears to be general, reliable, and replicable across a range of measures in the attitude domain. We close with a discussion of its theoretical implications, relevance to many areas of psychological science, as well as its heuristic and predictive value

Scientific principles are valuable because they highlight commonalities amongst many different empirical phenomena. In doing so, they not only create order within existing scientific knowledge (i.e., their heuristic or organizing function) but also point us towards new and previously undiscovered instances of that principle (i.e., their predictive function). In this paper, we introduce a new principle to the realm of psychology. This principle, which we refer to as the *shared features principle*, postulates that when stimuli share one feature, people will assume that those stimuli share other features as well. We first provide an overview of existing phenomena in which the shared features principle can be recognized. We then consider the principle itself in more detail by relating it to concepts that can be applied to a wide variety of phenomena. Finally, we illustrate the predictive value of the principle by demonstrating novel instantiations that control for alternative factors.

Let us first consider existing phenomena that seem to represent instances of the shared features principle. Take the minimal group effect in social psychology (Otten, 2016; Tajfel, Billig, Bundy, & Flament, 1971). Research shows that when individuals are arbitrarily assigned to the same group based on some shared feature (e.g., similar clothing item or preference for certain paintings), people assume that those individuals share other features as well (e.g., that others will share the participant's own traits; van Veelen, Otten, Cadinu, & Hansen, 2016). In the context of stigmatization, the mere proximity effect shows that when a stigmatized and non-stigmatized person share a similar physical location to one another, people assume that they also share other features (e.g., a normal weight individual will be stigmatized more when they stand next to an overweight individual; Hebl & Mannix, 2003).

In consumer and marketing psychology, research on counterfeit brands shows that these brands intentionally imitate the physical properties of (and thus share features with) high status brands in the hope this will influence assumptions about, and ultimately consumption of, the fake brand itself (e.g., assumptions that it is also high in quality, status, and worth purchasing; Phau & Teah, 2009). In moral psychology, when one person (John) is accountable for his past actions (e.g., membership of the Nazi party) and is said to share a feature with a second person (Tom) (e.g., John and Tom are part of the same family), this shared feature influences the assumptions people make about the latter's moral accountability (e.g., they assume that Tom is also morally responsible for his family member's actions; Uhlmann, Zhu, Pizarro, & Bloom, 2012). Finally, the shared features principle can also be recognized in learning psychology. In evaluative conditioning (EC), for instance, the fact that a neutral stimulus shares a similar time and location with a valenced stimulus often leads the former to acquire properties of the latter (Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010). This is also true for attribute conditioning: when an unknown person and a known athlete share one feature (spatio-temporal properties), people assume that they also share other features as well (the CS is viewed as being more athletic or healthy: Unkelbach & Högden, 2019).

If we take a step back, set the specific stimuli and responses to the side, and search for commonalities between the above, then we see that each phenomenon involves a broadly similar situation: one where people make assumptions about the properties of one stimulus (e.g., how positive or negative a person, group, or brand product is) based on the fact that it shares some other feature (e.g., physical or spatio-temporal properties) with a second stimulus (another person, group, or brand product).

If we are correct, then there are remarkable similarities between seemingly different domains in psychological science, which offers many new opportunities for cross-fertilization. Today the aforementioned effects are typically studied in isolation, with different researchers busy documenting the moderators and mediators of the behaviors they are interested in, and rarely interacting with their colleagues (or drawing on findings) from other intellectual domains. This does not have to be the case. In addition to illuminating previously hidden similarities and differences between psychological phenomena (heuristic value), the shared features principle also opens up entirely new avenues for study (predictive value). Before unpacking its heuristic and predictive value, we will first specify the principle itself. We will do so by drawing on a recently developed conceptual framework that centers on the idea of feature transformation (De Houwer, Richetin, Hughes, & Perugini, 2019).

Feature Transformation

Because the shared features principle applies to a wide range of known, and still to-be-discovered empirical phenomena, it is best described in abstract terms that do not refer to a specific phenomenon. De Houwer et al. (2019) recently introduced a framework that may prove useful in this regard. Their framework consists of four abstract concepts: source features, target features, source objects, and target objects. According to this perspective, 'objects' are broadly defined as any potential stimulus or behavior: they can refer to people, animals, inanimate items, and even responses. Features are defined as any assumed state of an object. These states can have multiple values and can relate to many different properties, from physical (e.g., height), to psychological (e.g., intelligence, valence), and behavioral (e.g., the way in which an object responds to its environment; also see Tversky, 1977).

The shared features principle is concerned with two types of features: *source features* and *target features.*[1] Target features are those features of an object about which assumptions are being made. Source features are those features of an object which give rise to assumptions about target features. The object that possesses the target feature is called the target object whereas the object that possesses the source feature is called the source object.

The target feature is typically the dependent variable whereas the source feature is the independent variable under investigation. The value of a source feature can be varied in order to investigate if this influences the corresponding value of the target feature. To illustrate this more clearly, take the halo effect. Here features of the source object (e.g., how attractive a person is) lead people to make assumptions about features of a target object (e.g., how intelligent that person is). The source and target feature can belong to the same object (e.g., judging how intelligent an attractive person is) or to different objects (e.g., judging how intelligent the partner of an attractive person is; see Forgas & Laham, 2016, for a review). In studies on the halo effect, the value of a source feature (perceived attractiveness) is varied to investigate if this influences the value of a target feature (perceived intelligence).

When a feature of a source object influences the assumptions made about a feature of a target object, *feature transformation* is said to take place. The term 'transformation' highlights that a source feature can give rise to assumptions about the features of a target object, and that the latter can change in ways that are similar or different to the former. To illustrate, let's return to the halo effect wherein features of a source object (e.g., how attractive a person is) influence assumptions about target object features (e.g., how intelligent a person is). It may be that an assimilative halo effect emerges for men (attractive males are thought to be competent in job hiring situations) whereas contrast effects emerge for women (attractive females are thought to be less competent in certain job hiring situations; see the 'beauty is beastly effect';

[1] We distinguish target and source feature from target and source objects because the target and source features can either be part of the same object or they can be part of different objects (for a more detailed treatment see De Houwer et al., 2019). In the case of shared features effects the source and target objects are typically different.

Paustian-Underdahl, & Walker, 2016). The term transformation captures both possibilities.

The Shared Features Principle

The shared features principle tells us why a wide variety of feature transformation effects occur: it implies that when source and target objects *share* some feature with one another people will make assumptions about other features of the target object. Verifying that two objects share a feature involves the identification of a feature that is part of both objects (also see Tversky, 1977). For instance, in minimal group effects, source and target objects both independently possess a common feature (e.g., the color of the clothing that they wear).[2]

The shared features principle is grounded in the phenomenon of generalization (see Ghirlanda & Enquist, 2003). Broadly speaking, generalization refers to the transfer of properties from one stimulus to another when those stimuli are similar along some dimension. The shared features principle implies that one can vary similarity by manipulating the features of objects (see Tversky, 1977, for a justification of this assumption) and builds on recent research showing that generalization is not limited to similarity with regard to perceptual features but also encompasses similarity at the symbolic level (Hughes et al., 2018). It also extends beyond simple generalization by allowing for a transformation of features (and not merely a transfer of features) to take place from one stimulus to another.

Finally, just like any scientific principle, the shared features principle does not *always* hold but does so only under certain conditions (e.g., it is likely that the shared feature needs to be salient). In fact, one of the aims of research is to uncover the moderators of scientific principles. Like all functional scientific principles (e.g., gravity), the principle does not specify the mechanism by which instances of the principle are brought about, nor does it assume that the same mechanism mediates all instances of the principle. Although we will speculate about the mental mechanism mediating shared features effects at the end of our paper, the main aim of the paper is to introduce the principle itself and illustrate its heuristic and predictive value.

Heuristic Value. The shared features principle has considerable heuristic value. As illustrated above, it can be applied to a wide variety of existing phenomena that were never previously viewed as being connected (e.g., it highlights commonalities and differences between effects in person perception and counterfeit branding). The unifying nature of the principle can be further strengthened by using the terms of the feature transformation framework. Until now, the social, persuasion, marketing, moral, and learning psychology literatures each adopted different terms when describing instances of the shared features principle. As a result, there is currently a multiplicity of concepts that can undermine our ability to 'see the forest through the trees' (i.e., to identify what is genuinely similar or different between various types of shared feature effects). For instance, evaluative and attribute conditioning research refers to conditioned (CS) and unconditioned stimuli (US), whereas operant conditioning research refers to discriminative stimuli (Sd) and reinforcers (Sr). These terms are rarely used in research on marketing, halo, moral accountability, and person perception. Rather these domains employ idiosyncratic terms that typically refer to the specific properties of the features and objects being studied (e.g., the status of products).

The feature transformation framework in general, and shared features principle in particular, circumvent this issue by providing a common or 'universal' language that allows one to describe and functionally explain many phenomena using a limited set of terms. Concepts such as source/target and object/feature can be used to describe shared feature effects that are typically studied (under different names) in different domains. For example, in impression formation research, we can refer to the fact that a source object (Bob) possesses a certain feature (e.g., is violent). When people then learn that a target object (Mike) shares some other feature with the source (e.g., they have physical characteristics in common), we can say that this shared feature leads people to make assumptions about the target object's features (e.g.,

[2] Note that this requirement excludes feature transformation effects that are due to related features. For instance, in endowment effects (Kahneman, Knetsch, & Thaler, 1990) and mere ownership effects (Beggan, 1992), people ascribe more value to things that they own. In this case feature transformation does take place (e.g., people assume that the object is higher in value or status because it is related to the self). However, this change in assumptions about the object is not due to the fact that the person and object share a feature. In fact, the person-object relation implies that the person and object have different features (i.e., the person is the owner whereas the object is owned). Although such related features (i.e., features that are different aspects of a single relation) could also underlie feature transformation, we differentiate related features from shared features as functional causes of feature transformation. One reason for this is that it is unclear how related features influence similarity whereas there are good arguments for assuming that shared features increase similarity (Tversky, 1977). Hence, whereas feature transformation because of shared features might be grounded in generalization (*see below*), feature transformation because of related features might not be.

that Mike is also violent). The same concepts can also be applied to conditioning research. Here too people learn that a source object (US) possesses a certain feature (e.g., is valenced as in EC or athletic as in attribute conditioning). They then learn that a target object (CS) shares some other feature with the source (e.g., both are presented together in space and time). As a result, people make assumptions about the target object's features (that the CS is also valenced or athletic). The very same concepts can be used when dealing with different types of learning, stigmatizing, prejudice, branding, and so on.

In short, the shared features principle allows researchers to conceptualize and speak about effects in ways that (a) enhance communication within and between intellectual domains, (b) prevent fragmentation, confusion, or conflict resulting from the use of multiple terms to describe the same underlying phenomenon, and (c) reveals similarities and differences between phenomena. While acknowledging important differences between domains, it argues that many effects involve four basic elements (source object, target object, source feature, and target feature), a situation wherein the source and target share one feature, and as a result, new assumptions are made about other target object features.

Predictive Value. The shared features principle also has predictive value and allows us to view old phenomena in new ways. Take EC, for instance, which can be defined as a change in evaluation due to regularities in the presence of two stimuli (see De Houwer & Hughes, 2020). Most researchers think of EC effects merely in terms of the spatio-temporal properties of stimuli, that is, the fact that stimuli are presented together in space and time. Yet our account takes a different perspective. It argues that EC effects may actually be due to the fact that the CS and US *share* a feature with one another, and in EC studies, this shared feature just so happens to be the time and location at which they are presented. If correct, then the crucial element in EC is the fact that stimuli share a feature and not the mere fact that they are paired in space and time. Note that this new way of thinking does not draw EC effects into question – simply our prior explanation of the observed changes in liking. In other words, we are not questioning that regularities in the presence of two stimuli can lead to changes in liking. Rather we are re-conceptualizing spatio-temporal contiguity as just one way to induce a shared feature effect. This new perspective leads to the prediction that EC-like effects can also be found when stimuli share a feature other than their spatio-temporal presence (e.g., the color in which stimuli are presented). Verifying this prediction would support the idea that EC is just one instance of a much broader class of share features effects and would illustrate the predictive power of the shared features principle.[3]

The Current Research

With the above in mind we carried out eight studies. Each employed a broadly similar format which we will preview here. We first asked participants to complete an acquisition phase. During this phase a series of trials were presented wherein three stimuli simultaneously appeared onscreen: a positive source object, a negative source object, and a neutral target object. We then manipulated the extent to which the target object shared a feature with a certain source object. In Experiments 1-3 the shared feature was the color in which stimuli were presented: half of the trials presented the neutral target object in the same color as the positive source whereas the other half presented the neutral target object in the same color as the negative source object. In Experiment 4 the shared feature was the size of the stimuli: half of the trials presented the neutral target object in the same size as a positive source whereas the other half presented the target object in the same size as a negative source. In Experiment 5 the shared feature was conceptual in nature. We first trained a class of conceptually related color stimuli (*Blue-Same-Yellow* and *Green-Same-Purple*) and then, during the acquisition phase, presented a neutral target object in either blue or green, along with a positive source in yellow and a negative source in purple. Experiments 6-8 excluded alternative explanations for this effect, replicated our prior findings with yet another dependent measure (evaluative priming), extended our findings into a socially relevant domain (person perception), and show that they hold even when people are told that there is no relation between the source and target, and that these object are being presented randomly together (further reducing the likelihood that the effect is driven by demand or represents a communication effect).

Following the acquisition phase, we assessed for evaluations using self-report ratings and an indirect procedure (either the Implicit Association Test [IAT] or the Evaluative Priming Task [EPT]). We added these latter procedures as scores obtained from these tasks are assumed to reflect more automatic

[3] Often, EC is defined as changes in liking that are due to the pairing of stimuli (e.g., De Houwer, 2007). Bar-Anan and Balas (2018) correctly pointed out that the concept 'pairing' could be understood in a broad sense as 'putting together' or 'joining to form a pair'. In that sense, also presenting two stimuli in the same color could be seen as a form of pairing. However, as is common in the literature on EC, we use the term 'pairing' only in the sense of 'spatio-temporal pairing' and thus limit EC to its standard meaning of changes in liking that are due to the spatio-temporal pairing of stimuli.

evaluations that can influence subsequent behavior in unique ways (De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009). If changes in liking are driven by mere spatio-temporal contiguity then we would expect to see similar and ambivalent evaluative responses towards both target objects (given that they were both repeatedly paired with positive *and* negative source objects). Yet if those same effects are driven by the fact that the target and source share another feature (e.g., color, size, location) then we would expect to observe positive evaluations of one and negative of the other. If our account is correct, changes in liking should be moderated by a range of different features that are shared by stimuli.

Experiment 1

Method

Participants and design. A total of 114 English-speaking volunteers (62 females; M_{age} = 33.12, SD = 8.39) participated online via the Prolific Academic website (https://prolific.ac) in exchange for a monetary reward (€1.50). The experiment was programmed in Inquisit 4.0 and hosted via Inquisit Web (Millisecond Software, Seattle, WA). It involved a single-factor between-subjects design (*Shared Feature*: target stimulus shared color with positive vs. negative source object), with self-reported ratings and IAT effects as the main dependent variables. Three method variables were manipulated between participants: *evaluative task order* (self-reports vs. IAT first), *IAT block order* (learning [acquisition] phase consistent vs. inconsistent first) and *stimulus assignment* (which target object appeared in the same color as positive or negative source objects). The sample size was determined prior to data collection on the basis of a power analysis. We stopped data-collection when 114 participants had completed all measures of the experiment to ensure that we would have sufficient statistical power to detect medium effects (planned sample size after exclusions = 110 which gives power = 0.80 to find an effect size of d = 0.47 at alpha = 0.05, two-tailed; or power = 0.95 to find an effect size of d = .63). Note that a similar analytic strategy was used in Experiments 1-8 and planned sample sizes were therefore similar (excepted where noted). The study designs were pre-registered, and are available, along with the raw data, and analytic plans for this and all other experiments on the Open Science Framework website (osf.io/pqm9v). We report all manipulations, measures, and studies run. All data were collected without intermittent data analysis.

Materials

Stimuli. Two nonwords (Morag and Struan) served as the target objects. Six positive (*rainbow, pleasure, smile, love, paradise, joy*) and six negative words (*war, cancer, hate, hell, misery, vomit*) served as the positive and negative source objects.

IAT. The two nonwords served as one set of target stimuli and the words "Good" and "Bad" as another. Eight positively valenced and eight negatively valenced adjectives served as one set of attribute stimuli (*fantastic, great, nice, good, pleasant, wonderful, amazing, happy versus terrible, disgusting, nasty, horrible, sick, awful, sad, unpleasant*) and the two nonwords served as the second set.

Procedure

Participants were first provided with a general overview of the experiment and then asked for their informed consent. The study consisted of three phases: acquisition phase, evaluative measures, and exploratory questions.

Acquisition phase. Prior to the acquisition phase participants were told the following: "you will encounter two new words: MORAG and STRUAN. You have probably never seen these words before. These words will appear on the screen together with two other words. The new word (MORAG or STRUAN) and the other words will initially appear in white. Then the color of the three words will change. Please pay close attention to the colors of each word and how they change. You will be asked some questions about this later in the study".

The acquisition phase then consisted of three blocks of 16 trials (48 total), with each block containing two types of trials: one trial in which one nonword was eventually presented in the same color as positively valenced words, and another trial in which a second nonword was eventually presented in the same color as negatively valenced words. Specifically, three stimuli simultaneously occurred onscreen during each trial: a neutral target object (either Morag or Struan) along with a positive and negative source object (i.e., a positively and negatively valenced word). All three stimuli were initially presented in white against a black background for 3000ms. On certain trials, the first nonword (e.g., Morag) and the positively valenced word both changed to the same color (e.g., blue) whereas the negatively valenced word changed to a different color (e.g., green). On other trials, the second nonword (e.g., Struan) and the negatively valenced word both changed to one color (e.g., yellow) whereas the positively valenced word changed to another color (e.g., purple). All stimuli remained onscreen for another 3000ms and were then removed, followed by an inter-trial interval of 1250ms, and the next trial. Stimulus color (i.e., blue, green, yellow and purple) was varied across trials, so that none of the colors could assume any specific positive or negative value (see Figure 1). Assignment of Morag or Struan to share a similar feature (color) as positive or negatively valenced words was randomly counterbalanced across participants.

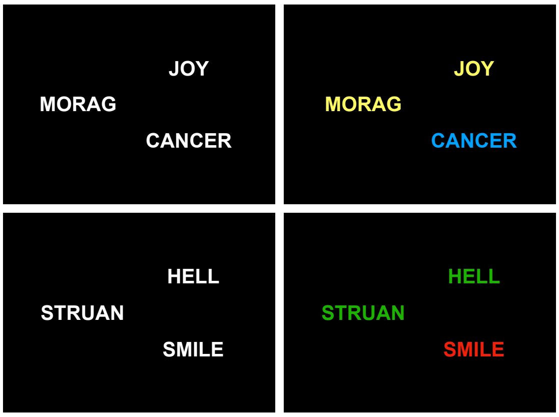

Figure 1. Illustration of the two types of trials during the acquisition phase of Experiment 1. During the first half of the trial (*left*) the neutral target and two source stimuli were presented in white. During the second half of the trial (*right*) the target and one of the sources changed to the same color whereas the second source stimulus changed to a different color.

Self-reported ratings. Self-reported evaluations of Morag and Struan was indexed using four semantic differential scales. On each trial, one of the two nonwords was presented and participants had to evaluate it using a scale ranging from -5 to +5 with 0 as a neutral point. The four end-points of the scales were: *Negative-Positive*, *Pleasant-Unpleasant*, *Good-Bad*, *I Like It-I Don't Like It.* A mean evaluative rating was calculated for each nonword by averaging scores from the four scales.

IAT. We sought to index automatic evaluations using an IAT. Participants were informed that a series of words would appear one-by-one in the middle of the screen and that their task was to categorize those items to their respective target (Morag or Struan) or attribute categories ('Good' and 'Bad') as quickly and accurately as possible. They were told that the two nonwords they had previously encountered (Morag and Struan) as well as the words "Good" and "Bad" (attributes) would appear on the upper left and right sides of the screen and that stimuli could be assigned to these categories using either the left ('E') or right keys ('I'). Each trial began with the presentation of a target or attribute stimulus. If the participant categorized the word correctly - by selecting the appropriate key for that block of trials - the stimulus disappeared from the screen and the next trial began. In contrast, an incorrect response resulted in the presentation of a red "X" which remained on-screen until the correct key was pressed. Overall, each participant completed seven blocks of trials. The first block of 20 practice trials required them to sort Morag and Struan into their respective categories, with Morag assigned to the left ('E') key and Struan to the right ('I') key. On the second block of 20 practice trials, participants assigned positive words to the "Good" category using the left key and negative words to the "Bad" category using the right key. Blocks 3 and 4 (20 and 40 trials, respectively) involved a combined assignment of target and attribute stimuli to their respective categories. Specifically, participants categorized Morag and positive words using the left key and Struan and negative words using the right key. The fifth block of 20 trials reversed the key assignments, with Struan now assigned to the left key and Morag to the right key. Finally, the sixth and seventh blocks (20 and 40 trials respectively) required participants to categorize Morag with negative words and Struan with positive words. Note that assignment of Morag and Struan to the left/right categories was counterbalanced across participants.

Behavioral Intentions. We also assessed if the acquisition phase altered behavior intentions towards the two nonwords. Participants were presented with two brand products labeled with either Morag or Struan. They were asked to indicate which of these products they would be willing to try in a supermarket and given the following five options: *I would try Morag*, *I would try Struan*, *I would try Morag and Struan*, *I would try neither*, *I don't know.*

Exploratory Questions. At the end of the study we asked a number of final questions for purely exploratory reasons. First, we probed to see if participants had *source valence awareness*: "*In the first part of the experiment (when words appeared initially in white and then switched their color), MORAG/STRUAN was always presented with two words. What was the meaning of those words?*', response options (Both words always had a positive meaning, Both words always had a negative meaning, One word always had a positive meaning and the other one a negative meaning, I don't remember). Second, we probed for *target-source contingency awareness*: "*In the first part of the experiment (when words appeared initially in white and then switched their color) did the color of MORAG/STRUAN always switch to the same color as…*" and given the following response options (The positive word that was also on the screen, The negative word that was also on the screen, I don't remember). A *manipulation check* was also included to examine if participants wrote down the target-source contingencies during the task: "*Think back to the first part of the experiment (i.e., when three words were paired onscreen). Did you ever take notes (or write down) what happened in order to help* *you figure out what was going on? Please be honest here (you will receive payment regardless of what you say*" and provided with an open-ended response option. We then probed for *demand compliance* – first for the self-reported ratings ("*Earlier you rated MORAG and STRUAN as being either positive, neutral, or negative. Did you base your ratings NOT on how you actually felt about those words but ONLY on what you thought the researchers wanted you to say*?") and then for the IAT ("*Earlier you completed the Implicit Association Test. Did you base your performance in that task NOT on your best efforts to perform the categorizations as quickly and accurately as possible but on your attempt to influence your speed or accuracy in order to go along with what you thought the researchers wanted you to feel about the words*?"). In both cases the response options were (Yes, No, I don't know). *Reactance* was probed in a similar way - first for the self-reported ratings ("*Earlier you rated MORAG and STRUAN as being either positive, neutral, or negative. Did you consciously resist what you thought the researchers wanted you to feel about those words*?") and then for the IAT ("*Earlier you completed the Implicit Association Test. Did you try to influence your speed or accuracy in order to consciously resist what you thought the researchers wanted you to feel about those words*"). In both cases the response options were (Yes, No, I don't know). Finally, we probed for *shared feature awareness*: "*During the first part of the study, did you notice that the color of MORAG and STRUAN switched to the same color as either positive or negative words*?", and *influence awareness*: "*Did this influence how you responded to MORAG and STRUAN*?". An open-ended response option was provided in each case. It should be noted that we did not examine if performance on these questions moderated evaluations within each individual study. Rather we did so meta-analytically (where power was available to answer such a question; for an overview of these exploratory questions see Table 1).[4]

Results

Analytic Strategy. A series of Welch's independent sample *t*-tests (along with Cohen's *d* effect sizes and their 95% confidence intervals) were carried out on the rating and IAT data to determine whether evaluations of a nonword (dependent variables) differed as a function of the features it shared with a source object (e.g., the fact that one nonword shared a feature [i.e., color] with positively valenced words, and the second nonword shared a feature with negatively valenced words; independent variable). The behavioral intentions data was analyzed using multi-nominal logistic regression models to assess whether participants were more likely to choose a certain nonword on the basis of shared features. A similar analytic strategy was used in Experiments 2-5.

Data Preparation.

Exclusions. We excluded data from eight participants who did not complete the entire session. The data of participants who had IAT error rates above 30% across the entire task or above 40% for any one of the four critical blocks, or who responded faster than 400ms on more than 10% of trials were excluded ($n = 3$). This led to a final sample of 103 participants.

Self-reported ratings. For each nonsense word, self-reported ratings from the four semantic differential scales were averaged. This led to two mean evaluative scores – one for the first nonsense word and another for the second nonsense word. A difference score was then calculated by subtracting the latter from the former. Positive values indicate a relative preference for the nonword that eventually shared a color with a positively valenced word over the nonword that shared a color with a negatively valenced word. Negative values indicate the opposite.

IAT. Following the recommendations of Greenwald, Nosek, and Banaji (2003), response latency data were prepared using the D scoring algorithm. The resulting D-IAT scores reflect the difference in mean response latency between the critical blocks divided by the overall variation in those latencies. The IAT score was calculated so that positive values reflected a relative preference for the nonwords that eventually shared a color with a positively valenced word relative to the nonword which eventually shared a color with a negatively valenced word. Negative values indicated the opposite response pattern.

Hypothesis Testing.

IAT. IAT scores differed as a function of whether a nonword shared its color with a positively or negatively valenced word, $t(98.12) = 6.63$, $p < .001$, $d = 1.31$, 95% CI = [0.88, 1.74], $BF_{10} > 10^5$. When Morag shared a color with a positively valenced word and Struan shared a color with a negatively valenced word, participants showed a relative preference for Morag over Struan ($M = 0.37$, $SD = 0.46$). When the color contingencies were reversed, participants preferred Struan over Morag ($M = -0.23$, $SD = 0.45$).

[4] During the review process we came to appreciate that several items could have been conceptualized better in our pre-registered documents (see osf.io/jer49). We decided to change how these questions are described in the paper itself to better reflect what we now believe the items to actually target (e.g., we now refer to one item as a '*shared feature awareness*' question instead of a '*hypothesis awareness*' question). These changes have been noted in the documentation attached to this OSF project.

Table 1. Percentage of sample who were source valence aware, target-source contingency aware, demand compliant, and reactant in each experiment.

			Demand compliant		Reactant	
Experiment	Source valence aware	Target-source contingency aware	IAT	Self-report	IAT	Self-report
1	65.0	71.8	17.5	15.5	19.4	16.5
2	65.0	62.1	14.6	11.7	9.7	15.5
3	79.4	79.4	18.6	22.7	11.3	14.4
4	78.1	62.0	7.5	17.6	9.6	15.0
5	74.9	45.3	14.0	19.6	8.4	15.1
6	-	-	1.7	8.2	7.4	15.2
7	78.1	56.8	7.0	14.1	10.5	10.0
8	94.9	-	1.4	13	10.9	22.5

Self-reported ratings. Self-reported ratings also differed as a function of whether a nonword shared its color with a positively or negatively valenced word, $t(98.32) = 8.33$, $p < .001$, $d = 1.65$, 95% CI = [1.20, 2.10], $\text{BF}_{10} > 10^6$. When Morag shared a color with a positively valenced word and Struan shared a color with a negatively valenced word, participants showed a relative preference for Morag over Struan ($M = 3.33$, $SD = 4.60$). When the color contingencies were reversed, participants preferred Struan over Morag ($M = -4.15$, $SD = 4.48$).

Behavioral intentions. Data from the behavioral intentions question were entered into a multinomial logistic regression with Morag as the reference category. Only results from the direct comparison of Morag and Struan were relevant to the shared feature hypothesis and are reported here (i.e., hypotheses do not refer to the selections of neither stimulus or both stimuli). Results demonstrated that participant's intentions towards Morag relative to Struan differed between the two shared feature conditions. The odds that a participant would choose the nonword that shared a feature with a positive word (versus the one that shared a feature with a negative word) were OR = 13.66, 95% CI = [3.56, 52.44], $p = .0001$.

Discussion

Results provide a novel demonstration of the shared features principle in the attitude domain that goes beyond the impact of mere spatio-temporal co-occurrence. Although a neutral target object was repeatedly paired with both positive and negative source objects, it acquired the valence of the source object that it shared a feature (color) with it. Specifically, target objects that appeared in the same color as positive source objects were rated positively whereas target objects that appeared in the same color as negative source objects were rated negatively. We also obtained evidence for this shared feature effect as indexed by automatic evaluations and behavioral intentions.

Experiment 2

In our second experiment we set out to replicate and extend our initial findings. In Experiment 1, participants completed an acquisition phase during which all stimuli were initially presented in white and only later changed to the same or a different color. In Experiment 2, however, we presented all stimuli in the same color during the first half of the trial. During the second half, we switched the color of one the valenced words, while keeping the color of the other valenced word the same as the neutral nonword. This modified design allowed us to differentiate between two explanations of our effects. The first (*shared feature hypothesis*) argues that an overlap in some stimulus feature (in this case color) will lead people to indicate that those same stimuli share other properties (valence). If so, then we should expect a similar pattern of findings as obtained in Experiment 1. A second possibility (*salience hypothesis*) entails that people's attention is fixated on any salient change in the context. Assuming that the effects of spatio-temporal contiguity is magnified when one or both of the contiguous events are salient (e.g., Rescorla & Wagner, 1972), one could argue that the change in liking for nonword could have resulted from the mere spatio-temporal contiguity between the nonword and salient valenced word (i.e., the valenced word whose color changed). This alternative account entails that the observed effect was an instance of EC (i.e., a feature transformation effect due to the sharing of spatio-temporal properties) rather than a feature transformation effect that is due to the sharing of color. If so, then the neutral nonword should acquire the valence of the valenced word which switches color within the trial (i.e., the salient source stimulus), leading to the opposite effect predicted by the shared features principle.

Method

Participants and design. 118 participants (67 females; $M_{age} = 32.3$, $SD = 8.6$) took part in the study via the Prolific Academic website.

Procedure

A similar procedure was used as in Experiment 1 with the exception of the acquisition phase.

Acquisition phase. Prior to the acquisition phase participants were told the following: "In the first part you will see two new words: MORAG and STRUAN. You have probably never seen these words before. These words will appear onscreen together with two other words. The new word (MORAG or STRUAN) and other words will initially appear in one color. Then the color of one of the words will change. Please pay close attention to the colors of each word and how they change. You will be asked some questions about this later on."

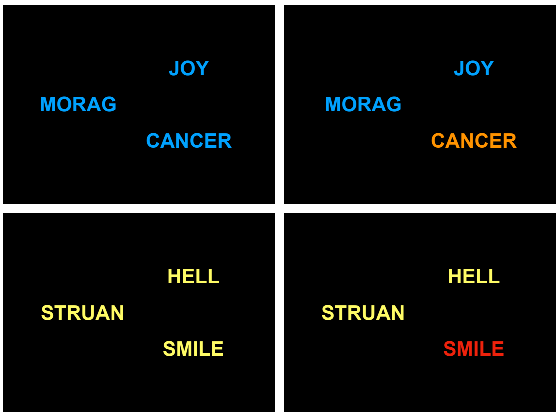

Figure 2. Illustration of the two types of trials during the acquisition phase of Experiments 2 and 3. During the first half of the trial (*left*) the neutral target and two source stimuli were presented in the same color. During the second half of the trial (*right*) the target and one of the sources remained the same color whereas the second source changed to a different color.

Training once again consisted of three blocks of 16 trials (48 total), with each block containing two different types of trials: one in which one nonword (e.g., Morag) stayed the same color as positive words, and another in which a second nonword (e.g., Struan) stayed the same color as negative words. Unlike Experiment 1, the nonword and two valenced words were initially presented in the same color for 3000ms. During one type of trial, one nonword and the positively valenced word remained in the same color (e.g., blue) whereas the negatively valenced word changed color (e.g., purple). During the second type of trial, the second nonword and the negatively valenced word remained in the same color (e.g., yellow) while the positively valenced word changed color (e.g., green). All stimuli remained onscreen for a further 3000ms before being removed, followed by an inter-trial interval, and the next trial (see Figure 2).

Exploratory questions. Exploratory questions were broadly similar to those reported in Experiment 1, with two exceptions. First, the *target-source contingency awareness* question was altered to fit the procedure used in Experiment 2: "*In the first part of the experiment, when MORAG/STRUAN appeared on the screen, which of the following words switched to a different color…*" response options (The positive words, The negative words, I don't remember). Second, the *shared feature awareness* question was altered for a similar reason: "*During the first part of the study, did you notice that the color of one of the two words presented on the right side of the screen changed, while the other word stayed the same color as MORAG or STRUAN?*'.

Results

Data Preparation. We excluded data from 12 participants who did not complete the entire experimental session, and a further three who failed to maintain IAT performance criteria. This led to a final sample of 103 participants.

Hypothesis Testing.

IAT. We did not obtain evidence that IAT scores differed as a function of the color that a nonword shared with a valenced word, $t(100.85) = -1.18$, $p = .24$, $d = -0.23$, 95% CI = [-0.62, 0.16], $\text{BF}_{10} = 0.38$. It seems that participants generally favored one nonword (Struan) over the other (Morag). This was true when (a) Morag remained in the same color as positive words (and the color of negative words changed; $M = -0.26$, $SD = 0.54$), or (b) when Struan remained in the same color as positive words (and the color of negative words changed; $M = -0.12$, $SD = 0.60$).

Self-reports. We did not obtain evidence that self-reported ratings differed as a function of the color that a nonword shared with a valenced word, $t(100.98) = -1.09$, $p = .28$, $d = -0.21$, 95% CI = [-0.61, 0.18], $\text{BF}_{10} = 0.35$. Once again, participants generally favored one nonword (Struan) over the other (Morag). This was true when (a) Morag remained in the same color as positive words (and the color of negative words changed; $M = -2.80$, $SD = 5.33$), or (b) when Struan remained in the same color as positive words (and the color of negative words changed; $M = -1.58$, $SD = 6.03$).

Behavioral intentions. Data were prepared and analyzed as in Experiment 1. Although participants intentions towards the two nonwords differed between the two shared features conditions, they did so in the opposite direction as predicted. That is, participants were less likely to choose the nonword that shared a feature with a positively valenced word (versus the one that shared a feature with a negatively valenced word), OR = 0.22, 95% CI = [0.07, 0.66], p = .007.

Discussion

The findings of Experiment 2 differ from those obtained in Experiment 1. During the acquisition phase a neutral nonword and two valenced words were first presented in the same color. One of the valenced words then changed to a different color while the other remained in the same color as the neutral nonword. We did not obtain evidence for the idea that self-reported and automatic evaluations emerged when such a procedure was used.

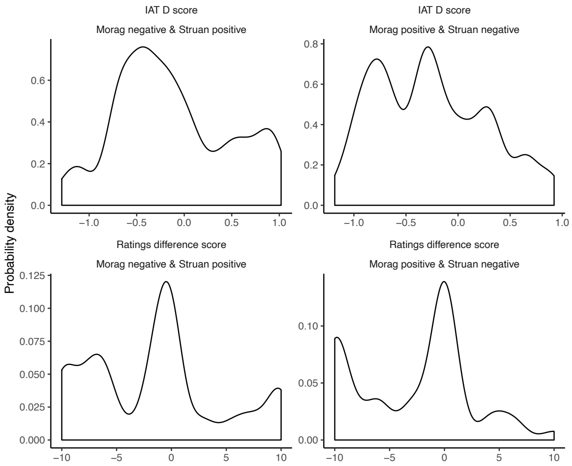

Figure 3. Kernel density plots illustrating distribution of IAT scores (upper panels) and self-reported ratings (lower panels) in Experiment 2.

Experiment 3

After inspecting the distribution of scores on the evaluative measures we came to appreciate that there was a high degree of variability in evaluative responding (see Figure 3). This suggests there may have been different groups of participants in our sample: those that did not show any evaluations towards either nonword (for potentially different reasons), those that showed evaluations in line with the shared features principle (i.e., nonword acquires the same valence as the valenced word it shares a color with) and a third group that showed evaluations in line with a salience hypothesis (e.g., nonword acquires the same valence as the valenced word which changes color). It appears that this latter group exerted more of an impact on the (overall) group level responses reported in Experiment 2 than the other groups.

In retrospect, we believe there may have been a relatively simple explanation for the difference in results of Experiments 1 and 2: the change in task instructions from Experiment 1 to 2. In Experiment 2 participants were told that “you will see two new words: Morag and Struan. These words will appear onscreen together with two other words. The new word (Morag or Struan) and other words will initially appear in one color. Then the color of one of the words will change...Please pay close attention to the color of each word and how they change”. These instructions may have encouraged people to focus greater attention on the *change*, rather than the overlap, in color, and thus treat changes in color as more diagnostic about nonword valence than the shared feature. If so, then modifying task instructions in a way that directs attention to the shared feature may lead to similar effects as seen in Experiment 1. With this in mind, we replicated Experiment 2 while modifying the instructions to emphasize that the nonwords and valenced words remained in the same color.

Method

Participants and design. 118 participants (70 females, M_{age} = 28.19, SD = 6.08) took part in the study via the Prolific Academic website.

Procedure

An identical procedure was used as in Experiment 2 with the exception of the instructions provided prior to the acquisition phase.

Acquisition phase. Prior to training participants were told the following: “You are going to see a new word appear on the screen (i.e., Morag or Struan). Morag or Struan will appear on the left of the screen. Two other words will appear on the right. Morag or Struan and other words will first appear in the same color. Morag or Struan will stay the same color as one of the words on the right. Please pay close attention to the colors of the words. You will be asked some questions about this later on.”

Exploratory questions. Although the exploratory questions were similar to those used in Experiment 2 we did make several changes. First, the phrasing of the *source valence awareness* question was revised to make it easier for the participant: “*In the first part of the experiment we presented MORAG/STRUAN along with two other words. Did…?*” and the response options were (Those two other words always have a positive meaning, Those two other words always have a negative meaning, One word always had a positive meaning and the other had a negative meaning, I don’t remember). The same went for the *shared feature awareness* question: “*During the first part of the study, did you notice that the color of one of the two words presented on the right side of the screen changed, while the other word stayed the same color as MORAG or STRUAN?*”, and the *influence awareness* question “*Did the fact that the words stay the same color or change color influence how much you liked or disliked MORAG or STRUAN?*”.

Results

Data Preparation. Nine participants did not complete the entire session whereas an additional twelve did not meet the IAT criteria. This led to a final sample of 97 participants.

Hypothesis Testing.

IAT. IAT scores differed as a function of the valence of the valenced word that had the same color as the neutral nonword, $t(93.42) = 3.29$, $p = .001$, $d = 0.66$, 95% CI = [0.25, 1.08], $BF_{10} = 20.10$. When Morag remained in the same color as the positive word (and the color of the negative word changed) and when Struan remained in the same color as the negative word (and the color of the positive word changed), participants demonstrated a relative preference for Morag over Struan ($M = 0.21$, $SD = 0.46$). When the color contingencies were reversed, participants demonstrated a relative preference for Struan over Morag ($M = -0.15$, $SD = 0.59$).

Self-reported ratings. Self-reported ratings also varied as a function of the color relation between target and source objects, $t(92.94) = 5.52$, $p < .001$, $d = 1.13$, 95% CI = [0.69, 1.56], $BF_{10} > 10^4$. When Morag remained in the same color as the positive word (and the color of the negative word changed), and when Struan remained in the same color as the negative word (and the color of positive word changed), participants showed a relative preference for Morag over Struan ($M = 2.92$, $SD = 5.25$). When the color contingencies were reversed participants demonstrated a relative preference for Struan over Morag ($M = -2.85$, $SD = 5.02$).

Behavioral intentions. Intentions towards the nonwords differed between the two shared features conditions, and in a way that was congruent with prior training. Specifically, the odds that a participant would choose the nonword that shared a feature with a positive word (over the one that shared a feature with a negative word) were OR = 6.94, 95% CI = [2.03, 23.77], $p = .002$.

Discussion

When task instructions directed attention towards (rather than away from) the shared feature, a shared features effect emerged. Specifically, neutral nonwords that shared a color with positive words were liked more than those which shared a color with negative words. We obtained evidence for the shared features effect on self-report, automatic (IAT), and behavioral intention measures. Importantly, the effect arose even though the nonword and valenced words shared their color from the start of each trial. Unlike the effect that was observed in Experiment 1, the effect in Experiment 3 can therefore not be explained in terms of mere salience.

Experiment 4

Until now we have seen how one particular shared feature (color) comes to moderate automatic and self-reported evaluations. Yet our account suggests that other shared features should function in a similar way. Indeed, a common size, direction, location, smell, or taste shared by two stimuli should lead people to act as if those stimuli share other features as well (e.g., valence). Therefore, in order to extend and generalize our findings, we swapped one shared feature (color) for another (size), to demonstrate that this second feature can also moderate likes and dislikes whenever two stimuli share it. In Experiment 4 participants once again encountered an acquisition phase in which three stimuli (neutral target, positive source, negative source) were presented onscreen. This time one neutral nonword and a positive word were presented in the same sized font whereas the negative word was presented in a differently sized font. Likewise, a second nonword and a negative word shared a common sized font whereas a positive word was always presented in a different sized font. If we are correct, then the nonwords should acquire the same valence as the valenced words with which they share a common size.

Method

Participants and design. 212 participants (103 females, $M_{age} = 30.33$, $SD = 6.18$) took part in the study via the Prolific Academic website. Given the change in shared feature from color to size (which we thought might be subtler) we decided to double our planned sample size relative to Experiments 1-3. The same was true for Experiment 5. A sample size of 200 participants provides sufficient power to detect effect sizes of $d >= 0.4$ (power = 0.80 at alpha = 0.05, two-tailed) or $d = .47$ (power = 0.95 at alpha = 0.05, two-tailed).

Procedure

A similar procedure was used as in Experiments 1-2 with the exception of the acquisition phase.

Acquisition phase. Prior to the acquisition phase participants were told the following:

"In the next part of the experiment you are going to see a novel word appear on the screen (i.e., MORAG or STRUAN). MORAG or STRUAN will appear on the left while two other words will appear on the right side of the screen. MORAG or STRUAN and other words will all appear in certain sizes. Please pay close attention to the sizes of the words. You will be asked some questions about this later on".

Training then consisted of three blocks of 16 trials (48 total) consisting of two different types of trials. During one type of trial the first nonword (e.g., Morag) was presented in the same sized font (e.g., 8% of screen height) as the positive word and a different sized font as the negative word (e.g., 4% of screen height). In another type of trial, the second nonword (e.g., Struan) was presented in the same sized font the negative word and a different sized font the positive word. Stimuli were always presented in the same color (white) and the sizes of the fonts was randomly counterbalanced across trials (e.g., sometimes a target and source share a small [4%] font and at other times they shared a large [8%] font; see Figure 4).

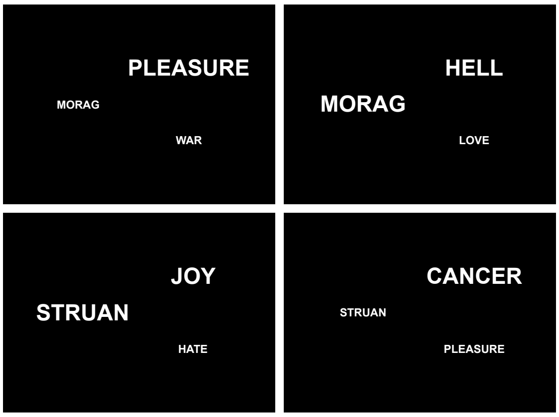

Figure 4. Illustration of the four types of trials during the acquisition phase of Experiment 4. Neutral targets were presented in the same or different size (sometimes in large and other times in small font) as positive or negative targets.

Exploratory questions. A similar set of exploratory questions were used as in Experiment 3, with the following exceptions. First, the *target-source contingency awareness* question was revised to fit the procedure:

"*Think back to the first part of the experiment (where the three words were presented together onscreen). Was MORAG always presented in*" and the response options were (The same size letters as POSITIVE Words, The same size letters as NEGATIVE words, I don't remember). A similar rationale was used to change the *shared feature awareness* question: "*Think back to the first part of the experiment. During that part of the study, we presented MORAG and Positive Words in the same sized letters and STRUAN and Negative Words in the same sized letters. Did you notice this during the study?*', and *influence awareness* questions: "*Do you think that the fact that MORAG and Positive words were presented in the same sized letters (and that STRUAN and Negative Words were presented in the same sized letters) influenced how you rated or otherwise responded towards MORAG or STRUAN?*'.

Results

Data Preparation. Fifteen participants did not complete the entire session whereas an additional nine did not meet the IAT criteria. This led to a final sample of 188 participants.

Hypothesis Testing.

IAT. IAT scores differed as depending on the feature shared by neutral nonwords and valenced words, $t(184.00) = 5.02$, $p < .001$, $d = 0.73$, 95% CI = [0.44, 1.03], $\text{BF}_{10} > 10^4$. When Morag was presented in the same size font as a positive word and Struan was presented in the same size font as a negative word, participants showed a relative automatic preference for Morag over Struan ($M = 0.16$, $SD = 0.48$). When the size contingencies were reversed, participants demonstrated a relative preference for Struan over Morag ($M = -0.18$, $SD = 0.46$).

Self-reported ratings. Self-reported ratings differed as a function of whether neutral nonwords and valenced words shared a feature, $t(179.27) = 8.51$, $p < .001$, $d = 1.25$, 95% CI = [0.93, 1.59], $\text{BF}_{10} > 10^6$. When Morag was presented in the same sized font as a positive word, and Struan was presented in the same sized font as a negative word, participants showed a relative preference for Morag over Struan ($M = 3.57$, $SD = 4.99$). When the size contingencies were reversed, participants showed a relative preference for Struan over Morag ($M = -2.18$, $SD = 4.21$).

Behavioral intentions. Intentions towards the nonwords differed between the two shared features conditions, in a way that was congruent with prior training. Specifically, the odds that a participant would choose the nonword that shared a feature (size) with a positive word (over the one that shared a feature with a negative word) were OR = 7.63, 95% CI = [3.11, 18.76], $p < .0001$.

Discussion

Results indicate that size can also function as a shared feature that moderates automatic and self-reported evaluations as well as behavioral intentions. During the acquisition phase a neutral nonword was presented with two valenced words – one positive and another negative. When a nonword was presented in the same size as positive word it was liked more than a nonword that was presented in the same size as a negative word. These findings replicate those obtained in Experiments 1 and 3 and demonstrate that different types of shared features lead to the transformation of evaluations and intentions.

Experiment 5

In Experiments 1-4, we exclusively focused on how physical features shared by stimuli (e.g., color or size) influence behavioral intentions, automatic and self-reported evaluations. Yet, as we highlighted in the introduction, there are many instances where the features that objects share are conceptual in nature. For instance, minimal group effects can emerge when people are said to share a conceptual relation with one another (e.g., are said to be 'overestimators' or 'underestimators' based on their prior behavior; e.g., Tajfel et al., 1971). Moral spill-over effects can occur when people are said to share a conceptual relation (e.g., they are family members; Uhlmann et al., 2012). Thus, the shared features principle accommodates feature transformation on the basis of physical and conceptual shared features.

In Experiment 5 we set out to experimentally model conceptual shared feature effects. Specifically, we first trained two conceptual categories that each consisted of two colors (e.g., *Blue-Same-Yellow* and *Green-Same-Purple*) followed by a similar acquisition phase to that used in Experiments 1-3. However, this time, we presented a neutral nonword in either blue or green along with a positive and a negative word that were presented in either yellow or purple. If a nonword is presented in blue and a positive word is presented in yellow (along with a negative word in purple) then participants should evaluate the nonword positively (given that blue and yellow were trained to be conceptually similar to one another in the first phase of the experiment). In contrast, if participants encounter a nonword in green along with a negative word in purple (and a positive word in yellow) then they should evaluate that nonword negatively (given that green and purple were trained to be similar to one another during the acquisition phase). Such a finding would further replicate our existing findings and expand the remit of the shared features principle by demonstrating that the shared feature moderating attitude formation can be conceptual rather than purely physical in nature.

Method

Participants and design. 214 participants (108 females, M_{age} = 30.65, *SD* = 6.08) took part in the study via the Prolific Academic website.

Procedure. The study consisted of four phases: color training, acquisition, evaluative measures, and exploratory questions.

Color training. Color training consisted of three blocks of 16 training trials followed by one block of 16 test trials. A Matching to Sample (MTS) task was used to establish relations between two sets of colors (e.g., *Yellow-Blue* and *Green-Purple*). On each trial, one color was presented at the top of the screen, and two at the bottom. Participants had to select the color at the bottom that went with the color at the top and were told that corrective feedback provided by the computer would help them do so. When a correct response was emitted then all stimuli were removed from the screen, a feedback message ('Correct') presented, followed by a 500ms ITI. If an error was made, stimuli were once again removed, corrective feedback provided ('Wrong'), an ITI followed by the next trial. Test trials were identical to training trials with the exception that corrective feedback was no longer provided (see Figure 5). Prior research on stimulus equivalence learning shows that such a MTS training procedure results in people responding as if the related stimuli are equivalent (see Hughes & Barnes-Holmes, 2016).

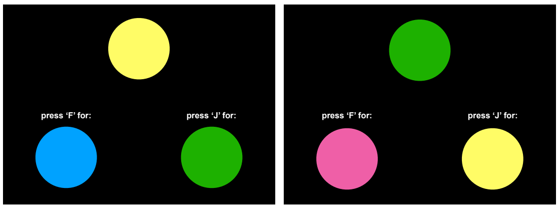

Figure 5. Illustration of the two types of trials during the color training phase of Experiment 5. Selecting blue in the presence of yellow, or purple in the presence of green (and vice-versa) was reinforced and any other color relation punished.

Acquisition phase. Prior to the acquisition phase participants were told the following: "In the next part of the experiment you are going to see a novel word appear on the screen (i.e., MORAG or STRUAN). MORAG or STRUAN will appear on the left while two other words will appear on the right side of the screen. MORAG or STRUAN and other words will all appear in certain colors. Please pay close attention to the color of the words. You will be asked some questions about this later on".

Training consisted of three blocks of 16 trials (48 total), with each block containing two types of trials: one type of trial where the first nonword (e.g., Morag) was presented in one color (e.g., blue), a positive word was presented in a second color (e.g., yellow), and a negative word was presented in a third color (e.g., purple). In another type of trial the second nonword (e.g., Struan) was presented in a fourth color (e.g., green), and the valenced words were presented in the aforementioned colors. Stimulus assignment to the various colors was counterbalanced across participants. All three stimuli were presented against a black background for 5000ms. Thereafter, all stimuli were removed, followed by an inter-trial interval of 750ms, and the next trial (see Figure 6).

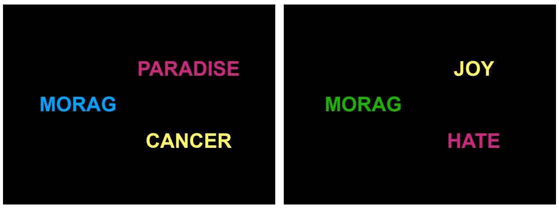

Figure 6. Illustration of the two types of trials during the acquisition phase of Experiment 5. During one type of trial (*left*) a neutral target and positive sources were presented in colors that had previously been related. During a second type of trial (*right*) a second target and negative sources were presented in colors that had also been related during the color training phase.

Evaluative measures. Evaluative measures were similar to Experiments 1-4.

Exploratory questions. Participants were asked a similar set of exploratory questions as in Experiment 4 with several exceptions. We also now probed for *color contingency awareness* (i.e., what the relationship was between the various colors): "*Think back to the first part of the experiment where you learned about the relationship between colors. What color was BLUE/YELLOW/GREEN/PURPLE related to*" and given the following options (Green, Yellow, Purple, I don't remember) (note: the name of the color in the question was never offered as a response option). We also assessed *target-source contingency awareness* (i.e., if they could recall what color the TOs and SOs were presented in): "*In the second part of the experiment, when MORAG/STRUAN appeared on the screen with two other words, what color was MORAG/STRUAN/Positive words/Negative words presented in*" and given the following options (Green, Blue, Yellow, Purple, I don't remember). Finally, we revised the *shared feature awareness* question: "*Think back to the first part of the experiment. During that part of the study, we trained you to relate Blue to Yellow and Green to Purple. In the second part (see below), we presented one of the words on the left in blue (or green) and the words on the right in yellow or purple. Did you notice this during the study?*', and *influence awareness* question: "*Do you think that the color that the words were presented in influenced how you rated or otherwise responded towards MORAG or STRUAN?*').

Results

Data Preparation. Fifteen participants failed to provide complete data. A further twenty failed to meet the IAT criteria. This led to a final sample of 179 participants.

Hypothesis Testing.

IAT. IAT scores differed depending on the valence of the word that shared a color connection with a neutral nonword, $t(168.75) = 3.79$, $p < .001$, $d = 0.57$, 95% CI = [0.27, 0.87], $\text{BF}_{10} = 109$. When Morag was presented in a color that was equivalent to the color a positive word was presented in, and Struan was presented in a color that was equivalent to the color a negative word was presented in, participants preferred Morag over Struan ($M = 0.14$, $SD = 0.46$). When the color contingencies were reversed, participants preferred Struan over Morag ($M = -0.12$, $SD = 0.46$).

Self-reported ratings. Self-reported scores differed depending on the valence of the word that shared a color connection with a neutral nonword, $t(169.77) = 7.66$, $p < .001$, $d = 1.15$, 95% CI = [0.83, 1.47], $\text{BF}_{10} > 10^6$. When Morag was presented in a color that was equivalent to the color that positive words were presented in, and Struan was presented in a color that was equivalent to the color that negative words were presented in, participants preferred Morag over Struan ($M = 2.34$, $SD = 4.12$). When the color contingencies were reversed, participants preferred Struan over Morag ($M = -2.38$, $SD = 4.09$).

Behavioral intentions. Intentions towards the nonwords differed between the two shared features conditions, in a manner that was congruent with prior training. Specifically, the odds that a participant would choose the target object that shared a feature with a positive word (over the one that shared a feature with a negative word) were OR = 5.00, 95% CI = [1.91, 13.06], $p = .001$.

Discussion

Experiment 5 extends our account further and shows that conceptual shared features give rise to automatic and self-reported evaluations in a similar way to physical shared features. Prior to the acquisition phase, two relations between colors were trained (i.e., *Blue-Similar-Yellow*, and *Green-Similar-Purple*). Thereafter a neutral nonword was simultaneously presented with two valenced words. Critically, one nonword was presented in blue whereas the other was presented in green. The positive words were presented in yellow and negative words in purple. Self-reported ratings, IAT effects, and behavioral intention measures all indicated that the nonword presented in blue was preferred relative to the nonword presented in green, supporting the idea that a shared conceptual feature can led to a transfer of other properties (i.e., valence).

Experiment 6

Thus far our findings have been framed as shared feature effects (i.e., as changes in evaluation that occur when stimuli a share feature with one another). Yet we thought it useful to entertain alternative theoretical ideas about our findings. Take, for instance, the possibility that the findings reported in Experiments 1-5 are instances of a *non*-shared features effect. According to this alternative account, stimuli which do not share a feature are evaluated in opposite directions. Hence, a neutral nonword does not acquire its valence from the valenced word it shares a feature with ('target-*same-as*-positive source') but from the valenced word that it does not share a feature with ('target-*opposite-to*-negative source').

Experiment 6 sought to replicate and extend our findings while controlling for a non-shared feature explanation. In our original acquisition phase three stimuli were always presented: a neutral nonword along with a positive and a negative word. We modified this phase so that eight stimuli were now

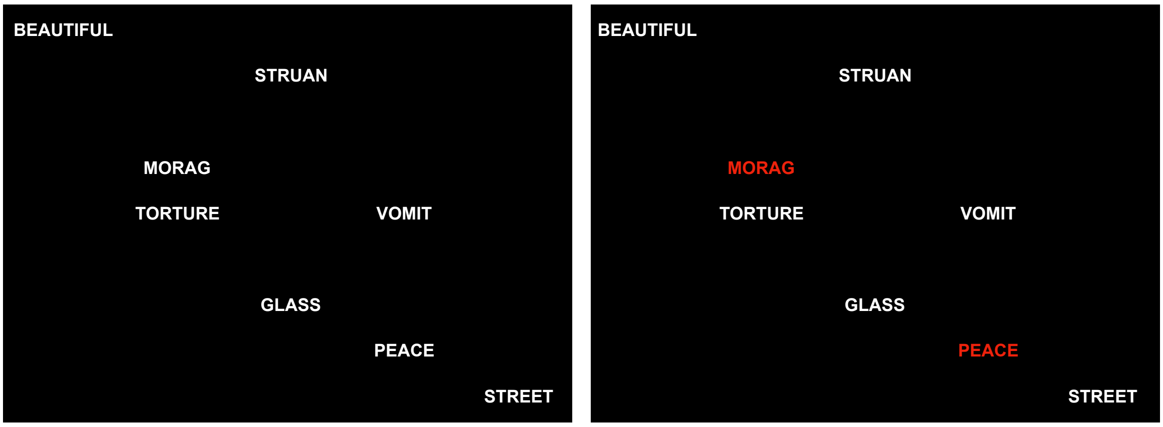

Figure 7. Illustration of the trial progression during the acquisition phase of Experiment 6. During the first half of the trial (*left*) all eight stimuli (two positive sources, two negative sources, two neutral sources, and two neutral targets) were presented in the same color. During the second half of the trial (*right*) one of the targets and a positive source changed to the same color.

presented: six stimuli (two positive, two negative, two neutral) along with two nonwords. During the first half of each trial all stimuli appeared in white. During the second half of the trial one neutral nonword and one valenced word were presented in the same color while all other stimuli remained in white. From a non-shared feature perspective, participants should not produce a strong evaluative response to that nonword because it does not share a feature with six stimuli of differing valence (i.e., the positive, negative, and neutral stimuli that, unlike the Morag, remain white). In contrast, the shared features principle would once again predict evaluations in-line with the valence of the single valenced stimulus whose color switches to the same color as the nonword.

Method

Participants and design. 262 participants (119 female, M_{age} = 27.30, *SD* = 7.36) took part in the study via the Prolific Academic website.

Procedure. The procedure was similar to Experiment 3 with several exceptions (*see below*).

Acquisition phase. Prior to the acquisition phase participants were told the following:

"In the next part of the experiment you are going to see a new word on the screen (i.e., MORAG or STRUAN). This new word (MORAG or STRUAN) will appear together with other words. These words will appear in certain colors. Please pay close attention to the color of the words. You will be asked some questions about this later on".

Training then consisted of three blocks of 16 trials (48 total), with each block containing two types of trials: one type of trial where the first nonword (e.g., Morag) and one of the positive words eventually shared a color, and another where the second nonword (e.g., Struan) and one of the negative words eventually shared a color. Specifically, each trial contained eight stimuli: two positive words (either *Love, Happy, Beautiful, Peace, Friendship,* and/or *Success*), two negative words (e.g., *Agony, Murder, Vomit, Disease, Cancer,* and/or *Torture*), two neutral words (*Table, Building, Glass, Street, Number,* and/or *Bowl*), as well as two neutral nonwords (*Morag* and *Struan*). Stimuli were selected from a large valenced word norm study (Moors et al., 2013). All stimuli initially appeared in the same color (white). After 3000ms the color of one nonword and one valenced word changed (e.g., blue) while the other six stimuli maintained the same color as before (white). Stimuli remained onscreen for another 3000ms before all stimuli are removed, followed by an inter-trial interval of 1000ms, and the next trial. Four stimulus colors were used (lime, fuchsia, yellow, and deep-sky-blue) and stimulus color was varied across trials so that no color could acquire a specific valence (see Figure 7).

Evaluative measures. Evaluative measures were similar to Experiments 1-5.

Exploratory questions. We modified the phrasing of certain items for clarity purposes. For instance, the *source valence awareness* question now asked: "*In the beginning of the experiment we initially presented MORAG/STRUAN together with several words. MORAG/STRUAN and these words all appeared in WHITE. Did the OTHER words...?*" response options (always have a positive meaning, always have a negative meaning, have different meanings [e.g., some were positive, some were negative, some were neutral], I don't remember). The *target source contingency awareness* question asked: "*During the first part of the experiment MORAG/STRUAN and another word changed color. Did MORAG/STRUAN...:*" response options (and POSITIVE WORDS subsequently share a color, and NEGATIVE WORDS subsequently share a color, and NEUTRAL WORDS subsequently share a color, I don't remember). The *demand question* for their self-reported ratings asked: "*Earlier you rated MORAG and STRUAN as being either positive, neutral, or negative. Did you base your response on how you actually felt about those words OR on what you thought the researchers wanted you to say?*" response options (How I actually felt about MORAG and STRUAN, What I thought the researchers wanted me to say [i.e., not on how I personally felt], I don't know), while the same question for the IAT asked: "*Earlier you completed the Implicit Association Test. Did you base your performance in that task on your best efforts to perform the categorizations as quickly and accurately as possible? Or did you attempt to influence your speed or accuracy in order to go along with what you thought the researchers wanted you to feel about the words*? Response options (I tried to perform the task as quickly and accurately as possible, I tried to alter my performance based on what I thought the researchers wanted to find, I don't know). The *shared feature awareness* question now asked: "*Think back to the first part of the experiment. We showed you MORAG and POSITIVE WORDS in the same color. We also showed you STRUAN and NEGATIVE WORDS in the same color. Did you notice this during the first part of the study?*', whereas the *influence awareness* question stated: "*Do you think that the fact that MORAG and POSITIVE WORDS were presented in the same color (and that STRUAN and NEGATIVE WORDS were presented in the same color) influenced how much you like or dislike MORAG or STRUAN?*'.

Results

Data Preparation. Twenty-three participants failed to provide complete data. A further eight failed to meet the IAT criteria. This led to a final sample of 231 participants.

Hypothesis Testing.

IAT. IAT scores differed depending on the valence of the valenced word that shared a color with a neutral nonword, $t(227.66) = 7.25$, $p < .001$, $d = 0.96$, 95% CI = [0.68, 1.23], $\text{BF}_{10} > 10^6$. When Morag was presented in the same color as positive words, and Struan was presented in the same color as negative words, participants preferred the former over the latter ($M = 0.31$, $SD = 0.51$). When the color contingencies were reversed, participants preferred Struan over Morag ($M = -0.18$, $SD = 0.52$).

Self-reported ratings. Self-reported scores differed depending on the valence of the valenced word that shared a color with a nonword, $t(227.44) = 12.08$, $p < .001$, $d = 1.58$, 95% CI = [1.29, 1.88], $\text{BF}_{10} > 10^6$. When Morag was presented in the same color as positive words, and Struan was presented in the same color as negative words, participants preferred Morag over Struan ($M = 3.61$, $SD = 4.72$). When the color contingencies were reversed, participants preferred Struan over Morag ($M = -3.40$, $SD = 4.09$).

Behavioral intentions. Intentions towards the two nonwords differed between the two shared features conditions, in a manner that was congruent with prior training. Specifically, the odds that a participant would choose the nonword that shared a feature with a positive word (over the one that shared a feature with a negative word) were OR = 14.85, 95% CI = [6.98, 31.63], $p < .0001$.

Discussion

Experiment 6 extends our shared feature account still further. It reveals that changes in evaluation occur due to the features that stimuli share with one another rather than the features they do not share. During the acquisition phase participants encountered eight words: two positive words, two negative words, two neutral words, and two neutral nonwords. During the first half of each trial all stimuli appeared in white. Later on, one nonword and one valenced word were presented in the same color while the other stimuli remained in white. According to a *non*-shared feature account, a nonword does not acquire its valence from the word it shares a feature with (e.g., 'target-same-as-positive source') but the word that it does not share a feature with (e.g., 'target-opposite-to-negative source'). Such an account cannot explain the effects reported here given that there were now six stimuli of varying valence that never shared the color-feature with the nonword during the second half of the trial, which should have led to ambivalent responses towards the nonword. In contrast, self-reported ratings, IAT effects, and behavioral intention measures all indicated that the nonword which shared a feature with positive words was evaluated more positively than the nonword that shared a feature with negative words.

Experiment 7

Experiments 1-6 demonstrate that when stimuli share features with one another changes in self-reported and automatic evaluations can occur. However, one might ask to what extent such effects are driven by demand characteristics (i.e., by participants recognizing the contingencies operating within the task and responding, not based on their actual evaluations of the stimuli, but rather on what they believe the researcher wants them to say). When designing the studies we were sensitive to such a possibility and decided to include an indirect procedure (the IAT) which is arguably less susceptible to demand than self-report measures. But one might still question the extent to which even IAT effects reflect *genuine* changes in automatic evaluation. Given these respective concerns we decided to once again attempt to replicate the shared features effect, but this time using the EPT as an indirect procedure. Although, like most (implicit) measures in psychology, evaluative priming effects are not entirely uncontrollable (e.g., Teige-Mocigemba & Klauer, 2013), they are generally considered to provide a valid index of automatic evaluation (e.g., De Houwer et al., 2009). Moreover, because of the structural differences between IAT and EPT procedures (De Houwer, 2003), a replication of our findings using the EPT would indicate that the observed effects on the IAT are not due to procedural elements that are unique to the IAT.

In Experiment 7 we focused on size as a shared feature given that we already replicated color-based effects in five of our six studies (and size only once). As such, replicating our prior findings would also strengthen the conclusion that our effects are not limited to color as a shared feature.

Method

Participants and design. 539 participants took part in the study via the Prolific Academic website.[5] After excluding participants with partial data, the final analytic sample consisted of 492 participants (242 women, 246 men, 5 did not provide gender data; $M_{age} = 28.28$, $SD = 7.43$). Note, that in this experiment, we opted to use a Sequential Bayes Factor (SBF) design with a maximal N (e.g., Schönbrodt & Wagenmakers, 2018). Specifically, we use a threshold of > 3 or < 0.33, a default Cauchy prior (r = .707), a minimum of 300 participants, an addition of 100 participants for each test, and a maximum of 500 participants (we ran an additional 39 participants to allow for 500 participants after exclusions).

Stimuli. The two nonwords MORAG and STRUAN were used as prime stimuli. Targets consisted of ten positive words (e.g., *fantastic, great, nice, good, pleasant, attractive, delight, smile, trust, positive*) and ten negative words (e.g., *terrible, disgusting, nasty, horrible, sick, abuse, failure, grief, negative, pain*).

Procedure

The procedure was similar to Experiment 4 with one notable exception: an EPT was used in the place of an IAT.

Evaluative priming task. Automatic evaluations were assessed using an EPT in which participants are asked to categorize target words as either positive or negative using the E and I keys of a computer keyboard. During all trials, the labels "good" and "bad" appeared on the lower left and right corners of the screen. In line with the procedures of earlier studies (e.g., Van Dessel, Gawronski, Smith, & De Houwer, 2017), a single trial consisted of a fixation cross presented for 500 ms, a blank screen for 500 ms, a prime for 200 ms, a blank screen for 50 ms, and the presentation of a target word. The inter-trial interval was set to vary randomly between 500 ms and 1500 ms. Whenever an incorrect response was made or participants did not respond prior to the response deadline of 1500 ms, 'Wrong' was displayed in the center of the screen for 250 ms before the next trial. Participants were asked to respond as quickly as possible without making too many errors. With two types of primes and two types of targets, there were four prime-target combinations. Participants first completed eight practice trials, which were then followed by eighty critical test trials. The test trials were presented in a single block, with each of the four types of prime-target combinations presented twenty times in random order.

Exploratory questions. Similar questions were provided as in Experiment 4.

Results

Data Preparation.

Evaluative priming. Latencies from incorrect responses in the EPT (4.26% of the final sample) were eliminated and outlier latencies longer than 1000 ms and shorter than 300 ms (5.92% of the correct responses in the final sample) were removed. From the remaining trials we calculated two mean reaction times for each participant – one for congruent and one for incongruent trials (i.e., trials where the valence of the target stimulus during the priming task was the same/opposite to the assumed valence of the prime stimulus presented on screen [MORAG or STRUAN]). This comparison of mean reaction times on congruent versus incongruent trials represents the standard scoring and analytic procedure for Evaluative Priming tasks.

Hypothesis Testing.

Evaluative priming. Following our pre-registered data collection and analysis plan, we analyzed the data after collecting 300 participants, but no conclusions could be made in favor of either the null of alternative hypotheses, $BF_{10} = 1.42$, $\mu = 4.21$ ms, 95% CI [0.95, 7.49], $\delta = 0.14$, 95% CI [0.03, 0.26]. Adding 100 additional participants did not change this, $BF_{10} = 0.76$, $\mu = 3.39$ ms, 95% CI [0.51, 6.29], $\delta = 0.11$, 95% CI [0.02, 0.21]. When sample size was increased to 500 moderate evidence in favor of the

[5] We should also note that our preregistration stated thresholds of > 3 and < 0.16. This was due to an oversight on our behalf – such asymmetries between cutoffs for H0 and H1 are uncommon in practice. Common convention (e.g., Jeffreys, 1961; Lee and Wagenmakers, 2014), indicates that the strength of evidence for one hypothesis compared to its competing hypothesis is regarded as noteworthy if BFs are above 3 or below 0.33. We therefore adopted these cutoff criteria. Notably, changing criterion for H0 (< 0.16 or < 0.33) did not influence the results or conclusions reported below.

alternative hypothesis was observed, $BF_{10} = 3.11$, $\mu = 3.88$ ms, 95% CI [1.28, 6.58], $\delta = 0.13$, 95% CI [0.04, 0.22]. For the sake of reader familiarity we also ran a (non-preregistered) frequentist between subjects Welch's *t*-test that also suggested evidence in favor of the expected differences between congruent and incongruent trials, $t(492) = 2.89$, $p = .004$, $d = 0.05$, 95% CI = [0.02, -0.08]. Participants' therefore demonstrated an Evaluative Priming effect in the expected direction.

Self-reported ratings. Whereas the sequential testing was used to determine sample size on the basis of the Evaluative Priming data only, data from the self-reported ratings and behavioral intentions were analyzed for the sample size only, as in our previous studies. Self-report ratings differed depending on the valence of the valenced word that shared a size with a neutral nonword, $t(479.98) = 12.72$, $p < .001$, $d = 1.15$, 95% CI = [0.96, 1.35], $BF_{10} > 10^6$. When Morag was presented in the same size as positive words, and Struan was presented in the same size as negative words, participants preferred Morag over Struan ($M = 2.17$, $SD = 4.35$). When the color contingencies were reversed, participants preferred Struan over Morag ($M = -2.66$, $SD = 4.02$).

Behavioral intentions. Intentions towards the two nonwords differed between the two shared features conditions, in a manner that was congruent with prior training. Specifically, the odds that a participant would choose the nonword that shared a feature with a positive word (over the one that shared a feature with a negative word) were OR = 4.33, 95% CI = [2.77, 6.78], $p < .0001$.

Discussion

Once again self-reported and automatic evaluations as well as behavioral intentions emerged when stimuli shared a feature with one another. We not only replicated our finding that size can function as a shared feature but also generalized them from one indirect procedure (IAT) to another (EPT). Notably, the evaluative priming effect was small, and smaller than the effect obtained on the self-reported ratings, or on the other indirect procedure (IAT) that we used in our earlier studies. That said, our findings with the EP task were in line with what we observed with the IAT, self-reports, and behavioral intentions. This suggests that our findings are not specific to one indirect procedure, and thus strengthens the conclusion that the implementation of shared features can lead to changes in automatic evaluations.

Experiment 8

In the introduction we argued that many phenomena in psychological science may represent instances of the shared features principle. Yet so far we have almost exclusively relied on artificial stimuli (nonwords) to demonstrate the principle itself. Although these stimuli bear similarity to certain real-world items (e.g., the names of novel brands or social groups) a more socially relevant demonstration seems warranted in order to support the larger claim being made here. With this in mind we created an entirely new task that used a more socially relevant set of stimuli (male faces) and yet another feature (common location). This task was designed to function as an experimental analogue of a classic social psychological manipulation: the minimal groups paradigm. This paradigm has often been used to study intergroup processes and set the stage for two highly influential theories in social psychology: Social Identity Theory (Tajfel, Turner, Austin, & Worchel, 1979) and Self Categorization Theory (Turner, Hogg, Oakes, Reicher, & Wetherell, 1987).

The minimal groups paradigm involves a situation where participants are assigned to different groups. This assignment is typically arbitrary (i.e., based on some trivial criteria such as a coin toss or a shared preference for certain paintings), novel (not based on any pre-existing criteria), anonymous (participants lack awareness of who is assigned to their group and never interact with them), and as such, intergroup evaluations or resource allocation serves no direct utilitarian self-interest. Despite this participants still demonstrate a relative preference for the group to which they have been assigned: they often allocate resources in ways that favor their own group, view other group members favorably, and as being more similar to them than outgroup members (for a recent review see Otten, 2016).

From a shared features perspective, certain minimal groups effects emerge when individuals are arbitrarily assigned to the same group based on some shared feature (e.g., similar clothing item or preference for certain paintings), and as a result, people then assume that those same individuals share other features as well (e.g., valence or personality traits). Drawing on this idea we created a task in which individuals did not share a feature with other individuals but rather with valenced events. Specifically, prior to the acquisition phase, participants were informed that the computer would pull images and words from two bags. Participants then saw a single stimulus onscreen (either an unknown male face or a positive or negatively valenced word) along with information highlighting from which bag that stimulus was pulled (e.g., "pulled from Bag 1"). Across trials they could learn that one neutral face (Target 1) was pulled from the same bag as positive words whereas a second face (Target 2) was pulled from the same bag as negative words. If the shared features principle is correct, then the fact that a target (neutral face) and source (valenced image) share one feature (common bag location) may lead participants to infer that they also share other features (valence).

It is worth noting that the acquisition phase was constructed so that stimuli appeared one at a time in random order. As such there was no spatio-temporal co-occurrence of the valenced words and neutral faces. Moreover, the task ensured that any contiguity between stimuli favored a neutral face being related to positive and negative words in equal measure. Thus it is unlikely that changes in liking would be driven by the spatio-temporal paring of stimuli (i.e., they do not represent instances of EC). Nevertheless, we decided to add a second condition in order to demonstrate that it was the shared feature and not mere spatio-temporal contiguity that moderated evaluations. Participants in this condition encountered a similar acquisition phase as outlined above but with one notable exception: information about which stimulus was pulled from which bag was never presented. An absence of evaluative responding in this condition would suggest that evaluations in the other condition are dependent on shared features.

Finally, even stronger support for our argument would emerge if people were to demonstrate evaluations even when they were told that the words and faces were being pulled from the two bags at random. This would imply that there was no logical connection between those stimuli and that there is no such thing as a good or bad bag. We therefore included a third condition which was similar to the first condition with one exception: prior to and during the acquisition phase participants were informed that bag assignment was random and that there was no such thing as a 'good' or 'bad' bag. If effects were to emerge here then it would suggest that shared features can guide evaluations even when people are told that those features are irrelevant and should be disregard.[6]

Method

Participants and design. 245 participants (140 female, M_{age} = 29.17, *SD* = 7.48) took part in the study via the Prolific Academic website. A single factor between-participants design with three levels (*Minimal Groups*: Related vs. Unrelated vs. Random) was used and the same method factors manipulated as in Experiments 1-7. Data-collection was terminated when 240 participants completed the experiment. Recruiting 204 participants provided sufficient power ($>$ 0.90) to observe a medium effect (f = 0.25) at alpha = 0.05. We decided to collect additional participants in order ensure that the necessary sample size was still obtained following exclusions.

Stimuli. Two unknown male images (labelled Chris and James) served as neutral stimuli while ten positive (*fantastic, wonderful, honest, kind, brave, amazing, nice, pleasant, selfless, great*), ten negative adjectives (*horrible, nasty, violent, terrible, hated, disgusting, mean, unpleasant, stupid, bully*) served as valenced stimuli.

Procedure. The procedure was similar to Experiment 3 with several exceptions (*see below*).

Acquisition phase. In all three conditions training consisted of three blocks of 12 trials (36 total) consisting of four different randomly presented trial-types: those that displayed one face (Chris), the second face (James), a positive word, or a negative word by itself in the middle of the screen for 4000ms. The stimulus was then removed, followed by a 2000ms ITI. The three conditions then differed in the following ways.

Shared feature (location) condition. Prior to the acquisition phase participants were told the following: "The computer will now pull a word or image from two bags: Bag 1 & Bag 2. We will tell you what word or image was pulled from Bag 1 or Bag 2. Please pay attention to the words and images pulled from each bag". During each trial a label was presented above the stimulus indicating from which bag it was pulled (i.e., "Pulled from Bag 1", "Pulled from Bag 2", "Pulled

[6] Most work has focused on a particular type of minimal groups *effect* - namely – one in which an individual is randomly assigned to one group and not another group on the basis of a stimulus (coin toss), stimulus property (they are wearing a similar clothing item to other group members), or response (they say they like one painter over another). In most cases, the individual being assigned to the group is the participant, and the outcome of interest is how that individual responds to other individuals (inside or outside their own group). The procedure we adopted also involves an individual being randomly assigned to a group on the basis of a stimulus (random allocation by the computer). As a result of this assignment behavior changes. Yet there are notable differences between the task used here and that used elsewhere: (a) the individual is no longer the participant but an unknown individual, (b) the group contains valenced stimuli rather than other people, and (c) the outcome is not intergroup bias but changes in person perception. One might question to what extent these effects are still minimal group effects. As far as we can see, there is no *a priori* reason why a minimal group effect has to involve a limited set of procedural parameters (e.g., participant as individual; group comprised of other people, focus on intergroup bias). Instead, when viewed through the lens of our shared features account, one sees that there may be many other minimal group *effects* that have not been studied and which could open up interesting new avenues for exploration (i.e., minimal group effects where the individual is not the participant, where the group members are not necessarily other individuals, and the outcome is not only intergroup bias). The task introduced here is one such example.

from Bag 3", "Pulled from Bag 4", "Pulled from Bag 5", "Pulled from Bag 6"). The label assignments were varied as a function of block number and stimulus identity, such that one neutral face (e.g., Chris) and positive words were assigned to Bag 1 in Block 1, Bag 3 in Block 2, and Bag 5 in Block 3 whereas a second neutral face (e.g., James) and negative words were assigned to Bag 2 in Block 1, Bag 4 in Block 2, and Bag 6 in Block 3. Note: assignment of the two faces to the same bags as positive and negative words was counterbalanced across participants.

Random condition. In addition to the instructions provided in the shared feature condition, participants in the random condition were also told the following: "The contents of each bag were randomly created. Thus there is no such thing as a 'good' or 'bad' bag, nor is there a connection between the words and images that are pulled from each bag." Participants were required to complete a manipulation check to ensure that they fully understood and processed these instructions before proceeding to a similar acquisition phase as in the shared features condition.

No contiguity condition. Prior to training participants were informed that the computer would display a word or image and that they should pay attention to these items. During the acquisition phase itself, no information about bag assignment was provided. Instead each stimulus was presented by itself in a non-contingent manner in exactly the same way as in the other conditions but without info about the bags..

Evaluative measures. Evaluative measures were similar to Experiments 1-6.

Exploratory questions. As similar set of exploratory questions were used as in Experiment 7, with several changes. First, the *source valence awareness* question asked: "*In the beginning of the experiment the computer pulled words and images from two different bags. Did the words that were pulled from the bags...?*" response options (always have a positive meaning, always have a negative meaning, have different meanings (e.g., some positive, some negative), I don't remember). Second, the *target source contingency awareness* question asked: "*Think back to the first part of the experiment. Chris/James and another word were pulled from the same bag. Was Chris/James pulled from the same bag as...*", response options (Positive words, Negative words, I don't remember). The *shared feature awareness* question asked: "*Think back to the first part of the experiment. In the first part of the experiment we told you that words and images were pulled from two bags. We pulled Chris and negative words from one bag and James and positive words from another bag. Did you notice that this was happening during the first part of the study?*, whereas the *influence awareness* question stated: "*Do you think that the fact that Chris and negative words were pulled from one bag (and that James and positive words were pulled from another bag) influenced how much you like or dislike Chris or James?*".

Results

Data Preparation. Twenty participants failed to provide complete data. A further three failed to meet the IAT criteria. This led to a final sample of 222 participants.

Hypothesis Testing.

IAT. IAT scores differed as a function of Minimal Group condition, $F(2, 219) = 28.26$, $p < .0001$, $\eta_p^2 = .21$. Planned pairwise comparisons with Bonferroni-Holm corrections for multiple testing indicated no evidence that IAT scores were different between the shared features condition (M = 0.28, 95% CI [0.18; 0.38]) and random conditions (M = 0.16, 95% CI [0.06, 0.26], pairwise $p = .10$), but provided evidence that both the shared features and random conditions differed from the no-contiguity condition (M = -0.20, 95% CI [-0.29; -0.12]; both pairwise $ps < .0001$).

Self-reported ratings. Ratings also differed as a function of Minimal Group condition, $F(2, 219) = 26.70$, $p < .0001$, $\eta_p^2 = .20$. Planned pairwise comparisons with Bonferroni-Holm corrections for multiple testing indicated no evidence that ratings were different between the shared features condition (M = 2.61, 95% CI [1.72, 3.49]) and random condition (M = 2.10, 95% CI [1.21; 3.00], pairwise $p = .39$), but provided evidence that both the shared features and random conditions differed from the no-contiguity condition (M = -1.14, 95% CI [-1.78; -0.50]; both pairwise $ps < .0001$).

Discussion

We once again obtained shared feature effects when a new feature (common location), set of socially relevant stimuli (faces), and procedure were employed. Not only did shared features guide intentions, self-reported and automatic evaluations, but seemed to do so even when people were told that they were irrelevant and should be disregard.[7]

[7] We were somewhat surprised to find an evaluative effect favoring one of the male images over the other in the no-contiguity condition given that neither image was paired with valenced items. We therefore carried out another study to replicate this condition with a different randomization method to determine if comparable evaluations once again emerged. Analyses revealed an evaluative effect on the IAT, $t(88) = -3.90$, $p = .0002$, $d = -0.41$, and self-reported ratings, $t(88) = -5.71$, $p < .0001$, $d = -0.61$ favoring the same male image over the other. Thus it seems the effect in the no-contiguity condition was likely driven by a relative pre-existing preference for one stimulus over another that is not a function

Meta-Analysis

In order to (a) better estimate the evidence for the magnitude of learning via shared features and (b) determine the likelihood of observing shared features effects under other experimental conditions (i.e., to provide information about the *generality* of the effect itself), we carried out meta-analyses of Experiments 1-8. In the case of Experiment 8, we included both conditions that involved a shared feature that one could learn from (i.e., the standard condition and the instructed randomness condition). Total sample size drawn from for the meta analyses was therefore n = 1525. Random effects meta-analysis models were fitted using the metafor R package (Viechtbauer, 2010) and the maximum likelihood estimator function. A separate meta-analysis was fitted for each outcome variable (IAT, self-report ratings, and behavioral intentions; note that not every experiment contained every measure). Our general strategy was to first fit a meta-analytic model and assess for heterogeneity. If heterogeneity was undesirably large then we tested for the presence of outlier experiments using metrics of both excessive influence on the meta analyzed effect size ($\Delta \text{SD}_{\text{effect size}}$) and excessive influence on heterogeneity ($\Delta\tau^2$) via leave-one-out analyses. Studies were only labeled as outliers if results from these tests were consistent across all three outcome variables (i.e., IAT, self-reports, & behavioral intentions). Analyses indicated that Experiment 2 was an outlier on the basis of undue influence on both the meta-analyzed effect size and heterogeneity and across all three outcome variables (full results from each metric available at osf.io/pqm9v). This was also congruent with our previous observation that the different instructions employed in Experiment 2 may have undermined the effect. As such, it was excluded and a second meta-analytic model was refit in each case. The results of both models are reported below. Forest plots of the robustness tests can also be found in Figure 8.

IAT

Fitting a meta-analytic model to the IAT revealed a significant effect of medium size (Cohen, 1988): k = 8, d = 0.75, 95% CI = [0.42, 1.07], 95% CR = [-0.14, 1.63], $p < .0001$. However, results were found to contain a high degree of heterogeneity, Q (df = 7) = 36.81, $p < .001$, $\tau^2 = 0.18$, $I^2 = 83.28\%$, $H^2 = 5.98$. Following the exclusion of Experiment 2 as an outlier, the meta-analyzed effect size was still found to be significant and had moved from medium to large effect size, k = 7, d = 0.86, 95% CI = [0.67, 1.06], 95% CR = [0.47, 1.26], $p < .0001$, and heterogeneity was found to lower, $Q(df = 6) = 11.24$, $p < .001$, $\tau^2 = 0.03$, $I^2 = 47.02\%$, $H^2 = 1.89$.

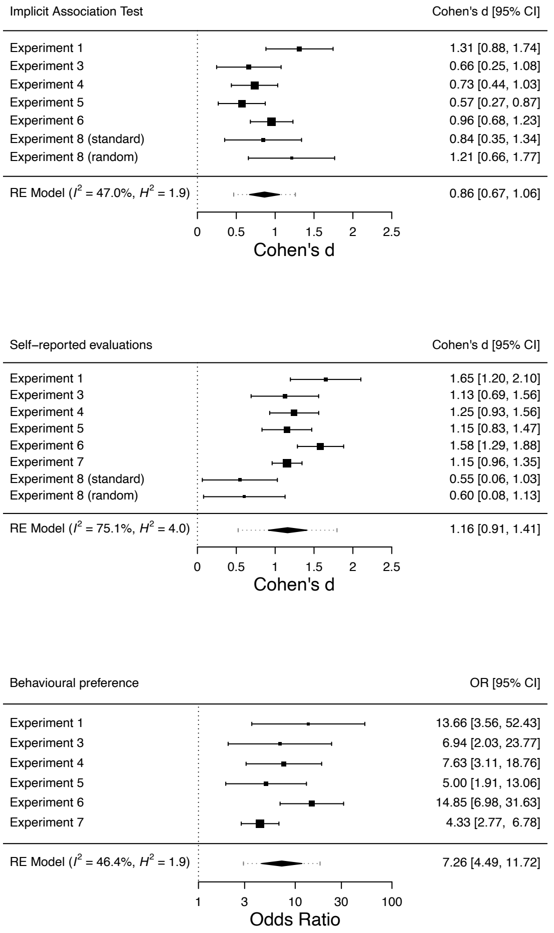

Figure 8. Forest plots for IAT (top), self-report (middle), and behavioral intentions (bottom). Note that Experiment 2 was excluded as an outlier following tests for excessive heterogeneity.

Self-Report Ratings

Fitting a meta-analytic model to the self-report ratings revealed a significant effect of large size, k = 9, d = 0.99, 95% CI = [0.62, 1.37], 95% CR = [-0.14, 2.12], $p < .0001$, but with a high degree of heterogeneity: $Q(df = 8) = 68.4$, $p < .001$, $\tau^2 = 0.29$, $I^2 = 90.65\%$, $H^2 = 10.70$. After excluding Experiment 2, the meta-analyzed effect size was still significant and of large size, k = 8, d = 1.16, 95% CI = [0.91, 1.41], 95% CR = [0.52, 1.80], $p < .0001$. Heterogeneity was also found to be reduced, $Q(df = 7) = 22.72$, $p < .001$, $\tau^2 = 0.09$, $I^2 = 75.08\%$, $H^2 = 4.01$.

of stimulus contiguity or contingency (i.e., this effect is not an instance of EC; for more details see osf.io/pqm9v).

Table 2. Moderation of shared feature effects as a function of source valence awareness, target-source contingency awareness, hypothesis awareness, reactance, and demand compliance.

		Subset included in the robustness tests		Subset excluded from the robustness tests		Difference between subsets	
DV	Moderator	*n*	Effect size	*n*	Effect size	χ^2	*p*
IAT							
	Source valence awareness	484	*d* = 0.83 [0.54, 1.12]	149	*d* = 0.85 [0.15, 1.54]	0.01	.933
	Target source contingency awareness	348	*d* = 1.11 [0.46, 1.77]	218	*d* = 0.33 [-1.54, 2.20]	1.59	.207
	Demand compliance	854	*d* = 0.80 [0.46, 1.13]	75	*d* = 0.76 [-0.16, 1.69]	0.01	.913
	Reactance	839	*d* = 0.85 [0.67, 1.04]	96	*d* = 0.83 [0.13, 1.52]	0.01	.911
Self-reported evaluations							
	Source valence awareness	865	*d* = 1.19 [1.02, 1.37]	256	*d* = 1.29 [0.80, 1.77]	0.36	.546
	Target source contingency awareness	625	*d* = 1.80 [1.21, 2.40]	429	*d* = 0.34 [-1.43, 2.11]	5.98	.014
	Demand compliance	1211	*d* = 1.05 [0.79, 1.30]	212	*d* = 1.80 [1.18, 2.43]	16.34	< .001
	Reactance	1222	*d* = 1.19 [0.94, 1.44]	201	*d* = 0.92 [0.33, 1.51]	2.44	.118
Behavioural intentions							
	Source valence awareness	805	OR = 1.89 [1.47, 2.32]	249	OR = 0.96 [-0.25, 2.18]	5.29	.021
	Target source contingency awareness	625	OR = 2.88 [1.95, 3.82]	429	OR = 0.03 [-2.64, 2.69]	10.47	.001

Notes: Subset employed in the robustness tests = results calculated using the subset of participants that was included in the robustness tests (i.e., those who were source-valence *aware*, target-source contingency *aware*, demand *non-compliant*, and *non-reactant*); Subset excluded from the robustness tests = results calculated using the subset of participants that was excluded for the robustness tests (i.e., those who were source-valence *unaware*, target-source contingency *unaware*, demand *compliant*, and *reactant*); Effect size = meta analytic effect size in that subset with 95% confidence intervals; χ^2 = chi-squared test for moderation (*df* = 1 for all tests).

Behavioral Intentions

Odds Ratios for behavioral intentions were converted to log-odds before being meta analyzed. Results were converted back to Odds Ratios for plotting and reporting. Results revealed a significant meta-analytic effect of medium size (Chen, Cohen, & Chen, 2010): $k = 7$, Odds Ratio = 4.71, 95% CI = [1.70, 13.20], 95% CR = [0.31, 72.97], $p = .003$, and with a high degree of heterogeneity: $Q(df = 6) = 42.25$, $p = .003$, $\tau^2 = 1.67$, $I^2 = 89.78\%$, $H^2 = 9.79$. After excluding Experiment 2, the meta-analyzed effect size was significant and now of large size, $k = 6$, Odds Ratio = 7.24, 95% CI = [4.48, 11.70], 95% CR = [2.92, 17.99], $p < .0001$, and heterogeneity was reduced, $Q(df = 5) = 9.31$, $p < .001$, $\tau^2 = 0.16$, $I^2 = 46.44\%$, $H^2 = 1.87$.

Robustness Tests

In our preregistered analytic plan we stated we would examine the robustness of the shared feature effect when only considering participants who were contingency aware. We also said we would examine if the effect was still present after people who were demand compliant or reactant were removed. These exclusions are common in the EC literature: contingency awareness is often a necessary condition to observe evaluative effects (thus it was required; Hofmann et al., 2010), whereas the other two exclusions ensure that our effects were not contaminated by undesirable sources. Given that each individual study lacked the power to address this question we opted for a meta-analytic approach instead. Each respective subset of participants was excluded and separate meta-analytic models were refit. As in the previous meta analyses, robustness tests excluded Experiment 2 as an outlier. Robustness of conclusions was defined as congruence in the acceptance/rejection of the null hypothesis between results obtained from the full sample and those obtained in a given subset.

Results were found to be robust to only including participants who were (a) aware (i.e., source valence awareness [77.5% of participants] and target source contingency awareness [59.6%]), (b) not demand compliant (80.9%), and (c) not reactant (85.8% on IAT, 89.7% on self-reports). This was found across all three outcome measures (IAT, self-reported evaluations, and behavioral intentions; all sensitivity analysis effect size *p*s < .002). In short, the general trend of evidence suggested that learning via shared features is robust to three common exclusions employed in the literature: requiring participants to be (contingency) aware, not demand compliant, and not reactant.

Moderation Analyses

Based on feedback from a reviewer, we also considered whether the shared feature effect was not merely robust to the aforementioned common exclusion criteria (i.e., source valence awareness, target source contingency awareness, demand compliance, reactivity), but was moderated by them. These analyses were therefore exploratory and not preregistered. Whereas our robustness tests excluded subsets of participants when calculating effect sizes, these multilevel moderator meta-analysis models instead calculated an effect size for both subsets (i.e., those subsets that were previously excluded and those were not) and examined moderation by the effect by this exclusion variable. The non-independence between the subsets from each study was acknowledged via a random intercept for study and a random slope for subset. In the case of some experiments, there were inadequate sample size in one subset to calculate an effect size. In such cases, the study was not included in the moderator meta analyses (i.e., only experiments where a pairs of subsets existed were included). Effect sizes for both subsets in each study and moderation tests are included in Table 2. Given the relatively large number of parameters being estimated in each of these models (i.e., effect sizes for each subset, estimation of moderation, as well as random slopes and intercepts) combined with the low sample sizes in the subsets that were previously excluded for the robustness tests, these analyses likely demonstrated relatively low power to detect effects (see Table 2 for *n*s and confidence intervals). No consistent evidence was found across the outcome variables for any of the moderators.

General Discussion

In this paper we introduced the shared features principle which postulates that when two stimuli share one feature people assume that they share other features as well. Although this principle may underpin many different phenomena in psychological science we sought to illustrate its predictive value in the domain of attitudes, test its potential boundary conditions, and demonstrate that it holds across different physical and conceptual features, outcome measures, and training procedures. Across eight studies we exposed participants to an acquisition phase that typically contained three stimuli: a neutral target, a positive source, and a negative source. We then manipulated each trial so that a target would either share a common color (Experiments 1, 2, 3, 6), size (Experiments 4 and 7) or location (Experiment 8) with one of the source objects. To demonstrate that our account speaks to both physical and conceptual features, Experiment 5 created two conceptual color categories (*Blue-Similar-Yellow*, and *Green-Similar-Purple*) and then, during the acquisition phase, presented a target and source stimulus in a different color from the same category (e.g., target in blue and source in yellow). In all experiments except Experiment 2, we observed that liking of the target object changed in the direction of the valence of the source object with which the target object shared a feature. The unexpected results of Experiment 2 were likely a consequence of instructions that directed attention away from shared features and towards changes in stimulus features (for more *see below*).

Taken together, our results provide strong and repeated support for the shared features principle. Changes in liking were not driven by mere contiguity but instead by the features that targets and sources shared with one another. This was demonstrated most clearly in Experiment 8 where stimuli were never paired. These shared feature effects were evident from self-reported ratings, behavioral intentions, evaluative priming, and IAT effects which consistently favored one target over the other. They also emerged regardless of the type (color, size, location) and nature (physical or conceptual) of feature manipulated and acquisition procedure used. These conclusions were further reinforced by our meta-analysis, in which shared features were shown to produce large effect sizes across measures of automatic evaluation, self-reported evaluation, and behavioral intentions. The relatively large degree of heterogeneity in the effect sizes between studies reflects the differing instances and implementations of shared features that were implemented between studies. Indeed, the credibility intervals (i.e., 95% CR, which reflect the likely range of effect sizes to be observed given both the estimated true effect size and the observed heterogeneity between studies) excluded zero by a large margin for all outcome variables (IAT, self-reports, and behavioral intentions), suggesting that learning via shared features is also highly likely to be observed in future studies under other as-yet unobserved implementations of the concept. We can therefore say that the shared features effect appears to be replicable, robust across a range of outcome measures and common exclusion criteria, and general across multiple instances and implementations.

Theoretical Implications

Until now we focused on the shared features principle itself and said little about why it actually emerges. There are two different levels at which to explain shared features effects (De Houwer, 2011; Hughes, De Houwer, & Perugini, 2016): (1) a mental level that aims to uncover the mental mechanisms that *mediate* the impact of the environment on behavior and (2) a functional level that aims to describe those elements of the environment that *moderate* behavior. We consider both in turn but will only briefly discuss one possible theoretical approach within each level. Undoubtedly, there are many other theoretical approaches that could be related to shared features effects. Although we look forward to the theoretical debates that our findings will spark, within the context of the present paper with its focus on establishing the principle itself, we limit speculations at the theoretical level and highlight only two approaches that appear particularly interesting to us.

The functional level of explanation: The sharing of features as a contextual relational cue. Without going into too much detail, functional explanations are not concerned with identifying mental representations and processes. Instead they seek to relate specific effects to more general behavioral principles using terms that refer to the function of events (for a detailed treatment see De Houwer & Hughes, 2020; Hughes & Barnes-Holmes, 2016). At this level shared features effects could be conceptualized as an instance of relational responding (i.e., a type of behavior that involves 'responding to the relationship between stimuli'). Relational responses are typically emitted in the presence of a stimulus called a *relational contextual cue*. This stimulus is a *contextual cue* in the sense that it signals (cues) how one should respond, and it is *relational* because it signals that a relational response should be emitted in that context. Take, for instance, a non-relational contextual cue such as a red traffic light at a busy intersection. This light signals how one should respond in that context (i.e., that walking across the street when the light is red will be dangerous for that person). Relational contextual cues require us to take this idea one step further. They also signal how one should respond. But instead of responding to just one stimulus they indicate that we should respond based to how stimuli are *related* to one another in that context.

To illustrate, imagine that you are presented with a positive word along with an unknown word. If this pair of stimuli is accompanied by the word 'SAME' this may signal to you that the unknown word has the same (evaluative) meaning as the positive word. As a result you will subsequently like the unknown word more than before. In this example the word SAME functions as a relational contextual cue: it signals that a relation of similarity exists between the unknown and positive word. One could conceptualize shared physical features such as color (Experiments 1,2,3,6), size (Experiments 4, 7) and location (Experiment 8) in much the same way: as a relational contextual cue which signaled a relation of similarity between two of the three stimuli presented in an acquisition trial (a neutral target and either a positive or negative source). Once such a relationship was formed other source features were transferred to the target (valence), thus leading to a change in evaluative responding. The fact that conceptual features also moderated evaluations and intentions (Experiment 5) is consistent with past work on the effects of relational contextual cues (Hughes & Barnes-Holmes, 2016). Thus, our shared feature effects are in line with modern (functional) conceptualizations of learning and behavior (e.g., Hayes, Barnes-Holmes, & Roche, 2001), and particularly with the idea of relational contextual cues.

The mental level of explanation: Inferential reasoning. At the mental level the goal is to identify the mental representations and processes that mediate between environment and behavior. We believe that shared feature effects fit well with an inferential perspective on human learning and attitudes (e.g., De Houwer, 2018; Van Dessel, Hughes, & De Houwer, 2019). The core conceptual unit of this perspective is a proposition, that is, an informational unit "that contains information about the nature of the relation between stimuli (e.g., A predicts B, A causes B, A co-occurs with B, ...)" (De Houwer, 2018, p.3). *Inferences* represent a sub-type of such propositions – namely – those generated from other momentarily available propositions. "The construction process that leads to the inference can be seen as an information generation procedure that involves the application of information generation (i.e., inference) rules to information that is currently entertained" (Van Dessel et al., 2019, p.4).[8]

In our studies the fact that the source and target object shared a feature may have caused participants to form a series of inferences about (a) how those stimuli were related and (b) the properties of the target object. It might have been these inferences which mediated subsequent evaluations and intentions. For instance, during the acquisition phase in Experiments 1-5, participants may have formed a number of simple propositions concerning the source and target objects (e.g., 'the target is presented in green', 'the positive source is presented green', and 'the negative source is presented in purple'). They may have also generated a number of propositions about the source and target object features (e.g., 'this word [target] is neutral', 'that word [source] is positive' and 'that word [source] is negative'). These basic propositions may have served as the 'raw ingredients' for a relational inference ('the target and source are similar in that they are both green') which could have led to an inference about the target objects features ('the source is good therefore the target is also good'). Note that these inferences were generated on the basis of both physical and conceptual features that objects shared. Thus, from an inferential perspective, the 'assumptions' at the core of shared feature effects are

[8] In this way the term 'inference' describes the outcome of the computation process rather than the computation process itself. The computational process of inferential reasoning can occur on the basis of many different inference rules (e.g., rules that encode general statements about the world [if-then rules] or rules based on mere similarity [analogical mapping rules]; for more see Van Dessel et al., 2019).

actually inferences that are constructed on the basis of past and present propositions about the source and target objects, their features, and how they are related. Also note that more research is needed to examine the exact nature of the inferences that people make, as well as the nature of the premises on which these inferences are based. Within the context of the present paper, we limit ourselves to highlighting inferential models as one possible theoretical approach that could help shed light on the mechanisms underlying shared feature effects.

Open Questions, Broader Implications, and Future Directions

In this paper we find that shared features can be used to establish self-reported and automatic evaluations of novel stimuli (Experiments 1, 3, 4, 5, 6, and 7) and influence people's first impressions of others (Experiment 8). Nevertheless, we recognize that there is still a gap between the specific findings reported here and the broader framing we offered in the introduction. In order to close this gap research will be needed showing that the shared features principle underpins both old and new phenomena in the field. Such work will need to show that the principle is responsible for changes in both evaluative and non-evaluative stimulus properties, and in ways that are relevant to clinical (e.g., fear, anxiety, disgust), social (e.g., accessibility), consumer (e.g., brand quality, brand identification), and cognitive psychology (e.g., attention). Researchers will also need to show that shared features effects not only drive evaluations but judgements, decisions, and behavior outside of the laboratory as well. Take, for instance, the brand mimicry effect studied in business and marketing: does the fact that a newly introduced brand product (e.g., a so called 'meatless' meat burger) shares a feature (shape, consistency, texture) with a known product (e.g., beef burger) lead people to assume that those products share other features (e.g., taste) and does this influence their decisions to purchase and consume the former more than the latter? Is the same true for other products – both legal (e.g., generic medications that share features with copyrighted mediations) and illegal (e.g., counterfeit luxury items that shared features with their authentic counterparts). In other words, does the fact that the target product shares one feature with a source product lead people to assume that both products share other features? If such findings were to emerge it would further support the idea that many known and to-be-discovered phenomena in psychological science represent instances of the same basic (shared features) principle.

We also limited our initial efforts to the formation of attitudes. Future work could investigate whether shared features can also be used to strengthen or revise existing evaluations as well. For instance, researchers could first establish a novel evaluation towards an unknown object, or take a pre-existing evaluations towards a known object (e.g., towards a celebrity, brand product, phobic, or clinically relevant stimulus such as spiders or alcohol). Those evaluations could then be modified using a similar task as used in Experiments 1-5. For instance, imagine that participants first complete the same acquisition phase as we used here to generate an evaluation and were then exposed to similar phase designed to reverse those initial evaluations. Researchers could vary this second task so that the target no longer shares a color with either type of source (similar to extinction in EC research; e.g., Gawronski, Gast, & De Houwer, 2015), swap the shared feature contingencies so that the target now shares a feature with the opposite source used in the first phase (similar to counter conditioning in EC research; Kerkhof, Vansteenwegen, Baeyens, & Hermans, 2011) or is exposed to the exact same contingencies as before, but between the formation and change phases, the source the target share a feature with is subjected to a revaluation procedure that changes its valence. In each case, they could examine if evaluations of the target change as a result of such manipulations.

When carrying out this work, researchers can also investigate the factors that moderate shared feature effects. to observe evaluative effects (thus it was required; Hofmann et al., 2010), whereas the other two exclusions ensure that our effects were not contaminated by undesirable sources. Given that each individual study lacked the power to address this question we opted for a meta-analytic approach instead. Each respective subset of participants was excluded and separate meta-analytic models were refit. As in the previous meta analyses, robustness tests excluded Experiment 2 as an outlier. Robustness of conclusions was defined as congruence in the acceptance/rejection of the null hypothesis between results obtained from the full sample and those obtained in a given subset.

Results were found to be robust to only including participants who were (a) aware (i.e., source valence awareness [77.5% of participants] and target source contingency awareness [59.6%]), (b) not demand compliant (80.9%), and (c) not reactant (85.8% on IAT, 89.7% on self-reports). This was found across all three outcome measures (IAT, self-reported evaluations, and behavioral intentions; all sensitivity analysis effect size *p*s < .002). In short, the general trend of evidence suggested that learning via shared features is robust to three common exclusions employed in the literature: requiring participants to be (contingency) aware, not demand compliant, and not reactant.

How shared features are established and changed may also matter: it may be easier to form and modify these effects via experience relative to observation or instruction. It might also be that these effects are subject to certain boundary conditions. For instance, in Experiments 2-3, directing attention towards the shared feature led to automatic and self-reported evaluations whereas directing attention away from that feature failed to produce such effects. One possibility is that people treat shared features as a cue that is 'diagnostic' for how they should respond to the target object (i.e., how they should evaluate the target). It may be that the impact of shared features on behavior is moderated by other cues in the environment which signal to what extent the shared feature is diagnostic or not when making a judgement. There may be still other conditions necessary for these effects to emerge and change that should also be examined. The current studies utilized only two types of procedure to document these effects and readers should be careful not to conflate the former with the latter. Many other procedures could be devised to study this class of effects.

We previously argued that shared features could be conceptualized as relational contextual cues. If so, then it should be possible to change the relational meaning of the fact that objects share features, and thus the assumptions people make about target object features on the basis of those shared features. Although a shared feature will typically signal that the source and target objects are similar to one another (and thus give rise to *feature transfer*) there is no reason why a shared feature cannot instead signal that the source and target are related in other ways (and thus give rise to *feature transformation*). For instance, it may be that the feature shared by a source and target object signals that those objects are opposite, hierarchical, different, or related in any number of other ways. This also seems like a promising research direction for future research.

Another interesting possibility is that opposite features may influence behavior as well. In a recent set of unpublished studies that were conducted in our lab participants were exposed to a simple learning task where participants had to assign valenced words to the left-side of the screen and unknown nonwords to the right-side of the screen. Following training the nonwords acquired an opposite valence to the valence items themselves. In this case, a source (valenced) and target (nonword) object did not share a feature with one another but possessed an opposite feature (i.e., one was linked to a left response and the other to a right response). The fact that they possessed opposite features (i.e., that they were assigned to opposite responses) may have led people to make certain assumptions about the target object based on the source object features (i.e., that the nonwords had an opposite valence to the source). If so then there may be an 'opposite features effect' waiting to be systematically explored.

Our findings also lead to new perspectives on human learning. In this area researchers have long made a distinction between acquisition (the emergence of a novel response in the presence of a stimulus) and generalization (the transfer of response-eliciting properties from one stimulus to another). In evaluative conditioning, for instance, researchers initially utilize an acquisition phase with contiguously presented stimuli to generate an evaluative response to a conditioned stimulus, and then use a generalization phase to test if this evaluative response transfers from one conditioned stimulus to another (e.g., Till & Priluck, 2000). Yet our work suggests that spatio-temporal contiguity might just be one way to induce a shared feature that provides the basis for feature transformation. We also argue that the shared features principle is itself rooted in the phenomenon of generalization. If we combine these two ideas we arrive at an interesting new possibility: acquisition and generalization may be more similar to one another than previously thought. For instance, in evaluative conditioning, EC effects can be conceived of as a transfer of evaluative properties based on the fact that stimuli share some *other* property with one another (common location in space and time). Evaluative generalization effects can also be conceived of as a transfer of evaluative properties based on the fact that stimuli share some *other* property with one another (common physical or conceptual property). In other words, acquisition and generalization can both be seen as feature transformation effects that occur when objects share features. When viewed in this way we see that acquisition research has often tended to focus on one shared feature (spatio-temporal properties) whereas generalization research has focused on others (perceptual or conceptual features). Yet from a shared features principle perspective, acquisition and generalization may both be instances of feature transformation.

Potential Limitations

One could ask if demand plays a role in our findings given that the experimental manipulations were relatively simple and salient (i.e., the color, size, or common location of a neutral and valenced stimulus varied in systematic ways). It may be that participants identified the manipulation and reported what they thought the researcher wanted to hear. While a reasonable suggestion, we see several reasons for why our effects are not driven by demand. First, we indexed our evaluative effects using multiple indirect tasks (IAT and evaluative priming) which are typically less sensitive to demand effects than direct measures such as self-report ratings. Second, we obtained similar effects across studies even after demand compliant participants had been removed from analyses (see the section on Robustness tests). Third, in Experiment 8, we still obtained evaluative effects even when we told people that the shared feature was irrelevant and invalid information for forming evaluations. In other words, even when people were told to disregard the key experimental factor (bag location) and reported that they knew this factor was irrelevant, they still showed an effect. This seems like a particularly strong test of the impact of shared features on evaluation. When taken together these three strands of evidence would argue against a simple demand or communication effect interpretation of our findings. Nevertheless, future work could seek to further control for this possibility (e.g., by using still other indirect measures, a better cover story, providing no instructions and just presenting the shared feature, or by adding a training phase that obscures the shared feature more extensively). That work could also examine why the effects were relatively large on one indirect procedure (IAT) and small on another (evaluative priming). It may be that additional training is needed, movement away from repetition of just two prime words, or other procedural parameters will increase the size of the effect on the latter procedure.[9]

Context of the Research

The co-authors on this paper share several features. Perhaps the most important is an interest in developing useful ways of speaking (definitions, terms, and concepts) and thinking (nomothetic systems) that can help stimulate better communication and interaction between researchers, both inside and outside of psychological science. Such an approach should be relatively broad in its scope, not limited to any one theory or content domain, and provide a common language that would facilitate dialogue between those that typically do not. This has recently given rise to a new conceptual tool-box (the Feature Transformation framework) that we hope will foster progress on at least some of these objectives. If nothing else, this new framework (and the ways of speaking and thinking it occasions) guided much of the current work, has started to shift how we view many other phenomena in psychology, and has unlocked exciting new research directions (e.g., on related and non-related feature effects). Building on these initial laboratory demonstrations of the basic principle, our next step is to showcase how previously identified phenomena (e.g., stigmatization, stereotyping, persuasion, moral, and certain marketing effects) are – fully or in part – driven by this principle, better explore its moderators and potential boundary conditions, and identify new instances of the shared feature effect.

Conclusion

In this paper we introduce the shared features principle and provide an initial test of its heuristic and predictive value. We found that when a valenced source and neutral target object shared one feature with one another (color or size), this was enough to influence assumptions about other features of the target (valence). This was true for both automatic and self-reported evaluations and when the type and nature of the shared feature was varied. Across experiments and in meta analyses, the principle was found to produce effects that were replicable, robust, and general. Although this paper focused on just one domain (attitudes) our conceptual account applies to many more, and offers a unified way to describe and analyze shared feature effects throughout psychological science. In all likelihood, there are many other domains and phenomena that could be conceptualized as instances of shared feature effects than covered here. We hope that our account will stimulate a new wave of research on this topic and contribute to a broader and deeper understanding of the way in which people arrive at assumptions about the features of objects in their environment.

Notes

This research was conducted with the support of Ghent University Grant BOF16/MET_V/002 to JDH. Correspondence concerning this article should be sent to sean.hughes@ugent.be. All materials and data are available at osf.io/pqm9v

[9] In Experiment 8 participants were informed that the contents of each bag were randomly created. Thus there is no such thing as a 'good' or a 'bad' bag, nor was there a connection between the words and images that were pulled from each bag. Although we obtained an impact of these instructions, it is not entirely clear if participants (a) viewed contingencies between the source and target objects emerging from each bag as random (when in fact they were not), or if they (b) viewed the targets as having been randomly assigned to the same bag as sources (which can be a perfectly random allocation). Future work could examine this in more detail. Such an approach would involve presenting participants with information indicating that events are randomly related which is then countered by personal experiences with those events that are not random. This would provide an interesting opportunity to test the relative power of instructions versus experiences on stimulus evaluations.

References

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics - Simulation and Computation, 39*(4), 860-864.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

De Houwer, J. (2003). A structural analysis of indirect measures of attitudes. In J. Musch & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 219–244).

De Houwer, J. (2011). Why the cognitive approach in psychology would profit from a functional approach and vice versa. *Perspectives on Psychological Science, 6*(2), 202-209.

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

De Houwer, J., & Hughes, S. (2020). *The psychology of learning: An introduction from a functional-cognitive perspective.* The MIT Press.

De Houwer, J., Richetin, J., Hughes, S., & Perugini, M. (2019). On the assumptions that we make about the world around us: A conceptual framework for feature transformation effects. *Collabra: Psychology, 5*(1), 43. doi:10.1525/collabra.229.

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin, 135*(3), 347-368.

Forgas, J. P., & Laham, S. M. (2016). Halo effects. In R. F. Pohl (Ed.), *Cognitive illusions: Intriguing phenomena in judgement, thinking and memory* (2nd ed., pp. 276–290). New York, NY: Psychology Press.

Gawronski, B., Gast, A., & De Houwer, J. (2015). Is evaluative conditioning really resistant to extinction? Evidence for changes in evaluative judgements without changes in evaluative representations. *Cognition and Emotion, 29*(5), 816–830. doi:10.1080/02699931.2014.947919

Ghirlanda, S., & Enquist, M. (2003). A century of generalization. *Animal Behaviour, 66*(1), 15–36. doi:10.1006/anbe.2003.2174

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.

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (2001) *Relational frame theory. A post-Skinnerian approach to language and cognition.* Kluwer Academic/Plenum.

Hebl, M. R., & Mannix, L. M. (2003). The weight of obesity in evaluating others: A mere proximity effect. *Personality and Social Psychology Bulletin, 29*(1), 28-38.

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: a meta-analysis. *Psychological Bulletin, 136*(3), 390-421.

Hughes, S., & Barnes-Holmes, D. (2016). Relational frame theory: The basic account. In R. D. Zettle, S. C. Hayes, D. Barnes-Holmes, & A. Biglan (Eds.), *The Wiley Handbook of Contextual Behavioral Science* (pp. 129-178). West Sussex, UK: Wiley Blackwell.

Hughes, S., Barnes-Holmes, D., Van Dessel, P., de Almeida, J. H., Stewart, I., & De Houwer, J. (2018). On the symbolic generalization of likes and dislikes. *Journal of Experimental Social Psychology, 79*, 365–377. doi:10.1016/j.jesp.2018.09.002

Hughes, S., De Houwer, J., & Perugini, M. (2016). The functional-cognitive framework for psychological research: Controversies and resolutions. *International Journal of Psychology, 51*(1), 4-14.

Kerkhof, I., Vansteenwegen, D., Baeyens, F., & Hermans, D. (2011). Counterconditioning: : An effective technique for changing conditioned preferences. *Experimental Psychology, 58,* 31-38.

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., van Schie, K., Van Harmelen, A.-L., De Schryver, M., De Winne, J., & Brysbaert, M. (2013). Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior Research Methods, 45*(1), 169–177. doi:10.3758/s13428-012-0243-8

Otten, S. (2016). The Minimal Group Paradigm and its maximal impact in research on social categorization. *Current Opinion in Psychology, 11*, 85-89.

Paustian-Underdahl, S. C., & Walker, L. S. (2016). Revisiting the beauty is beastly effect: examining when and why sex and attractiveness impact hiring judgments. *The International Journal of Human Resource Management, 27*(10), 1034-1058.

Phau, I., & Teah, M. (2009). Devil wears (counterfeit) Prada: a study of antecedents and outcomes of attitudes towards counterfeits of luxury brands. *Journal of Consumer Marketing, 26*(1), 15-27.

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

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. doi:10.3758/s13423-017-1230-y

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology, 1*(2), 149-178.

Tajfel, H., Turner, J. C., Austin, W. G., & Worchel, S. (1979). An integrative theory of intergroup conflict. *Organizational Identity: A Reader*, *56*, 65.

Teige-Mocigemba, S., & Klauer, K. C. (2013). On the controllability of evaluative-priming effects: Some limits that are none. *Cognition and Emotion*, *27*(4), 632–657. doi:10.1080/02699931.2012.732041

Till, B. D., & Priluck, R. L. (2000). Stimulus generalization in classical conditioning: An initial investigation and extension. *Psychology & Marketing, 17*(1), 55-72.

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory.* Basil Blackwell.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84(4),* 327–352. doi:10.1037/0033-295X.84.4.327

Uhlmann, E. L., Zhu, L. L., Pizarro, D. A., & Bloom, P. (2012). Blood is thicker: Moral spillover effects based on kinship. *Cognition, 124*(2), 239-243.

Unkelbach, C., & Högden, F. (2019). Why Does George Clooney Make Coffee Sexy? The Case for Attribute Conditioning. *Current Directions in Psychological Science*, 1-7.

Van Dessel, P., Gawronski, B., Smith, C. T., & De Houwer, J. (2017). Mechanisms underlying approach-avoidance instruction effects on implicit evaluation: Results of a preregistered adversarial collaboration. *Journal of Experimental Social Psychology.* doi:10.1016/j.jesp.2016.10.004

Van Dessel, P., Hughes, S., & De Houwer, J. (2019). How do actions influence attitudes? An inferential account of the impact of action performance on stimulus evaluation. *Personality and Social Psychology Review, 23*(3), 267-284. doi:10.1177/1088868318795730.

van Veelen, R., Otten, S., Cadinu, M., & Hansen, N. (2016). An integrative model of social identification: Self-stereotyping and self-anchoring as two cognitive pathways. *Personality and Social Psychology Review, 20*(1), 3-26.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1-48.