Abstraction in Attitude Acquisition: A Cognitive-Ecological Perspective on the Generalization and Robustness of Likes and Dislikes

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Kathrin Reichmann
aus Blaubeuren

Tübingen
2023

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	08.03.2024
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter/-in:	Prof. Dr. Mandy Hütter
2. Berichterstatter/-in:	Prof. Dr. Klaus Fiedler

Danksagung

Allen voran möchte ich mich bei Prof. Dr. Mandy Hütter für ihre fortwährende Unterstützung, ihr Vertrauen und ihre Zugänglichkeit während der Erstellung dieser Arbeit bedanken. Die Mitarbeit in unserer Arbeitsgruppe war mir eine besondere Freude und ich bin nach wie vor sehr dankbar über diese Gelegenheit. Gleichermaßen gilt mein Dank Prof. Dr. Barbara Kaup und Dr. Michael Ramscar für die Zusammenarbeit und Ihre Hilfe, sowie Prof. Dr. I-Ching Lee für die Betreuung meines Forschungsaufenthaltes in Taipeh.

Dankbar erwähnt seien auch die Mitglieder der Forschergruppe „Modal and Amodal Cognition", von denen ich sehr viel lernen konnte. Hier seien insbesondere Prof. Dr. Diane Pecher und Prof. Dr. Volker Franz für das Mentoring während der Promotion erwähnt. Auch bei unseren Studierenden, die mit vielen tollen Beiträgen zur vorliegenden Arbeit beitrugen, bedanke ich mich ganz herzlich. Ella Zapanta danke ich für ihre Mitarbeit und Hilfe während des letzten Jahres. Dank gebührt auch den Probanden und Probandinnen, die viel Zeit für unsere Experimente aufbrachten.

Besonders möchte ich mich bei Freunden und Freundinnen sowie Kolleginnen und Kollegen für ihre stets aufmunternden Worte, ihre Unterstützung und die vielen schönen gemeinsamen Stunden und Erlebnisse bedanken. Dazu gehören beispielsweise etwaige After-Work Drinks und besondere Zeiten auf Konferenzen mit Johanna, Tobias, Zach, und Thomas; erlebnisreiche Wanderungen mit Anne, Sven, Tanja und Sarah; sowie die vielen schönen Mittagspausen in der Rümelinstraße. Viel Halt boten außerdem unser Stammtisch und die Wochenenden mit den Lieblingspsychos. An dieser Stelle danke ich auch Johannes dafür, dass er mich zum Lachen bringt und mir immer viel Kraft gegeben hat.

Letztlich gilt mein größtes Dankeschön meiner Familie für alles, was sie mir ermöglicht; für ihre bedingungslose Liebe und Akzeptanz.

Erklärung über Eigenanteil

Chapter 1 beruht auf Manuskript Nr. 1:

Reichmann, K., Hütter, M., Kaup, B., & Ramscar, M. (2023). Variability and abstraction in evaluative conditioning: Consequences for the generalization of likes and dislikes. *Journal of Experimental Social Psychology*, *108*, 104478. Doi: https://doi.org/10.1016/j.jesp.2023.104478

Chapter 2 beruht auf Manuskript Nr. 2:

Reichmann, K., Lee, I-C., Hütter, M. (preprint). Are attitudes towards outgroup members more resistant to change? On the role of social categories in attitude change via evaluative conditioning.

Chapter 3 beruht auf Manuskript Nr. 3:

Reichmann, K., & Hütter, M. (under review). Abstract representations of attitudes: Do they make evaluative conditioning resistant to US revaluation? A study of ecological conditions.

Eigenanteil am jeweiligen Manuskript:

Manuskript Nr.	Erstautoren-schaft	Entwicklung der wissenschaftlichen Idee (%)	Datenerhebung (%)	Analyse und Interpretation (%)	Verfassen des Manuskriptes (%)
1	Ja	40	90	80	70
2	Ja	80	60	80	70
3	Ja	60	80	80	70

Abstract

In between acquisition and retrieval, information must be stored and represented in the human mind. Mental representations can vary in their level of abstractness, depending on generative conditions of the learning environment. Importantly, abstraction has consequences for learning outcomes, such as the generalization and updating of knowledge. The present thesis studied abstraction in the domain of attitude acquisition. Providing a cognitive-ecological perspective, the interplay between intrapsychic processes (i.e., abstraction) and the learning environment (i.e., learning conditions that make abstraction particularly likely) was considered to predict the generalization and robustness of likes and dislikes. Three empirical projects employed evaluative conditioning (EC) as an experimental paradigm to induce attitudes via the pairing of stimuli. Each project relied on a different theoretical perspective on abstraction to derive cognitive-ecological factors that facilitate the formation of abstract representations.

Chapter 1 focused on abstraction via the discriminative learning of cues. The results of three experiments (*N* = 505) showed that variability in attitude objects facilitate the extraction of predictive cues, leading to abstract representations of attitude objects and an increase in the generalization of acquired attitudes to novel stimuli. **Chapter 2** studied abstraction as a function of psychological distance. While members of socially distal groups (outgroups) were represented more abstractly than members of socially proximal groups (ingroups), two experiments (*N* = 222) showed that this did not affect the degree of attitude change. Lastly, **Chapter 3** investigated the way evaluative experiences themselves are represented in memory. Three experiments (*N* = 727) tested the ecological conditions that facilitate the abstraction of valence during conditioning, making attitudes resistant to a revaluation of initial evaluative experiences. The findings can be better explained by abstraction via comparison than predictive learning.

Overall, the present work allows for theoretical advancements by combining findings on abstraction with research on attitude acquisition. The projects highlight a necessary shift from studying the operating principles of evaluative learning to a focus on the format and content of acquired representations. They also offer practical implications regarding the design of interventions targeting attitude change.

Keywords: attitude acquisition, abstraction, evaluative conditioning, attitude change, generalization, mental representations

Zusammenfassung

Zwischen dem Erwerb und dem Abruf von Informationen müssen diese im Gedächtnis gespeichert und repräsentiert werden. Mentale Repräsentationen können sich in ihrem Grad der Abstraktion unterscheiden, je nachdem, in welcher Lernumwelt die Informationen erworben wurden. Der Abstraktionsgrad kann dabei Konsequenzen für den Abruf von Informationen haben, insbesondere in Bezug auf die Generalisierung und Aktualisierung des erworbenen Wissens. In der vorliegenden Arbeit wurde Abstraktion während des Erwerbs von Einstellungen und Präferenzen untersucht. Aus einer kognitiv-ökologischen Perspektive wurde hierbei das Zusammenspiel zwischen intrapsychischen Prozessen (d.h. Abstraktion) und der Lernumgebung (d.h. Lernbedingungen, die Abstraktion besonders wahrscheinlich machen) betrachtet, um die Generalisierung und Robustheit von Einstellungen vorherzusagen. Drei empirische Untersuchungen verwendeten die evaluative Konditionierung (EC), um Einstellungen durch die Paarung von Stimuli zu induzieren. Die Kapitel berücksichtigen unterschiedliche theoretische Perspektiven, um Umweltfaktoren abzuleiten, die Abstraktion fördern könnten.

Kapitel 1 konzentrierte sich auf Abstraktion durch diskriminatives Lernen. Die Ergebnisse dreier Experimente (*N* = 505) verdeutlichten, dass Variabilität in den präsentierten Einstellungsobjekten die Extraktion prädiktiver Cues und damit die Bildung abstrakter Repräsentationen von Einstellungsobjekten fördern kann. Folglich war eine stärkere Generalisierung der erworbenen Einstellungen auf neue Stimuli zu beobachten. **Kapitel 2** untersuchte Abstraktion als Funktion psychologischer Distanz. Während Mitglieder sozial entfernter Gruppen (“Outgroups”) abstrakter repräsentiert wurden als Mitglieder sozial naher Gruppen („Ingroups“), zeigten zwei Experimente (*N* = 222), dass dies keinen Einfluss auf das Ausmaß der Einstellungsänderung hatte. Zuletzt betrachtete **Kapitel 3**, wie valente Lernerfahrungen selbst im Gedächtnis repräsentiert werden. Drei Experimente (*N* = 727) testeten Umweltbedingungen, die zu einer Abstraktion von Valenz während der Konditionierung führen und Einstellungen erzeugen, die sich gegenüber einer Revaluation der initialen Lernerfahrungen resistent zeigen. Die Ergebnisse können besser durch Abstraktion über Vergleichsprozesse, als Abstraktion über prädiktives Lernen erklärt werden.

Zusammenfassend kombiniert die vorliegende Arbeit Erkenntnisse zu Abstraktionsprozessen mit Forschung zum Einstellungserwerb. Die Projekte veranschaulichen die Notwendigkeit, neben kognitiven Prozessen des Einstellungslernens auch das Format und den Inhalt der erworbenen Repräsentationen zu berücksichtigen. Darüber hinaus lassen sich aus den Ergebnissen praktische Implikationen für die Gestaltung von Maßnahmen zur Einstellungsänderung ableiten.

Keywords: Einstellungserwerb, Abstraktion, Evaluative Konditionierung, Einstellungsänderung, Generalisierung, mentale Repräsentationen

Contents

Abstraction in attitude acquisition: A cognitive-ecological perspective on the generalization and robustness of likes and dislikes

Understanding human learning is at the heart of understanding human behavior and the human mind. The very general question of how humans learn has been discussed for decades and in various domains of research. One part of the question refers to the way information is stored and represented in memory (Kaup et al., 2023). Mental representations can vary not only in their sensory qualities (Barsalou, 2008; Meteyard et al., 2012), but also in their abstractness (Burgoon et al., 2013; Gilead et al., 2020; Reed, 2016; Trope & Liberman, 2010). Abstract representations are less detailed than concrete ones, but widely applicable (Gentner & Hoyos, 2017; Trope & Liberman, 2010). The ability to abstract information from a larger body of experiences explains why humans perform so well in making inferences from only sparse data sets (Tenenbaum et al., 2011), can engage in prospective thought (Gilead et al., 2020), and are able to move beyond an egocentric reference point (Liberman & Trope, 2008). “The blessing of abstraction” thus lies in a top-down route of cognitive processing that allows for efficient learning and the transcendence of the here-and-now (Gilead et al., 2020; Tenenbaum et al., 2011).

Research on abstraction has focused extensively on both the causes and consequences of abstract knowledge. Causes of abstraction are anchored in generative conditions of the learning environment (Smith, 2014). For example, abstraction becomes likely under learning conditions that allow cues to compete for relevance (Ramscar et al., 2010), or when the psychological distance of a reference object to the self increases (Trope & Liberman, 2010). Thus, specific aspects of the learning environment can determine the occurrence of abstraction. Importantly, abstraction has consequences for learning outcomes. The degree of knowledge generalization can increase with an increasing abstractness of representations (e.g., Gentner & Smith, 2013; Pearce, 1987; Ramscar et al., 2010; Shepard, 1987), and acquired knowledge can become more robust towards changes in the environment and external influences (e.g., Dayan & Berridge, 2014).

Within social psychology, abstraction processes have been considered relatively rarely in the last few years (but see Ledgerwood, 2014, and McCrea et al., 2012, for notable exceptions). This is rather surprising, considering that learning outcomes such as the generalization and robustness of knowledge play a central role in social contexts – especially when it comes to the generalization and robustness of attitudes. Attitudes can be defined as “summary evaluations of an object” (Fazio et al., 2007; p.608) or “global evaluative assessments” (Hütter, 2022; p. 640), with consequences for the way (attitude) objects are evaluated.[1] Attitudes generalize when an evaluation of a new stimulus corresponds to the evaluation of a previously learned stimulus. For example, generalization occurs when a novel person looks like a familiar one (FeldmanHall et al., 2018), or when a novel person is categorized as a member of a known social group (Glaser & Kuchenbrandt, 2017; Park & Hastie, 1987; Ranganath & Nosek, 2008). While the ability to generalize is an important one to navigate in an uncertain and complex word (Wu et al., 2018), it can have negative consequences such as the emergence of prejudice and discrimination (Dovidio & Gaertner, 1999; Gilmour, 2015). Moreover, the uncertainty and complexity of the environment also requires that attitudes can be flexibly updated in the light of novel information. Several studies showed that likes and dislikes are sensitive to revaluations of the initial evaluative experiences (Baeyens et al., 1992; Jensen-Fielding et al., 2018; Peters & Gawronski, 2011; Sweldens et al., 2010; Walther et al., 2009). In addition, novel information on attitude objects can change existing attitudes (Calanchini et al., 2013; Kerkhof et al., 2011; Olson & Fazio, 2006; Rydell et al., 2007), but this is restricted to specific circumstances (e.g., Bettencourt et al., 1997).

Whereas previous research showed that attitudes generalize or are updated in the light of novel information, as of now a deeper understanding of the conditions promoting or diminishing either learning outcome is still lacking. The central aim of the present thesis is to consider abstraction during attitude acquisition to deepen our understanding of the generalization and robustness of likes and dislikes. Taking on a cognitive-ecological perspective (Fiedler, 2014), evaluative judgements (i.e., generalization and robustness) are considered as outcomes of the interplay between intrapsychic processes (i.e., abstraction) and the learning environment (i.e., learning conditions that make abstraction particularly likely). This approach offers new perspectives on instances where attitudes overgeneralize (e.g., in prejudice), are difficult to modify (e.g., when combating intergroup biases), or do not change when initial information are revaluated. Moreover, it allows for theoretical advancements by combining long-standing findings on abstraction from cognitive psychology with research on attitude acquisition, to help us understand how the format and content of mental representations influences evaluative judgements.

[1] *Attitude objects* can be generally defined as the entity that is evaluated. Such an entity can be “anything that is discriminable or held in mind” (Eagly & Chaiken, 2007, p. 583). Attitude objects differentiate attitudes from other concepts (e.g., mood involves an evaluative reaction but is not directed towards an entity; Eagly & Chaiken, 2007).

Abstraction

Abstraction, in general, describes the cognitive process of "identifying a set of invariant central characteristics of a thing" (Burgoon et al., 2013, p. 502; see Barsalou, 2003; Gilead et al., 2020; Reed, 2016; for discussions of a definition of abstraction). For example, abstraction occurs during the formation of a social category, when various individuals are grouped together according to their common features (e.g., categorization based on race or gender). The resulting representation has a lower number of corresponding features in the environment than the original learning experiences, making the representation more abstract.[2] At the same time, just because abstract representations are less detailed than concrete ones, they do not convey less information. Instead, abstract representations hold *additional* information, namely regarding features and relations that are applicable across instances (Trope & Liberman, 2010). Thus, while abstraction decreases the specificity of a concept, it increases the scope of it (Gentner & Hoyos, 2017; Liberman & Förster 2009). For example, an attitude towards a social group is applicable to a higher number of instances than an attitude towards a specific person. In other words, mental representations can be both relatively more abstract (less detailed but broader) or concrete (more detailed but narrowly applicable).

Different theoretical accounts exist that specify which features are retained and which ones get omitted during abstraction. For example, abstraction via the *discriminative learning of cues* (e.g., Hoppe et al., 2020; Ramscar et al., 2010; Ramscar, 2021) predicts that abstract representations hold the features most predictive of an outcome, while less predictive ones get discarded. Alternatively, *Construal Level Theory (*Trope & Liberman, 2010) posits that features invariant across different dimensions of psychological distance (e.g., the social distance of an attitude object to the self) make up an abstract representation. Lastly, *abstraction via comparison* suggests that relational commonalities across instances are abstracted in learning (Gentner, 2005; Gentner & Markman, 1997), leading to the acquisition of relational concepts like *above* or *positive*. Importantly, these different accounts of abstraction make complementary predictions for generative conditions of the learning environment that should facilitate abstraction. Thus, they allow for a specification of cognitive-ecological factors that should lead to abstraction in learning.

[2] Interchangeably with the notion of "abstract representations", one could also speak of a "high-level construal" (Trope & Liberman, 2010), a "high-level representation" (Shapira et al., 2012), a "summary representation (Barsalou, 2003), or an "abstract mindset" (Gilead et al., 2014). "Features" can refer to both observable and non-observable characteristics of objects and individuals (e.g., the way people look like, but also their nationality, ethnicity, and social group membership).

Abstraction via discriminative learning

First, abstraction via discriminative learning can occur when the learning environment allows cues to compete for relevance (Ramscar et al., 2010; Ramscar, 2021). When cues compete for relevance ("cue competition"), learners can discriminate between relevant and irrelevant cues (Hoppe et al., 2022; Ramscar, 2021), which facilitates the extraction of invariant characteristics across stimuli (Burgoon et al., 2013). Cue competition, in general, describes a learning mechanism that reinforces reliable cues and devalues unreliable ones via the occurrence of prediction errors. When expected outcomes differ from received outcomes, predictions are adjusted according to the novel experience. For example, this can be the case in a social context when a generous response is expected from another person but a less generous one is provided (Hackel et al., 2015).

Prediction errors can occur in information ecologies where various complex stimuli predict a set of discrete outcomes. As an example, consider a situation where multiple attitude objects with one common feature (e.g., products with the same brand name) repeatedly co-occur with positive experiences. As the common feature of the attitude objects becomes predictive of the positive outcome (i.e., the brand name would predict a positive experience), abstraction can occur. To summarize, variability in attitude objects in the environment is an aspect that can trigger cue competition and thus abstraction (Chapter 1; see Apfelbaum & McMurray, 2011; Raviv et al., 2022 for a similar argument in language acquisition). In addition, the presentation sequence of stimuli during learning can either diminish or facilitate the occurrence of cue competition (Hoppe et al., 2020; Ramscar et al., 2010). The stimulus that is presented first generates a hypothesis that is then confirmed or updated by the second stimulus. For example, the presentation of an attitude object might generate the expectation of a positive outcome that is then either confirmed or disconfirmed upon experience of the actual outcome. In the context of conditioning, such a procedure would resemble "forward conditioning", whereby conditioned stimuli (CSs) occur before unconditioned stimuli (USs; Baeyens et al., 1993; Hammerl & Grabitz, 1993; Miller et al., 1995). Stimulus presentations in the reversed sequence do not allow for these predictions, and thus should also diminish abstraction during learning (Chapter 3).

Abstraction and psychological distance

Whereas cue competition should result in abstraction in variable learning environments and in dependence of the presentation sequence of stimuli, another cognitive-ecological factor that drives abstraction is psychological distance. According to Construal Level Theory (CLT; Trope & Liberman, 2010), psychologically distal objects, situations and individuals are generally represented in more abstract terms than proximal ones. Psychological distance is egocentric, in a way that distance dimensions such as time, space, social distance and hypotheticality refer to the distance between a reference point (the self in the here and now) and an object removed from that point. To maintain perceptual consistency across distance, abstract representations are required that hold features invariant across distance dimensions (Trope & Liberman, 2010). While invariant features are present at both near and far distance points, the processing of concrete features varies with distance (Kim et al., 2009; Soderberg et al., 2015). For example, an object such as a chair might be represented in terms of its overall shape and proportion in spatial distance, and in terms of specific details in spatial proximity. Another example provides the *outgroup homogeneity effect* (Boldry et al., 2007; Park & Judd, 1990). According to this effect, people judge members of outgroups as more similar to one another than members of the ingroup. The finding is in line with the idea that outgroup members are represented in a more abstract way, as they are socially more distant than ingroup members (Hess et al., 2018; Linville et al., 1996). Thus, learning environments with psychologically distal rather than proximal objects should result in the encoding and retrieval of abstract knowledge representations (Chapter 2).

Abstraction via comparison

Lastly, learning environments that foster comparison should also foster abstraction. Comparison, like cue competition and psychological distance, can facilitate abstraction by determining the relevance of specific attributes in the environment. While cue competition and psychological distance do so via the predictive value of features or their consistency across distance, comparison leads to the extraction of common relational structures via the structural alignment of two or more events (*structure-mapping theory*; Gentner, 1983; Gentner, 2005; Markman & Gentner, 1993). Structural alignment goes beyond the identification of surface similarities across stimuli and highlights their relational commonalities. For example, a bicycle shares surface features with a pair of glasses (i.e., both have round shapes) and relational commonalities with a skateboard (i.e., both are vehicles). Structure-mapping theory would predict a preference for matching the bicycle with the skateboard over the glasses (the systematicity principle; Namy & Gentner, 2002). In other words, comparison via structural alignment leads to the extraction of relations that go beyond perceptually similar features of training exemplars.

Accordingly, learning conditions that make relational similarities salient across learning instances should also facilitate abstraction. Such learning conditions include an increase in the number of objects per category (Chapter 1; Christie & Gentner, 2010; in line with the predictions of the discriminative learning account of abstraction mentioned above). Common labels of objects should facilitate comparison as well (Christie & Gentner, 2014; Namy & Gentner, 2002). For example, when one attitude object co-occurs reliably with various positive outcomes, the overall valence of experiences should become salient and get abstracted. The object highlights the common relational structure of the experiences (i.e., their positive connotation), even if the experiences do not share surface similarities. As a result, the attitude object gets linked to positivity (Chapter 3). In an applied context, such a situation could occur, for instance, when marketers use different brand ambassadors to advertise their product. If the brand ambassadors do not share surface features (e.g., gender or hair style), but are all well-known, popular, and perceived positively, these common relational aspects should get abstracted during learning. The resulting representation would link the product to the positive attributes of the brand ambassadors rather than their individual perceptual features.

Interim summary

The different theoretical accounts of abstraction highlight that the formation of abstract representations is anchored in generative conditions of the learning environment. They make converging predictions for some aspects of the information ecology (e.g., variability in attitude objects), while making contradictory predictions for other ones (e.g., the presentation sequence of stimuli; see Chapter 3). Nevertheless, they allow for a specification of cognitive-ecological factors that determine the format and content of acquired representations. Consequently, this influences how representations are applied during judgement. For example, Ledgerwood (2014; p. 438) argues that “the way we think about […] an object can influence the way we evaluate it”, describing how abstraction in the context of attitudes can influence evaluative responding. Two learning outcomes seem to be particularly relevant here: First, the generalization of acquired knowledge and second, the robustness of knowledge representations to novel information.

Abstraction and generalization

First, because abstract representations are less detailed than concrete ones and are widely applicable across stimuli (Gentner & Hoyos, 2017; Liberman & Förster, 2009), they allow for inferences regarding novel instances and thereby facilitate generalization (Gentner & Markman, 1997). In fact, some authors even argue that abstraction is necessary for meaningful generalization, as abstract representations guide inference from incomplete data sets (Tenenbaum et al., 2011). Generalization, in general, occurs when a subject responds to a new stimulus in the same way as to a previously learned stimulus (Jäkel et al., 2008; Shepard, 1987). In contrast to abstract representations, concrete representations serve as a weaker basis for generalization. When concrete representations are activated, their irrelevant features are retrieved as well (Smith, 2014). On the other hand, when abstract representations are activated, no additional extraneous features are retrieved and thus the overlap between the new stimulus and the activated representation increases. As a result, the degree of generalization increases (Shepard, 1987). For example, Son and colleagues (2008) presented simplified versus complex training instances to 12- and 24-month-old children. Generalization increased for simplified relative to complex training instances, which directly demonstrates the link between abstraction and generalization. Other examples stem from category learning (Bowman & Zeithamova, 2020; Ramscar et al., 2010; Son et al., 2008), analogical reasoning (Christie & Gentner, 2010; Gentner & Hoyos, 2017), stereotype formation (Park & Hastie, 1987; Ranganath & Nosek, 2008), and evaluative learning (Chapter 1).

Abstraction and robustness

Secondly, because abstract representations can be seen as more detached from specific referents in the environment than concrete representations, they are more robust to environmental changes and external influences (Dayan & Berridge, 2014; Ledgerwood, 2014; Ledgerwood et al., 2010). For example, when participants were led to adopt an abstract mindset, evaluations of an attitude object became less susceptible to incidental social influences than when participants were led to adopt a concrete one (Ledgerwood et al., 2010; Experiments 2 and 4). Moreover, representations acquired during operant conditioning might only store outcome values but not specifics of the outcomes, which should leave conditioned responses unaffected by post-conditioning changes in the reward structure of the environment (Daw et al., 2005; Dayan & Berridge, 2014). On the other hand, operant conditioning might also lead to the acquisition of detailed representations of external incentives. Here, conditioned responses should reflect the current knowledge of the environment (Dayan & Berridge, 2014; Dayan & Niv, 2008).

To conclude, the cognitive-ecological factors identified above should have direct consequences for the way knowledge is generalized and updated, as both learning outcomes depend on the abstractness of mental representations.

Abstraction in attitude acquisition

In the context of attitudes, the generalization and updating of knowledge plays a central role as well, as both aspects can have positive side-effects (e.g., influencing the effectiveness of interventions targeting attitude change; Paluck et al., 2021) but also negative ones (e.g., leading to prejudice and discrimination; Gilmour, 2015). The different theoretical accounts of abstraction considered above offer insights into the learning conditions that should facilitate or diminish abstraction during attitude acquisition. They can thus be seen as important boundary conditions for evaluative judgements, influencing the generalization and robustness of likes and dislikes.

Learning conditions, in general, vary broadly between experimental paradigms leading to the formation of attitudes. For example, the *impression formation paradigm* presents evaluative verbal information together with faces of individuals (Rydell & Gawronski, 2009), *persuasion* induces attitudes via instructions (Petty & Cacioppo, 1986), and *evaluative conditioning* refers to attitude acquisition via stimulus pairings (De Houwer et al., 2001; Hofmann et al., 2010). The paradigms do not only differ in the modality of the presented material (verbal, pictorial etc.), but also in the way attitude objects co-occur with evaluative experiences (the number of times a stimulus is presented, the number of training stimuli etc.). Nevertheless, they all can lead to the formation or modification of evaluative representations in memory. Evaluative representations consist both of an attitude object (the entity an attitude refers to) and the evaluative meaning attached to the attitude object. Theoretically, both elements of evaluative representations can vary in their level of abstractness. An attitude object can be a concrete exemplar, a category of exemplars, a perceptual cue, or another entity of varying abstractness (see Eagly & Chaiken, 2007). Similarly, evaluative meaning can refer to a concrete learning experience or to an overall notion of valence (positive versus negative; see also Gawronski & Bodenhausen, 2018). Evaluative representations form the basis of evaluative responding (Hütter & Rothermund, 2020), with observable outcomes on direct (e.g., rating scales) and indirect (e.g., performance-based) measures of attitudes (Gawronski & Brannon, 2018). A theoretical distinction between the acquisition stage (formation of evaluative representations) and the retrieval stage (activation and application of evaluative representations) of evaluative learning helps to disentangle processes acting upon encoding from those acting upon retrieval (Hütter & Rothermund, 2020). The present work concentrates on abstraction at the acquisition stage, but abstraction might play a role during retrieval as well (see the General Discussion).

Evaluative conditioning

One of the most intensely studied and most straightforward and effective ways to induce attitudes is via evaluative conditioning (EC; De Houwer et al., 2001; Hofmann et al., 2010; Levey & Martin, 1975). The EC paradigm has been employed to study various phenomena in social psychology (e.g., implicit evaluative learning; Olson & Fazio, 2001), consumer science (e.g., the controllability of attitudes; Hütter & Sweldens, 2018) and health psychology (e.g., interventions for healthy eating behavior; Hollands & Marteau, 2016;

Masterton et al., 2021; Zerhouni et al., 2019). In EC, attitude objects (conditioned stimuli; CSs) co-occur in spatiotemporal contiguity with stimuli of positive or negative valence (unconditioned stimuli; USs). As a result of the pairings, evaluations of CSs change in the direction of the US valence, which is also referred to as the *EC effect* (Hofmann et al., 2010). In EC, evaluative learning occurs via the acquisition or modification of evaluative representations that link attitude objects (CSs) to evaluative meaning (USs) in one mental episode. Upon exposure to the CS, the evaluative representation gets activated and forms the basis for evaluative responding, leading to a bias in the evaluation of CSs (De Houwer, 2011).

A lot of past studies on EC focused on the operating conditions and operating principles of EC. Operating conditions refer to the automaticity of cognitive processing (Corneille & Stahl, 2019; Hütter & Rothermund, 2020), while operating principles circumscribe the quality of cognitive processes and make a distinction between processes that operate in an associative versus propositional manner (De Houwer, 2018; Gawronski & Bodenhausen, 2018; Hütter, 2022). Propositional learning, as opposed to associative learning, involves the encoding of particular relations between co-occurring stimuli as well as the assessment of their truth value (Gawronski & Bodenhausen, 2009, 2018). In the last years, the discourse converged on the assumption that (largely non-automatic), propositional processes alone can explain outcomes in EC studies (Corneille & Stahl, 2019). Nevertheless, many open questions remain on the kind of information that is stored and represented specifically during evaluative conditioning (Gawronski & Bodenhausen, 2018; Hütter, 2022). Here, the distinction between propositional versus associative learning provides little information. Associative learning may lead to the acquisition of links between stimulus identities (S-S learning), or links between stimulus identities and an evaluative response elicited by the US (S-R learning). In a similar vein, propositional learning may produce representations of stimulus identities (similar to S-S learning), or representations that capture the inferred valence of a CS (similar to S-R learning; Gawronski & Bodenhausen, 2018). Thus, the format and content of evaluative representations remains vague under different operating principles. One way to approach this manner is to study evaluative representations in dependence of the learning environment (Hütter, 2022). For example, both S-S and S-R learning seem to play a role in EC, and specific aspects of the conditioning procedure can strengthen the role of either one (Gast & Rothermund, 2011; Sweldens et al., 2010; Walther et al., 2018). As an example, Gast and Rothermund (2011) suggested that evaluative responding during conditioning leads to S-R learning, while passively attending the stimuli facilitates S-S learning. Importantly, S-S learning and S-R learning differ in the way evaluative meaning is represented in memory.

Because S-S learning involves the representation of stimulus identities whereas S-R learning only retains evaluative responses, the latter representation can be considered as more abstract than the former one.

At the same time, not only evaluative meaning (USs) but also attitude objects (CSs) may be represented on varying levels of abstraction. For example, Glaser and Kuchenbrandt (2017) demonstrated that attitudes towards CSs ("aliens") generalized to whole categories of CSs ("alien tribes"). In a similar vein, Hütter and colleagues (2014) made a distinction between evaluative identity conditioning and evaluative cue conditioning. In evaluative cue conditioning, evaluative meaning gets attached to CS cues (e.g., gender), that are predictive of US valence beyond individual CSs. Conditioned evaluations towards cues generalized to novel stimuli sharing the cue with the original stimuli. The finding again showed that representations of attitude objects (CSs) are not restricted to specific stimuli but can take on the form of abstract entities.

To summarize, these past studies demonstrated that EC can involve higher-order cognitive processes that store knowledge in less detailed, but widely applicable representations. This applies both to the ways attitude objects (CSs) and their evaluative meaning (USs) are represented in memory.

The present thesis

The present thesis moves beyond this previous work and considers abstraction in interplay with the learning environment to predict evaluative judgements. More specifically, the different theoretical perspective on abstraction considered above (discriminative learning, psychological distance, comparison) allows one to derive cognitive-ecological factors leading to the formation of abstract representations, thereby influencing the generalization and robustness of acquired preferences. With that, the present work does not only touch on the question whether EC effects generalize (e.g., Hütter & Tigges, 2019) or whether they prove robust towards novel information (e.g., Walther et al., 2009), but also tests boundary conditions of each learning outcome.

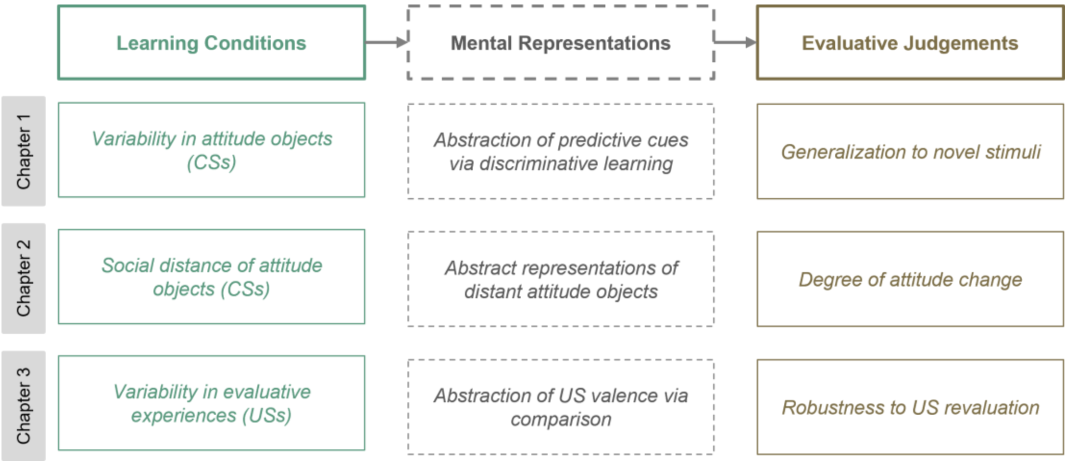
Chapter 1

First, while a number of past studies demonstrated *generalization effects* in EC, it is still unclear what kind of learning experiences facilitate or diminish generalization. Generalization effects were obtained for stimuli similar to CSs (e.g., CSs displayed from a different angle; Hütter & Tigges, 2019), for novel stimuli of the same category (Glaser & Kuchenbrandt, 2017; Jurchiş et al., 2020; Luck et al., 2020), and stimuli symbolically related to CSs (Hughes et al., 2018), as well as for stimuli containing the same cue as the CSs (Hütter et al., 2014). The occurrence of generalization effects in EC not only strengthens the external validity of the paradigm (Hütter & Tigges, 2019), but also allows for inferences regarding phenomena of overgeneralization such as prejudice (Glaser & Kuchenbrandt, 2017). Moreover, generalization plays an important role in the design of effective interventions aiming to modify attitudes not only towards training, but also novel stimuli (e.g., interventions targeting prejudice via EC; French et al., 2013; Lai et al., 2014; General Discussion).

The first empirical project presented here focused on the cognitive-ecological conditions of the generalization of likes and dislikes (see *Figure 1*). Drawing on discriminative learning as one means of abstraction, we manipulated the variability of attitude objects (CSs) to create learning conditions that enhance or diminish the occurrence of cue competition. Under high variability of CSs, cue competition should facilitate the abstraction of predictive cues, leading to attitudes towards abstract entities rather than concrete instances. In contrast, when the variability of CSs is low, specific CSs should acquire evaluative meaning as cues cannot compete for relevance in this learning environment. As a result, generalization towards novel stimuli should be stronger when CS variability is high rather than low. Three experiments manipulated CS variability by including either one, or multiple different exemplars of a category as CSs in the conditioning phase. We measured participants' evaluations of the CSs as well as novel exemplars of the categories (generalization stimuli) and included a recognition memory task and evaluative measures of CS components as manipulation checks.

Figure 1

Schematic Overview of the Present Thesis


Learning Conditions
Mental Representations
Evaluative Judgements
Chapter 1
Variability in attitude objects (CSs)
Abstraction of predictive cues via discriminative learning
Generalization to novel stimuli
Chapter 2
Social distance of attitude objects (CSs)
Abstract representations of distant attitude objects
Degree of attitude change
Chapter 3
Variability in evaluative experiences (USs)
Abstraction of US valence via comparison
Robustness to US revaluation

Note. Three empirical projects studied how learning conditions relate to evaluative judgements by considering abstraction during learning. CSs = conditioned stimuli. USs = unconditioned stimuli.

Chapter 2

Next to generalization, the robustness of likes and dislikes was the second learning outcome considered in the present study. Robustness plays a role in the context of attitudes from two different points of view. First, the *degree of attitude change* describes the sensitivity of evaluative representations to novel information. Measured as the change in linking before and after conditioning, it indicates how effectively EC can modify existing attitudes. In practical settings, it is crucial to understand the learning conditions that facilitate attitude change. A prime example are interventions trying to modify (maladaptive) attitudes, for example towards unhealthy food (e.g., Bui & Fazio, 2016; Hensels & Baines, 2016; Masterton et al., 2021), or the self (Thew et al., 2017).

The second project included in this thesis studied the degree of attitude change in dependence of the social distance of attitude objects to the self (*Figure 1*). Considering that psychological distance of objects can be related to the abstractness of representations (Trope & Liberman, 2010), socially proximal CSs should be represented more concretely than socially distal CSs. For example, this is the case in the context of social categories, where members of one’s own group were found to be perceived as less similar to one another than members of another group (outgroup homogeneity; Boldry et al., 2007; Judd & Park, 1988; Linville et al., 1989; Park & Rothbart, 1982). As a consequence of the way CSs are represented as a function of social distance, attitude change might vary accordingly. In particular, existing attitudes might be easier to modify for socially proximal than distal CSs, as past studies showed larger degrees of evaluative learning for distinct than similar stimuli (e.g., Glaser & Kuchenbrandt, 2017; Hütter et al., 2014). Two experiments used faces as CSs that were labeled as either from the ingroup, or an outgroup prior to conditioning. The faces then co-occurred with positive or negative USs during learning, and participants’ attitudes towards the faces were measured prior and post conditioning via indirect (the evaluative priming task; Fazio et al., 1995) and direct measures (a continuous rating scale) of attitudes.

Chapter 3

In addition to attitude change, likes and dislikes can also prove robust when they do not reflect modifications of initial evaluative experiences. In EC, *US revaluation* procedures allow for a test of the sensitivity of attitudes towards retrospective changes of the US valence. In US revaluation, USs are paired with information contradicting their inherent valence after the conditioning phase (e.g., a smiling face could occur with the statement “makes fun of others”). If evaluations of CSs reflect the revaluated US valence rather than their initial valence, one can infer that evaluations are sensitive to a postconditioning change of evaluative learning experiences. From a theoretical perspective, this implies that evaluations of CSs must depend on specific US identities (in line with “S-S learning”; Walther et al., 2009). From a practical perspective, such outcomes can have (un)desirable consequences. For example, in a consumer context the downfall of a celebrity endorser of a brand would lead to the devaluation of the brand (Sweldens et al., 2010). Interestingly, while evaluative conditioning was generally found to be sensitive to US revaluation procedures (Gast & Rothermund, 2011; Jensen-Fielding et al., 2018; Walther et al., 2009), certain procedures yielded EC effects that were insensitive to US revaluation (Sweldens et al., 2010; Gast & Rothermund, 2011). As of now it remains an open question what underlying process qualifies the relative influence of US revaluation on the size of EC effects (but see Walther et al., 2018).

The third research project explored the hypothesis that abstraction on the side of the US can make EC effects resistant to US revaluation (*Figure 1*). Building on the findings of Sweldens and colleagues (2010; Experiment 1), we first compared a conditioning procedure that presented one CS with different USs of the same valence (“one-to-many” pairings) with a procedure that presented one CS with the same US (“one-to-one” pairings). In line with the findings of Sweldens and colleagues (2010), smaller US revaluation effects were expected for “one-to-many” than “one-to-one” pairings. Taking abstraction into account, such findings would be compatible both with abstraction via predictive learning (CSs predict the US valence with higher accuracy than specific USs) and abstraction via comparison (multiple USs allow learners to align the exemplars and identify valence as their common element). However, the two accounts make opposing predictions regarding the presentation sequence of stimuli. Thus, two additional experiments presented CSs and USs sequentially, either in a forward, or a backward manner. This manipulation should only matter for abstraction via predictive learning, but not via comparison. With the outcomes of this line of research, inferences can be drawn regarding the learning conditions that should make EC effects more robust to US revaluation.

To summarize, the present thesis provides a cognitive-ecological perspective on the generalization and robustness of likes and dislikes by considering procedural aspects of EC in interaction with abstraction during learning. Whereas only three learning outcomes are considered here (generalization, attitude change, sensitivity to US revaluation), a similar approach might be applied to other outcomes as well (e.g., the context-dependency of attitudes). Moreover, it is likely the case that abstraction plays a role in attitude acquisition beyond evaluative conditioning (see the General Discussion).

Methodological considerations

Studying mental representations comes with the difficulty that they cannot be observed directly, and their format and content needs to be inferred from appropriate measures (Burgoon et al., 2013). The empirical projects presented here mostly (but not only) relied on memory measures to test whether details of the learning episode, one hallmark of abstract representations, are retained during learning. It is important to note that this does not mean that the link between learning conditions and evaluative judgements should be fully mediated by memory. The performance in the memory measures may be impacted by numerous different factors independent of abstraction. Nevertheless, the results of the memory measures allow one to test whether different cognitive-ecological factors indeed yield representations of varying content (as a manipulation check; Fiedler et al., 2021). It is an important task for future research to develop and employ additional measures for the way evaluative experiences are stored in memory (see the General Discussion).

Next to an appropriate choice of measures, drawing conclusions from given data requires the selection of suitable statistical approaches (McElreath, 2016). The present thesis mostly relied on multilevel models that can account for complex variance structures (heteroscedasticity) in the data and take inter-individual differences and item-specific differences into account (Behm et al., 2013; Judd et al., 2017; Vanbrabant et al., 2015). Moreover, Bayesian analyses were conducted in Chapter 3 to test whether null effects provide evidence for the null hypothesis, or indicate data insensitivity (Dienes, 2014; Hoijtink et al., 2019; Kruschke, 2018). All experiments included in this thesis adhere to open science guidelines (e.g., Klein et al., 2018). The data of the experiments, analysis scripts, and pre-registrations are publicly available on OSF. The links are included in the methods sections of each chapter.

CHAPTER 1: STIMULUS VARIABILITY AND GENERALIZATION

Variability and abstraction in evaluative conditioning: Consequences for the generalization of likes and dislikes

Kathrin Reichmann, Mandy Hütter, Barbara Kaup, & Michael Ramscar
Eberhard Karls Universität Tübingen

Article accepted for publication at the *Journal of Experimental Social Psychology*.
Doi: https://doi.org/10.1016/j.jesp.2023.104478

Author Notes

Kathrin Reichmann, Mandy Hütter, Barbara Kaup, Michael Ramscar, Department of Psychology, Eberhard Karls Universität Tübingen, Germany.

The present research was supported by a grant from the Deutsche Forschungsgemeinschaft DFG (381713393) awarded to Mandy Hütter, Barbara Kaup, and Michael Ramscar as part of the research group *Modal and Amodal Cognition* (RU 2718).

We have no conflict of interest to disclose.

Correspondence concerning this article should be addressed to Kathrin Reichmann or Mandy Hütter at the Fachbereich Psychologie, Eberhard Karls Universität Tübingen, Schleichstraße 4, 72076 Tübingen, Germany. Electronic mail may be sent to kathrin.reichmann@uni-tuebingen.de or mandy.huetter@uni-tuebingen.de.

Abstract

The present work examines whether the variability of attitude objects at attitude acquisition increases the generalization of likes and dislikes. In particular, variability might enhance the discriminative learning of cues, resulting in attitudes towards abstract entities rather than concrete instances. Using evaluative conditioning as an experimental paradigm to study attitude acquisition, we manipulated the variability of conditioned stimuli (CSs) that were paired with unconditioned stimuli (USs) of negative or positive valence. CSs resembled Chinese characters that could be grouped into categories by one common component. In the *invariable* condition, one item per category served as CSs. In the *variable* condition, multiple items per category were used as CSs. We measured participants' evaluations of the CSs and novel Chinese characters (generalization stimuli) and included a recognition memory task and evaluative measures of CS components. As compared to the invariable condition, the learning condition that introduced variability among CSs facilitated generalization towards novel stimuli, diminished recognition memory performance, and produced evaluative ratings of CS components that were more extreme for common components. The findings suggest the formation of attitudes towards abstract cues rather than concrete instances in the variable relative to the invariable condition and propose that high variability facilitates the generalization of likes and dislikes. We discuss mechanistic explanations as well as practical implications with regard to the formation of prejudice and intergroup biases.

Keywords: evaluative conditioning, variability, abstraction, generalization, cue competition

Attitudes and preferences are important determinants of human behavior, guiding social decision-making in various situations such as hiring novel employees (Von Helversen et al., 2014) or choosing a candidate to vote for (Galdi et al., 2008; I. C. Lee et al., 2016). Humans often rely on likes and dislikes towards similar individuals, objects, and situations when making judgments and decisions on novel ones. At the same time, the generalization of attitudes can also have negative side effects. For instance, evaluations that are generalized from an individual to a social group can lead to prejudice and discrimination (Gilmour, 2015; Le Pelley et al., 2010). Considering the far-reaching consequences of attitude generalization, it is not surprising that the topic has received much research interest (e.g., Glaser & Kuchenbrandt, 2017; Högden et al., 2020; Hütter et al., 2014; Luck et al., 2020; Von Helversen et al., 2014; Walther, 2002).

While much of this work has sought to understand how attitudes are generalized at the judgment stage (e.g., Högden et al., 2020; Von Helversen et al., 2014), little is known about the learning conditions that promote or diminish the generalization of attitudes. To address this question, one must consider (a) that different learning experiences can result in different cognitive representations of attitudes, and (b) that evaluations of familiar and novel attitude objects might depend in central ways on how attitudes are represented in memory (Hütter, 2022; Hütter & Rothermund, 2020).

Drawing on findings in other domains of learning research (e.g., Christie & Gentner, 2010), the present work focuses on the variability of training input as one means of generalization. Exposing learners to variable inputs has been described as an effective way of improving generalization in learning (Apfelbaum & McMurray, 2011; Estes & Burke, 1953; Hahn et al., 2005; Raviv et al., 2022). We propose that the variability of stimulus objects encountered at attitude acquisition influences how attitudes are represented in memory, with consequences for the generalization of likes and dislikes. By taking on this cognitive-ecological perspective on the generalization of attitudes (Fiedler, 2014), we ascribe environmental conditions (here, stimulus variability in the environment) a key role in explaining evaluative learning and generalization. Understanding how environmental conditions relate to the generalization of likes and dislikes can contribute to our understanding of the acquisition of prejudice and stereotypes (Park & Hastie, 1987), and has implications for the design of interventions targeting attitude change (e.g., interventions to induce negative evaluations towards smoking; Măgurean et al., 2004; or negative evaluations towards unhealthy foods; Masterton et al., 2021; Bui & Fazio, 2016).

Variability and generalization

The relation between variability and generalization was documented in various domains of learning research, proposing that generalization is positively influenced by variability in training input (Apfelbaum & McMurray, 2011; Estes & Burke, 1953; Hahn et al., 2005; Raviv et al., 2022). For example, in category learning Posner and Keele (1968) reported an increase in the generalization of category knowledge after participants were exposed to variable rather than invariable training sets. In problem solving, the variability of worked examples increased the transfer of acquired skills to novel problems (Paas & Van Merriënboer, 1994). In concept learning, infants generalized a novel sound presented with animal categories only to unknown category exemplars after they experienced multiple (vs. single) animals per category (Vukatana et al., 2015). Similar outcomes were reported in research on language acquisition (e.g., speaker variability; Rost & McMurray, 2009), and inductive reasoning (e.g., premise diversity; Osherson et al., 1990). The studies highlight the relationship between variability of training exemplars and generalization at test. Importantly, because manipulations of variability produced similar results across domains of learning research, the underlying principles seem to be comparable (Raviv et al., 2022). Various accounts exist that attempt to explain the relation between variability and generalization.

One account suggests that variability fosters the formation of abstract representations during learning and thereby increases generalization (Apfelbaum & McMurray, 2011). Abstraction, in general, refers to the "process of identifying a set of invariant characteristics of a thing" (Burgoon et al., 2013; p. 502). Thus, abstract representations retain only those features that are relevant for a learning outcome, while irrelevant ones are ignored (see also Ramscar et al., 2010; Reed, 2016). For example, a representation of several individuals in terms of their social group membership can be seen as abstract, as the representation highlights the common characteristics across individuals (e.g., fans of a soccer club wearing club merchandise). Because variability in training stimuli emphasizes invariant characteristics across training exemplars, it can facilitate abstraction. For example, variable training sets in reward learning help learners to identify the cues that are most predictive of a reward across instances. Later at test, learners can predict rewards based on the presence or absence of the cues in novel instances (Ramscar et al., 2010). At the same time, the formation of abstract, simplified representations has the drawback of diminishing memory for specific details. For example, abstract representations seem to make it harder for learners to distinguish between seen and unseen exemplars (Bowman & Zeithamova, 2020; Garagnani et al., 2021; Hahn et al., 2005; Tussing & Greene, 1999).

One way to determine the relevance of features of exemplars is via cue competition. Cue competition, in general, describes the process by which cues compete for relevance in prediction of a particular outcome (Hoppe et al., 2022; Miller et al., 1995; Ramscar et al., 2010; Rescorla, 1968; Siegel & Allan, 1996). Positive weights are formed for cues that produce little or no error for an outcome, while negative weights are acquired for cues that result in prediction errors (Ramscar, 2021). The overarching function of cue competition is that of reducing prediction errors, and hence uncertainty (Hohwy, 2020; Kiefer & Hohwy, 2019; Rescorla, 1968). Learning from variable stimuli allows cues to compete for relevance, which results in the cues that most reliably predict outcomes being emphasized. Put differently, variability improves the discrimination between cues in stimuli. By contrast, learning from stimuli that lack a rich cue structure hinders cue competition and thus also learning to discriminate between cues (Ramscar et al., 2010). Accordingly, this perspective explains increased generalization with higher variability in training via the formation of more abstract representations, with cue competition as the underlying principle.

An alternative to this account suggests that variability increases generalization via the number of exemplars that represent a concept. With an increasing number and diversity of training examples, the likelihood that a new stimulus resembles a known one increases as well (Bowman & Zeithamova, 2020; Hahn et al., 2005; Homa et al., 1981; Nosofsky, 1988; 2011). For example, diverse training stimuli in category learning offer learners the chance to draw broad inferences, because the training stimuli demonstrate the scope of the category (Homa et al., 1981; Nosofsky, 1988; 2011). Importantly, this broadness account proposes that variability increases the *number* and *diversity* of representations making up a concept, but not their *abstractness*. As a consequence, and in contrast to the abstraction account, memory for specific details of training items should not be affected by variability (Bowman & Zeithamova, 2020).

To summarize, both accounts try to explain how variability affects generalization by specifying the way knowledge is stored in memory, either as abstract entities or as multiple concrete knowledge representations.

Variability and the generalization of attitudes

In the domain of attitude acquisition, the variability of attitude objects encountered during learning could also constitute a central determinant for the generalization of likes and dislikes. As an example, the co-occurrence of different female faces with positive images might trigger the acquisition of an association between a cue (“female”) and valence (“positive”), fostering generalization at test (Hütter et al., 2014). To our knowledge, no prior research exists that investigated the link between variability and generalization in attitude acquisition directly.

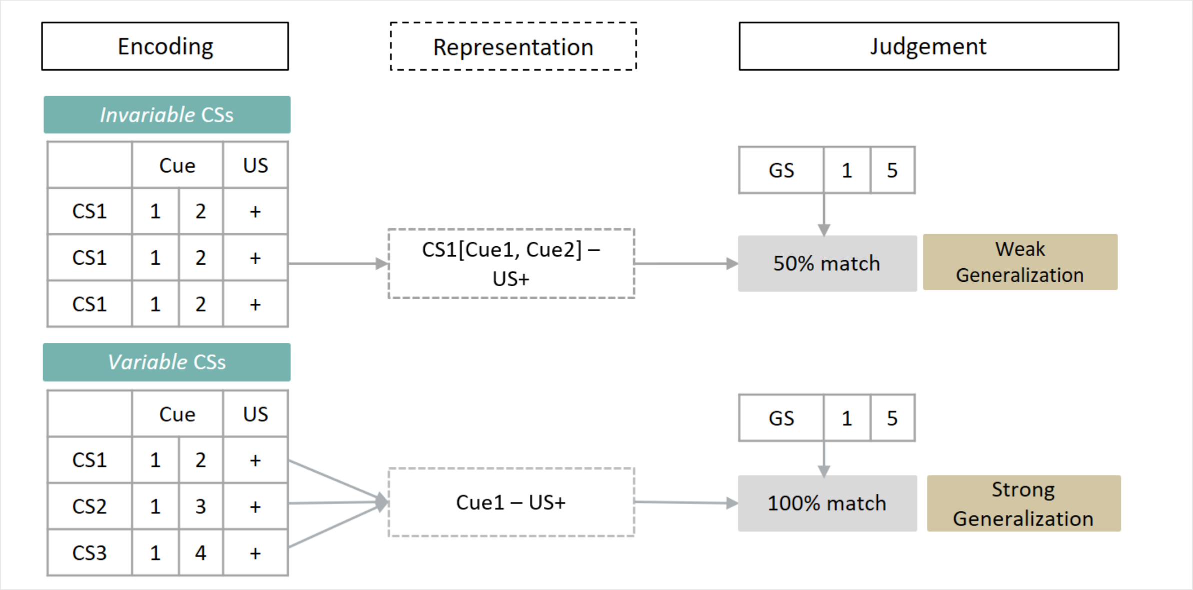
The present research aims to fill in this gap. We employed evaluative conditioning (EC) as an experimental procedure to study the acquisition and generalization of attitudes (EC; De Houwer et al., 2001; Hütter & Fiedler, 2016) and used recognition memory measures and evaluations of stimulus components to distinguish between the abstraction versus broadness accounts. In EC procedures, conditioned stimuli (CSs) that are neutral in valence occur in spatiotemporal contiguity with unconditioned stimuli (USs) of negative or positive valence. The evaluations of the CSs typically change in the direction of the US valence, which is also referred to as the *EC effect* (for reviews, see De Houwer et al., 2001; Hofmann et al., 2010; Walther et al., 2005). Importantly, attitudes acquired via EC can generalize to novel stimuli never seen during learning. For example, conditioning of a specific image of a person changes evaluations of modified displays of the person (e.g., the same person photographed from a different angle; Hütter & Tigges, 2019). Moreover, using category exemplars as CSs subsequently changed evaluations of similar stimuli and the whole stimulus category (Glaser & Kuchenbrandt, 2017; Jurchiş et al., 2020; Luck et al., 2020). While little is known about procedural aspects that might facilitate these generalization effects, the variability of attitude objects (CSs) encountered during conditioning might be a potential moderator.

In accordance with the “abstraction” account, high variability in CSs may result in abstract representations of CSs that only contain the cues that are most predictive of US valence across CSs. As a consequence, generalization towards novel instances increases. For example, imagine two distinct learning scenarios that vary in the variability of CSs. As displayed in *Figure 1*, the first scenario (upper panel, “invariable CSs”) entails only a single CS (CS1) that consists of two components (Cue 1 and Cue 2). This CS is repeatedly paired with a US of the same valence (e.g., a positive image, US+). In this learning environment, the resulting representation constitute a link between the concrete CS and US valence (CS1-US+ associations).

The second scenario (lower panel, “variable CSs”) consists of CSs that overlap in one component (Cue 1), but vary in their second component (Cue 2, Cue 3, or Cue 4). All CSs would again be paired with USs of the same valence (e.g., positive valence, US+). In this condition, only one CS cue predicts US valence across stimuli (i.e., Cue 1). As a consequence, an abstract representation might form that entails the most predictive cue (i.e., Cue 1), while disregarding less predictive ones (i.e., Cues 2, 3, 4).[3] In other words, the abstraction account predicts the formation of a link between the fixed cue and US valence (Cue1-US+ associations), at the cost of unique CS components.

Figure 1

An illustration of Variable versus Invariable Learning Conditions



	Cue		US
CS1	1	2	+
CS1	1	2	+
CS1	1	2	+

	Cue		US
CS1	1	2	+
CS2	1	3	+
CS3	1	4	+

Note. The upper panel ("*invariable CSs*") displays a conditioning procedure that repeatedly presents the same conditioned stimulus (CS) with a positive unconditioned stimulus (US+). The CS consists of two cues, Cue 1 and Cue 2. The resulting representation would consist of a link between the CS as a whole and the respective US. In turn, this representation constitutes a 50% match with a generalization stimulus (GS), resulting in weak generalization. The lower panel ("*variable CSs*") displays a conditioning procedure that includes CSs that are consistent in Cue 1 and vary in their second cue. In this context, the learner should abstract away from the varying cue and encode a link between Cue 1 and the respective US+. Here, the match between the representation and the GS amounts to 100%, which should result in strong generalization.

Importantly, the two learning scenarios depicted in *Figure 1* should have consequences for the evaluation of novel, generalization stimuli (GS). Considering that generalization is generally driven by the perceptual overlap between a knowledge representation and a novel stimulus (Shepard, 1987), it becomes evident that the perceptual overlap is higher after the variable than the invariable conditioning procedure. The perceptual overlap between a GS that consists of the familiar Cue 1, and a novel Cue 5, and an abstract representation (containing only Cue 1) amounts to 100%, because Cue 1 is present both in the representation and in the generalization stimulus. On the other hand, the overlap between a specific CS representation (containing both familiar Cues 1 and 2) and the novel stimulus

[3] This prediction can also be expressed in quantitative terms by calculating associative strengths between cues and US valence with the Rescorla-Wagner (RW) model (Rescorla & Wagner, 1972). The RW model predicts associative values of equal size for each CS cue for invariable CSs, and higher associative values for Cue 1 relative to the other Cues for variable CSs. See Supplement A for a detailed description of the calculations.

amounts to only 50% – because only Cue 1 is present in the GS, but not Cue 2.[4] Thus, generalization should be stronger in the variable (versus invariable) condition.

Another possibility is that the various CSs encountered in the variable condition provide a broader basis for generalization due to their higher number, rather than abstractness. This would imply that learners represent all of the features of the individual CSs in the variable condition as well. Testing learners' memory for details, and their evaluations of single CS components provides one way to distinguish between the two accounts. First, learners should have greater difficulties distinguishing between old and new stimuli in the variable than the invariable condition if variability fosters abstraction in the representation of CSs. Second, evaluative judgments of the most predictive cues should be more extreme than evaluations of less predictive ones, if the resulting representation entails only the most predictive cue (e.g., Cue 1) rather than the CSs used for learning. If these patterns are not observed, it would follow that the broadness account provides a better explanation for an increase in generalization than the abstraction account.

The present study

In this article, we report three studies in detail that manipulated the variability of CSs included in an EC procedure. CSs resembled Chinese characters that could be grouped into four categories by one common component. In the invariable condition, one item per category served as CSs. In the variable condition, multiple items per category were employed as CSs. Novel characters from the categories served as generalization stimuli (GSs). Generalization was expected to be more pronounced in the variable compared to the invariable condition. All experiments included both a direct (visual rating scales) and indirect measure (affect misattribution procedure; Payne et al., 2005) of attitudes. Because indirect measures infer attitudes from performance on a behavioral measure, they are less prone to demand effects and social desirability biases in responding.

We also included two additional measures to test the content of acquired representations of CSs. First, we included a variant of the Deese/Roediger-McDermott paradigm (DRM; Roediger & McDermott, 1995) to test participants' recognition memory performance. Participants were expected to make more recognition errors following an EC procedure with variable rather than invariable CSs. Second, we included evaluations of

[4] One could argue that the overlap only amounts to 50% for abstract CS representations as well, because the novel stimulus consists of two cues and only one matches the representation. However, we assume that the two representations differ in the number of cues that need to be present for an *activation* of the representation: the specific representation requires two matching cues (Cue 1 and Cue 2), the abstract representation requires only one (Cue 1). Thus, we expect stronger activation of the underlying representation in the variable relative to the invariable condition, enhancing generalization towards the novel stimulus.

individual CS cues to test whether predictive cues are evaluated more extremely than less predictive ones for variable CSs. Evaluations should not differ between the two cues in the invariable condition.

The first experiment tested generalization alone, and Experiments 2 and 3 included the recognition memory measure and evaluations of individual CS components. Further, Experiment 2 controlled for the number of CSs used at test, and Experiment 3 held the total number of CSs included in the learning phase constant across the two learning procedures. We conducted one additional experiment that employed a similar experimental procedure as Experiment 2 but presented the dependent measure in a different sequence. We report this additional experiment in the supplemental material and as part of internal meta-analyses that integrate the findings from all experiments.

For all experiments, we report how we determined sample sizes, all data exclusions and all manipulations and measures employed. Pre-registrations (for Experiments 2 and 3), data files, analysis scripts and stimulus material are publicly available on OSF via https://osf.io/tafy9/?view_only=1ef497132b5d456a8f5ec940911bfca9. The studies were approved by the ethics committee for psychological research at the authors' institution.

Experiment 1

This experiment sought to investigate our initial hypothesis that high variability in CSs during learning increases the generalization of likes and dislikes. That is, we tested whether the presentation of variable CSs increases the generalization of evaluations towards novel stimuli, relative to a condition that presents one specific CS repeatedly. The first experiment was not pre-registered.

Method

Participants

Data collection was conducted online. Participants received a study link via the university mailing list. They could sign up for a raffle (10 x 25€ vouchers for a local book store) as a reward for their participation. Participation took about 15 minutes. The study link expired after 14 days and all data sets collected until then were included in the data analysis. Of originally 238 participants, 38 were excluded because they spoke Chinese or reported that they had not paid attention during the learning phase. This resulted in a sample of 200 university students (149 female, 48 male, 1 diverse, 2 no response) of different majors, aged between 18 and 69 years (*M* = 24.14, *SD* = 6.51). The sample size provided an 80% power to detect a standardized beta coefficient of $\beta = 0.29$ or greater (two-sided t-test against zero) for the three-way interaction of US valence, stimulus type, and CS variability on direct evaluative ratings (which reflects our main hypothesis about the effect of CS variability on generalization), with a 5% false-positive rate (simulation-based approach with *simr* in R; Green & MacLeod, 2016).

Design

The study employed a 2 (US valence: negative vs. positive) × 2 (stimulus type: CS vs. GS) × 2 (CS variability: invariable vs. variable) mixed design with repeated measures on the first two factors.

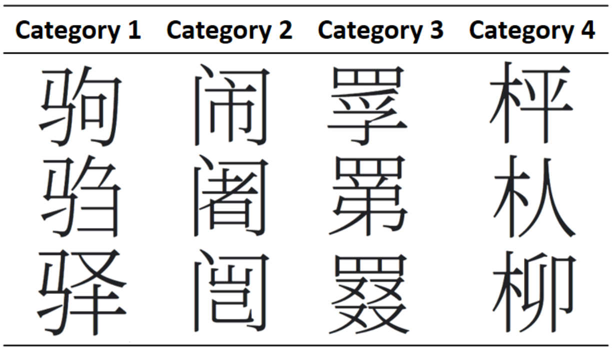
Materials

We used 40 pleasant and 40 unpleasant pictures from the International Affective Picture System (IAPS; Lang et al., 1997) as USs. Pleasant versus unpleasant pictures differed significantly in valence, $t(78) = -48.74$, $p < .001$, but not in arousal, $t(78) = 0.61$, $p = .545$.

A selection of stimuli akin to Chinese characters served as CSs in this and all subsequent experiments (see *Figure 2*). The characters were composed of two components and could be classified into four categories. One component repeated across characters of a category (*fixed* component). The second component was unique for each character (*varying* component). Each category consisted of ten characters, and CSs were randomly chosen from this stimulus pool. In the variable condition, five characters were selected per category as CSs. In the invariable condition, only one character was selected per category for the learning phase. Additionally, three characters were selected from each category to serve as GSs in the testing phase.

Figure 2

Examples of Conditioned and Generalization Stimuli



Category 1	Category 2	Category 3	Category 4
驹	闹	罦	枰
骀	阇	罤	朳
驿	闿	罬	柳

Note. CSs resembled Chinese characters. Every character consists of one component that repeated across characters of a category (*fixed* component), and one component that varied between characters (*varying* component). Characters of two categories were assigned to positive US valence, and characters of the other two categories to negative US valence. The figure illustrates only 3 of the 10 stimuli per category. Three generalization stimuli were selected from each category that did not serve as CSs during conditioning.

Procedure

All experiments were programmed in jsPsych (De Leeuw, 2015). Participants first went through the conditioning phase, and then completed the evaluative measures (direct ratings and the Affect Misattribution Procedure [AMP]).

Conditioning phase. CS categories were randomly assigned to positive and negative US valence with the restriction that two categories were paired positively and two were paired negatively. For each CS, a US image was randomly selected from the image pool of the respective valence. Thus, every CS was paired with a unique US. Every CS-US pair was presented five times per learning block in the *invariable* condition, and once per learning block in the *variable* condition. With four learning blocks à 20 trials, participants went through a total number of 80 learning trials in both conditions. After each learning block, they had the chance to take a break from the presentation. Each CS-US pair was presented simultaneously on the screen for 2000ms, with an inter-trial-interval (ITI) of 500ms. The order of pairs was randomized within blocks, and left-right assignment of CSs and USs was also randomly determined on a trial-by-trail basis. Prior to the learning phase, participants were informed that they would see a sequence of stimuli that they should simply attend to (see Supplement C for the complete task instructions).

Evaluative measures. After the learning phase, participants rated all CSs and GSs on a rating scale from -100 (unpleasant) to 100 (pleasant) and completed the AMP (Payne & Lundberg, 2014; Payne et al., 2005). The sequence of direct and indirect evaluative measures was counterbalanced between participants. Every CS and GS appeared once in the rating task and once in the AMP (resulting in 16 trials in the invariable condition, and 32 trials in the variable condition per measure). Stimuli from the CS and GS categories were presented in an interspersed manner in both measures.

In the AMP, CSs and GSs were used as primes and three-letter syllables from the Chinese language were used as targets (e.g., “tao”, “sha”). Participants were instructed to guess and press the letter “e” (unpleasant) or “k” (pleasant) if the syllable could mean something pleasant or unpleasant in Chinese, respectively.[5] Participants were also informed that characters and syllables are randomly presented together (see Supplement C). On every trial, primes were displayed for 90ms, followed by a 125ms inter-stimulus-interval and the presentation of the syllable for 125ms. Then, a grey mask appeared until participants made their response. Trials timed out after 4000ms. After an ITI of 125ms, the next trial started.

Results

Data analyses were conducted in R (R Core Team, 2023), version 4.1.1, using the packages tidyverse (Wickham et al., 2019), lme4 (Bates et al., 2015), and ggplot2 (Wickham & Chang, 2014). Supplement D includes the full statistics of every fixed and random parameter coefficient for the calculated models for this and the subsequent experiments.

Direct evaluative ratings

Results of the continuous evaluative ratings are displayed in *Figure 3* as a function of CS variability (variable vs. invariable), US valence (positive vs. negative), and stimulus type (CS vs. GS). The ratings were submitted to a linear mixed-effects model that accounted for inter-individual differences in responding. This form of data analysis has been shown to be especially useful in research on generalization effects (J. C. Lee et al., 2021; Vanbrabant et al., 2015). The model included the factors US valence, stimulus type, and CS variability as fixed effects, and random by-subject intercepts and slopes for the level 1 variables, US

[5] Participants are usually asked to evaluate the targets for pleasantness in the AMP (Payne et al., 2005; Payne et al., 2014). However, because the present study used syllables as targets, we asked participants to guess the meaning as it seems counterintuitive to evaluate unknown syllables on their pleasantness. Nevertheless, affective responses to primes should influence the (presumed) affective meaning of the targets.

valence and stimulus type. Fixed effects were effect coded (-0.5, 0.5).[6] To improve the interpretability of the results, and to provide scale-free indicators of effect sizes, we report standardized regression coefficients. They were calculated by fitting the same mixed model to z-standardized rather than raw evaluative ratings.

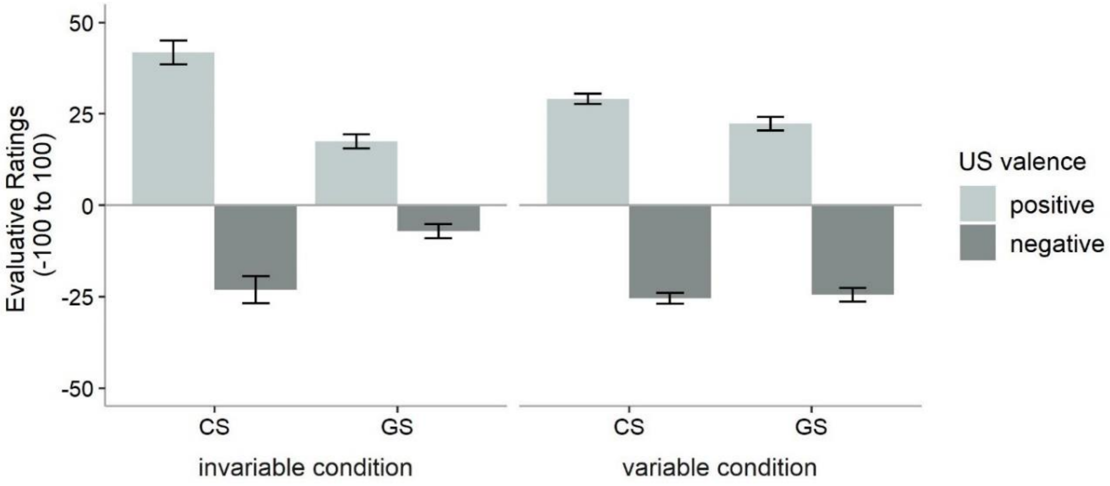
We observed a significant main effect of US valence, *B* = 47.70, *SE* = 3.16, *t*(217.12) = 15.10, *p* < .001, β = 0.91, demonstrating an overall EC effect. Evaluative ratings were on average 47.70 points (on the scale ranging from -100 to 100) higher for CS+ and GS+ compared to CS- and GS-. The EC effect was reduced for GSs in comparison to CSs, as revealed by a significant two-way interaction of US valence and stimulus type with a negative parameter estimate, *B* = -24.03, *SE* = 2.72, *t*(4388.62) = -8.82, *p* < .001, β = -0.46. Importantly, the three-way interaction of US valence, stimulus type, and CS variability was significant, *B* = 32.55, *SE* = 5.45, *t*(4388.62) = 5.97, *p* < .001, β = 0.62, indicating that the EC effect differed between learning conditions (variable vs. invariable) depending on stimulus type.

To resolve the three-way interaction, we calculated separate interaction effects (US valence × CS variability) for each type of stimulus, by setting the stimulus type of interest to zero and re-estimating the model. For CSs, the two-way interaction between US valence and CS variability did not reach significance, *B* = -10.39, *SE* =7.18, *t*(360) = -1.45, *p* = .148, β = -0.20. Thus, CS variability did not influence the size of the EC effect for CSs so that levels of evaluative learning were largely equated between conditions. For GSs, the interaction effect was significant, *B* = 22.16, *SE* = 6.57, *t*(253.47) = 3.37, *p* < .001, β = 0.43, and in the expected direction: generalization was stronger in the variable than the invariable condition. To further assess whether this was the case for both GS+ and GS-, we calculated simple slopes for the effect of CS variability per US valence, for GSs only. For GS+, the difference between the variable and invariable condition was not significant, *B* = 4.78, *SE* = 3.75, *t*(241.63) = 1.27, *p* = .204, β = 0.09. For GS-, generalization was more pronounced in the variable than invariable condition, *B* = -17.39, *SE* = 4.23, *t*(231.70) = -4.11, *p* < .001, β = -0.33. The results indicate that the differences in generalization between learning conditions were mainly driven by negative US pairings, suggesting valence-specific influences of variability on generalization.

[6] The model was specified as lmer(ratings ~ USvalence * stimulus type * CS variability + (USvalence + stimulus type | subject) in R, using the lme4 package. Effect coding: US valence (-0.5 negative, 0.5 positive), stimulus type (-0.5 CS, 0.5 GS), CS variability (-0.5 invariable, 0.5 variable).

Figure 3

Mean Evaluative Ratings in Experiment 1



Note. CS = Conditioned Stimuli, GS = Generalization Stimuli. Error bars display standard errors.

AMP

A significant main effect of US valence indicated an overall EC effect also on this measure, *B* = 2.17, *SE* = 0.11, *z* = 7.36, *p* < .001, demonstrating that the odds for answering "pleasant" were on average 2.17 times higher for CS+ and GS+ than CS- and GS-. The EC effect was reduced for GSs in comparison to CSs, *B* = 0.76, *SE* = 0.15, *z* = -1.87, *p* = .061, but this effect failed to reach significance. The parameter estimate for the three-way interaction of valence, CS variability, and stimulus type was not significant, *B* = 1.33, *SE* = 0.29, *z* = 0.97, *p* = .330, even though the differences were in the expected direction.

Simple slopes were calculated for the US valence × CS variability interaction for each type of stimulus separately. For CSs, the EC effect was not significantly qualified by CS variability, *B* = 1.21, *SE* = 0.28, *z* = 0.70, *p* = .486. However, the strength of generalization depended on CS variability, *B* = 1.61, *SE* = 0.23, *z* = 2.07, *p* = .039. Thus, and in line with the findings for the direct evaluative measures, the difference in odds to respond "pleasant" between positive and negative valence was larger in the variable than the invariable condition. Again, the difference between the learning conditions was significant only for GS-, *B* = 0.62, *SE* = 0.15, *z* = -3.24, *p* = .001, but not for GS+, *B* = 0.99, *SE* = 0.15, *z* = -0.04, *p* = .968.

Discussion

The main goal of Experiment 1 was to examine whether the variability of CSs in EC influences the generalization of evaluations via abstraction. Results of the direct evaluative ratings demonstrated stronger generalization of evaluations when many CSs per category were included (*variable* condition), relative to a learning procedure that included only one CS per category (*invariable* condition). At the same time, the size of the EC effect in the CSs did not depend on the learning condition. This indicates that variability in CSs mainly exerted an influence on generalization rather than on the strength of evaluative learning per se. Moreover, we observed this effect for negative, rather than positive pairings. The same tendency was observable for the AMP, even though the three-way interaction did not reach significance. Tentatively, this result indicates that the differences in generalization were not only due to demand effects.

The conclusions are limited to the extent that the number of presented CSs per category during learning was confounded with the number of CSs presented at test in this experiment. Because all CSs that occurred during learning were also presented during testing, participants evaluated five CSs per category in the variable condition, and only one CS per category in the invariable condition. To ensure that any differences in generalization were due to differential processing at the encoding stage (learning) and not due to differential processing at the retrieval and judgment (testing) stage, only one CS per category was included in evaluations for both CS variability conditions in Experiment 2.

Experiment 2

Experiment 2 was conducted to examine whether the results of Experiment 1 replicate when the confound noted above were controlled for. In Experiment 2, the number of CSs that were evaluated after the learning phase was held constant across CS variability conditions. The experiment also included two additional measures to test whether participants formed more abstract representations of CSs in the variable condition: First, participants completed a recognition memory task immediately after the learning phase. They were expected to have greater difficulties distinguishing between "old" stimuli (CSs) and "new" stimuli (GSs and distractors) in the condition including many CSs per category, relative to the condition including only one CS per category. This would support the assumption that abstraction takes place as a result of a process that omits irrelevant details. Second, participants also evaluated CS components. In the invariable condition, participants should evaluate CS components in about equal terms if they formed a concrete representation that does not distinguish between stimulus features. In the variable condition, more extreme evaluations of the fixed CS components were expected relative varying components, indicating more abstract representations that emphasize the CS features most predictive of US valence. The pre-registration of the experiment can be accessed via https://osf.io/g7nkw?view_only=d7180cb703c343279ccc8fcdedd51780.

Method

Participants

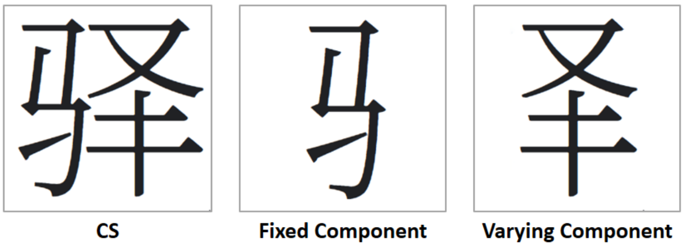
The anticipated sample size in Experiment 2 was set to *N* = 132 participants (based on the effect found in Experiment 1 for the three-way interaction of US valence × stimulus type × CS variability on evaluative ratings, *B* = 32.55, *SE* = 5.45, and to achieve a power greater than .95, plus an additional 20% to account for data exclusions). The study was conducted online, and participants were recruited via Prolific (www.prolific.co). The participant pool was restricted to participants from Germany, with German as a first or fluent language. After excluding three participants who did not pay attention during the learning phase (self-reported), a total of *N* = 129 data sets were included in the analysis. With this sample size, the experiment provided an 80% power to detect a standardized beta coefficient of β = 0.50 or greater (for the three-way interaction of US valence, stimulus type and CS variability on direct evaluations, with an alpha-level of .05). Participants were between 18 and 69 years old (*M* = 30.62, *SD* = 10.39). 65 participants were female, 61 male, and 2 diverse (one participant did not respond to this question). In total, the study took about 15 minutes. Participants received 2.00 GBP.

Materials and procedure

The study materials and procedure were largely the same as in Experiment 1, except for a few exceptions. First, we included only one CS per category for both learning conditions in the testing phase. Thus, in the invariable condition, one CS was randomly selected out of the five CSs from the learning phase to be included in the test phase. As a result, in this experiment the number of CSs and GSs (total: 16) in the evaluative rating task and AMP was equated between CS variability conditions. Moreover, participants also evaluated the two components of the CSs on a direct evaluative rating scale (-100 to 100) in isolation. One component was the category-defining component that stayed the same across characters of a category (*fixed* component) and the other varied between characters of a category and was thus unique for each character (*varying* component). See *Figure 4* for an example of fixed and varying components of a CS. Because 4 CSs were included in the test phase, participants evaluated 4 fixed, and 4 varying components of the respective stimuli in both CS variability conditions. Participants evaluated the CS components at the end of the experiment.

Figure 4

Examples of Fixed and Varying Components that were Evaluated at the End of Experiments 2 and 3



Note. CSs consisted of two components that were evaluated in isolation at the end of Experiments 2 and 3. The *fixed component* occurred for all exemplars of a category. The *varying component* occurred for only one exemplar of a category and thus varied across category exemplars.

In Experiment 2, participants also completed a recognition memory task. Here, the four CSs, 12 GSs, and four additional distractors (Chinese characters not related to any categories from the conditioning phase) were included as stimuli. Participants classified each stimulus as either “new” or “old” by pressing the left or right arrow key, respectively. They were instructed to respond “new” if they did not recognize the stimulus from the learning phase, and to respond “old” when they did. They were asked to respond quickly and accurately (see Supplement C for the complete task instructions). Trials timed out after four seconds. Stimuli were presented in the center of the screen, and four practice trials were included at the beginning of the experiment with completely novel characters as stimuli.

Before this experiment, we conducted an additional experiment that employed a similar experimental procedure but presented the recognition memory task at the end of the experiment. That posed the problem that some participants reported confusion over the task instructions. Namely, it was unclear whether “old” responses referred to stimuli presented during learning (CSs only, as anticipated), or stimuli presented during learning or testing (CSs and GSs, because GSs occurred when evaluating the stimuli). Therefore, recognition memory performance was measured directly after the conditioning phase in Experiment 2. We report the results of the additional experiment in Supplement B and as part of the internal meta-analyses.

Results

Direct evaluative ratings

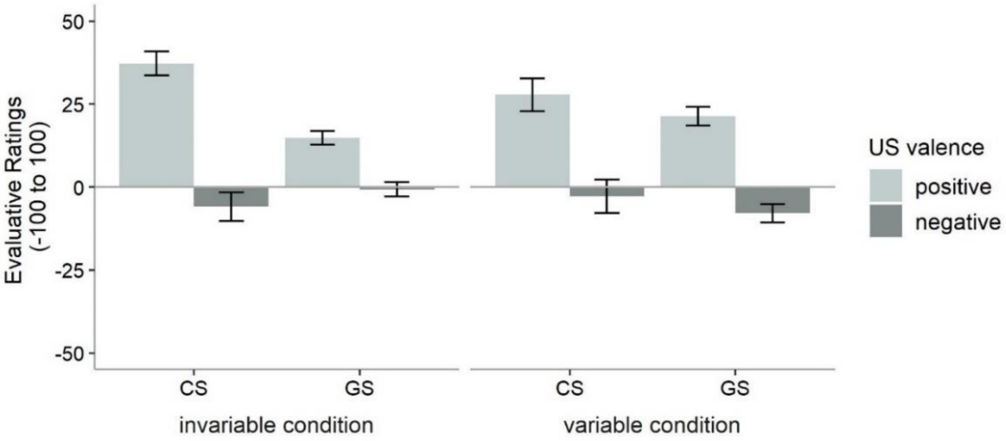
Results of the direct evaluative ratings of CSs and GSs are displayed in *Figure 5*. Evaluative ratings were submitted to the same linear-mixed effect model as specified for Experiment 1. The model included US valence, CS variability, and stimulus type as fixed effects as well as their interactions, and random by-subject intercepts and slopes for US valence and stimulus type. The overall EC effect was significant, *B* = 29.65, *SE* = 4.09, *t*(148.61) = 7.24, *p* < .001, β = 0.60, and reduced for GSs as compared to CSs, *B* = -14.53, *SE* = 4.28, *t*(1806) = -3.39, *p* < .001, β = -0.29. The three-way interaction of US valence, CS variability, and stimulus type was significant, *B* = 26.23, *SE* = 8.57, *t*(1806) = 3.06, *p* = .002, β = 0.53.

To decompose the three-way interaction, we assessed the interaction of US valence and CS variability separately for each type of stimulus. The two-way interaction did not reach significance for CSs, *B* = -12.57, *SE* = 10.19, *t*(345) = -1.23, *p* = .218, β = -0.25. Although not significant, the negative parameter estimate indicated a larger EC effect in the invariable relative to the variable condition. For GSs, this relation was reversed, *B* = 13.66, *SE* = 8.19, *t*(148.61) = 1.67, *p* = .097, β = 0.28. Here, the EC effect was on average larger in the variable than the invariable condition (indicating greater generalization for variable than invariable CSs), even though this difference was again not significant. A test of valence-specific effects of variability on generalization revealed a non-significant difference between the variable and invariable condition for GS+, *B* = 6.50, *SE* = 5.00, *t*(142.18) = 1.30, *p* = .196, β = 0.13, and for GS-, *B* = -7.16, *SE* = 5.53, *t*(140.05) = -1.30, *p* = .197, β = -0.14. Thus, generalization effects were present in both variability conditions, independent of the specific US valence.

We also tested valence-specific effects of generalization within each variability condition. Evaluations decreased for GSs compared to CSs for positive valence in the invariable condition, *B* = -22.38, *SE* = 3.81, *t*(1342.22) = -5.87, *p* < .001, β = -0.45, but not in the variable condition, *B* = -6.46, *SE* = 4.79, *t*(1342.22) = -1.35, *p* = .178, β = -0.13. Thus, positive evaluations were generalized more after exposure to variable than invariable CSs. For negative US valence, differences between CSs and GSs were non-significant for both the invariable condition, *B* = 5.27, *SE* = 3.81, *t*(1342.22) = 1.38, *p* = .168, β = 0.11, and the variable condition, *B* = -5.05, *SE* = 4.79, *t*(1342.22) = -1.05, *p* = .292, β = -0.10. Thus, while generalization effects did not differ significantly between CS variability conditions, simple slopes revealed that evaluations were less extreme for GSs than CSs for positive valence in the invariable condition, while there were equally extreme in the variable condition (see *Figure 5*).

Figure 5

Mean Evaluative Ratings in Experiment 2



Note. CS = Conditioned Stimuli, GS = Generalization Stimuli. Error bars display standard errors.

AMP

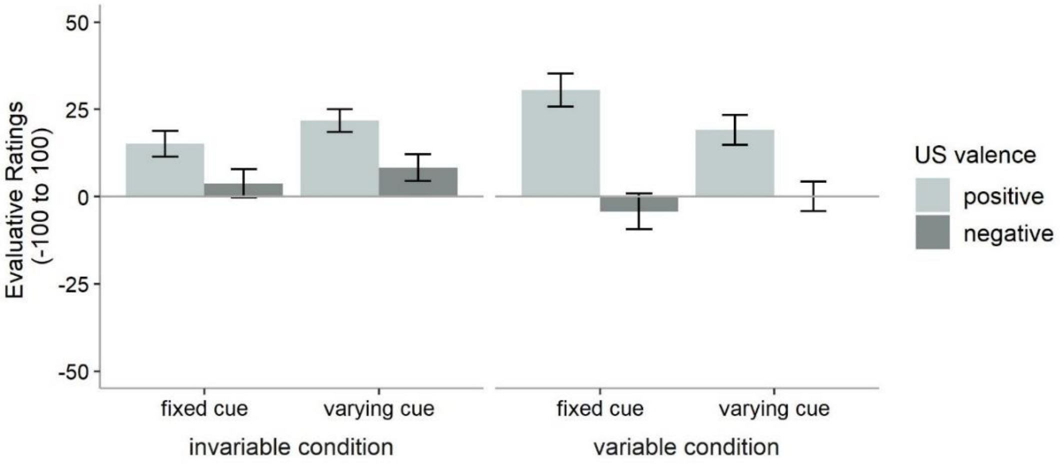
A generalized linear mixed-effect model on “pleasant” (1) versus “unpleasant” (0) responses obtained in the AMP (0.15% of timed out trails were excluded) revealed a significant overall EC effect, *B* = 2.07, *SE* = 0.12, *z* = 5.95, *p* < .001. The odds to respond “pleasant” were higher for positive than negative valence. The EC effect was qualified by stimulus type, *B* = 0.59, *SE* = 0.22, *z* = -2.41, *p* = .016, indicating a smaller EC effect for GSs than CSs. The three-way interaction of US valence, CS variability and stimulus type was not significant, *B* = 1.68, *SE* = 0.44, *z* = 1.19, *p* = .235.

Evaluative ratings of stimulus components

Evaluative ratings obtained for the CS components are displayed in *Figure 6*. The ratings were submitted to a linear mixed-effect model with the same model structure as described for the analysis of CS and US ratings, but in this model the factor component type (fixed vs. varying) replaced the factor stimulus type.[7] On average, CS components were evaluated more positively when they were part of a CS previously paired with positive rather than negative valence, *B* = 19.69, *SE* = 3.91, *t*(129.01) = 5.03, *p* < .001, β = 0.41. The impact of US valence on evaluative ratings did not depend on the type of component and CS variability, *B* = -17.87, *SE* = 10.79, *t*(774) = -1.66, *p* = .098, β = -0.38. However, because the effect was in the expected direction, we further analyzed the three-way interaction. Simple slopes showed that the size of the EC effect was larger for the fixed component than the varying component in the variable condition, *B* = -15.76, *SE* = 8.44, *t*(774) = -1.87, *p* = .062, β = -0.33, but did not differ between components in the invariable condition, *B* = 2.11, *SE* = 6.72, *t*(774) = 0.32, *p* = .753, β = 0.04. In the variable condition, the difference between components was primarily driven by positive US valence, *B* = -11.44, *SE* = 5.99, *t*(732.53) = -1.91, *p* = .056, β = -0.24, rather than negative US valence, *B* = 4.32, *SE* = 5.99, *t*(732.53) = 0.72, *p* = .471, β = 0.09, even though both effects were non-significant.

Figure 6

Mean Evaluative Ratings of CS Components in Experiment 2



Note. Fixed cue = CS component fixed across characters of a category, varying cue = CS component varying between characters. Error bars display standard errors.

[7] The model was specified as lmer(ratings ~ US valence * component type * CS variability + (US valence + component type| subject)) in R, using the lme4 package. Effect coding: US valence (-0.5 negative, 0.5 positive), component type (-0.5 fixed, 0.5 varying), CS variability (-0.5 invariable, 0.5 variable).

Recognition memory task

Table 1 provides an overview of the proportions (and standard deviations) of "old" responses in the invariable versus variable condition, separately for CSs (correct responses), GSs (false alarms) and distractors (false alarms). "Old" and "new" responses were used to calculate the signal detection measure *d'* for every participant. Four practice trails, responses to distractors, and trials that timed-out (0.33% of trials) were excluded from the calculations. *d'* reflects individual participant's sensitivity to distinguish between old (CSs) and new (GSs) items and is generated by subtracting aggregated and standardized "old" responses for GSs from those of CSs (Macmillan & Creelman, 2004).[8] The index *d'* takes the general tendency to respond "old" rather than "new" into account and thus reflects sensitivity rather than response biases. When an individual cannot discriminate between the old and new stimuli at all, *d'* equals 0. Increasing *d'* values indicate increasing performance in distinguishing between CSs and GSs.

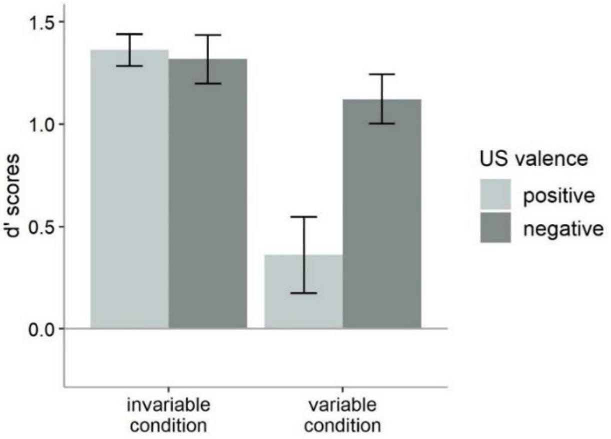
Figure 7 presents the averaged *d'* values for each US valence and CS variability condition. Overall, *d'* values were higher in the invariable condition (M = 1.34, SD = 0.89) than the variable condition (M = 0.74, SD = 1.17). This difference was significant, B = -0.60, SE = 0.13, t(128.99) = -4.75, p < .001.[9] The parameter estimate indicated an average decrease of *d'* of 0.6 in the variable condition, compared to the invariable condition. This effect depended on US valence, B = -0.81, SE = 0.24, t(128.99) = -3.30, p = .001. For positive valence, *d'* was larger in the variable than the invariable condition, B = -0.92, SE = 0.18, t(257.82) = -5.23, p < .001, while this difference was not significant for negative US valence, B = -0.28, SE = 0.18, t(257.82) = -1.57, p = .118. An analysis of raw "old" versus "new" responses indicated a similar result pattern (*Table 1*). See Supplement E for a more detailed description of this analysis.

[8] *d'* = z ("old" responses to CSs) – z ("old" responses to GSs). Z-scores of 0 and 1 were substituted by 1/(2/N) values and 1-1/(2N) values, respectively, to avoid infinite *d'* values (Macmillan & Creelman, 2004). N denotes the number of trials on which the proportion is based. We do not report standardized parameter coefficients as *d'* values rely on standardized input values.

[9] Model specified as lmer (d' ~ CS variability * US valence+ (1|subject)) in R, using the lme4 package. Effect coding: CS variability (-0.5 invariable, 0.5 variable), US valence (-0.5 negative, 0.5 positive).

Figure 7

Mean d' scores in Experiment 2



Note. Error bars denote standard errors.

Table 1

Proportion (standard deviations) of 'old' responses in the invariable versus variable condition for CSs, GSs, and distractors across experiments, with simple slopes testing for the difference between CS variability conditions

	'old' responses	'old' responses				
	Invariable	*Variable*	*OR*	*SE*	*z*	*p*
Exp. 2						
CS	0.82 (0.20)	0.73 (0.19)	0.62	0.25	-1.93	.054
GS	0.21 (0.18)	0.44 (0.19)	3.31	0.16	7.35	< .001
Distractor	0.06 (0.18)	0.14 (0.20)	2.58	0.34	2.80	.005
Exp. 3						
CS	0.82 (0.24)	0.68 (0.25)	0.43	0.22	-3.80	< .001
GS	0.32 (0.22)	0.44 (0.21)	1.77	0.14	3.99	< .001
Distractor	0.18 (0.24)	0.12 (0.19)	0.61	0.25	-1.97	.049

Note. Standard deviations are presented in parentheses. 'Old' responses to CSs are correct responses, while 'old' responses to GSs and distractors are false alarms. Simple slopes were calculated on 'new' (0) and 'old' (1) responses (with a generalized linear mixed model). Odds ratios (OR) indicate the odds for responding 'old' in the variable relative to the invariable condition. See Supplement E for a full description of the outcomes.

Discussion

The results of Experiment 2 were consistent with Experiment 1 with regards to the evaluations of CSs and GSs. The present experiment thereby substantiated the notion that the generalization of evaluative responses towards novel stimuli of the categories is stronger when there is variability among CSs. As opposed to Experiment 1, differences in generalization effects were mainly driven by positively paired stimuli.

Moreover, evaluations of the CS components further supported our assumption that the levels of abstraction at which participants represented the CSs differed between the CS variability conditions. In the invariable condition, there was no difference in the evaluation of the two types of CS components. In the variable condition, the EC effect was larger for the CS component fixed across CSs than the component varying between CSs. Lastly, participants' ability to discriminate between CSs and GSs was diminished in the variable condition as compared to the invariable condition for positively paired stimuli. In line with our prediction that participants did not simply store more exemplars in the variable condition, but rather abstracted away from varying CS components, participants in the variable condition had greater difficulties to distinguish "old" and "new" stimuli.

Experiment 3

In all of the previous experiments, the number of CSs per category that were included in the conditioning phase was confounded with the total number of CSs. That is, participants in the variable condition saw 20 CSs together with the USs, while participants in the invariable condition saw only 4 CSs. Thus, a potential alternative explanation for the results of the previous experiments might posit that the total number of CSs rather than the number of CSs per category was the crucial determinant of our findings. The results of the recognition memory task might be particularly affected by this possible confound, because the task becomes more difficult when one has to memorize a higher number of CSs in total. To rule out this alternative explanation, we conducted a third experiment that held the number of CSs constant across the two learning procedures via the use of filler stimuli. The pre-registration for the experiment is available via https://osf.io/dhf2s?view_only=ae98a2a9d47341edac0c716bc77c7051.

Method

Participants

To determine the sample size required for a replication of the effect of CS variability on the continuous evaluative ratings reported in Experiment 2 (three-way interaction of US valence, CS variability and stimulus type, B = 26.23, SE = 8.57) an a-priori power analysis was conducted using the mixed-effect model reported for Experiment 2. To achieve a power of .95 and account for 10% expected data exclusion, the anticipated sample size was set to N = 176.[10]

Participants were recruited via Prolific (www.prolific.co), and the sample was restricted to those who had not participated in Experiment 2, live in Germany, and speak German fluently. After excluding 10 (of originally N = 177) participants who reported that they had not paid attention during the learning phase, a total of N = 167 data sets were included in the data analysis. Participants (64 female, 98 male, 4 diverse, 1 no response) were between 18 and 71 years old (M = 30.28, SD = 9.83). This sample size provided an 80% power to detect a standardized beta coefficient of β = 0.42 or greater (for the three-way interaction of US valence, stimulus type and CS variability on direct evaluations, with an alpha-level of .05, and a model that was fitted to the data of Experiment 3). In total, the study took about 20 minutes, and participants received 2.55 GPB for participation.

[10] We based the a-priori power analysis on the results of Experiment 2 rather than Experiment 1 because Experiment 2 is procedurally closer to Experiment 3. We accounted for 10% data exclusion (instead of 20% as in Experiment 2) because only around 4% had to be excluded in Experiment 2.

Materials and procedure

The study design and materials were the same as Experiment 2, aside from one detail. To keep the total number of CSs constant across CS variability conditions, filler CSs were selected from a pool of Chinese characters that did not correspond in any of their parts to the characters of the categories. 20 filler CSs were selected for the invariable condition, and 4 filler CSs were selected for the variable condition. Filler CSs were paired with neutral USs, which were selected from the IAPS (Lang et al., 1997) and THINGS database (Hebart et al., 2019).

In total, the evaluative conditioning phase consisted of 160 learning trials in each of the learning conditions. In the invariable condition, each CS-US pair was presented twenty times when the CS belonged to one of the categories (resulting in 80 trials) and four times when the CS was a filler CS (resulting in another 80 trials). In the variable condition, each CS-US pair was presented four times when the CS belonged to one of the categories (resulting in 80 trails), and twenty times when the CS was a filler CS (resulting in another 80 trails). Keeping the number of learning trials constant also made sure that fixed CS components are presented equally often in both learning conditions.

Results

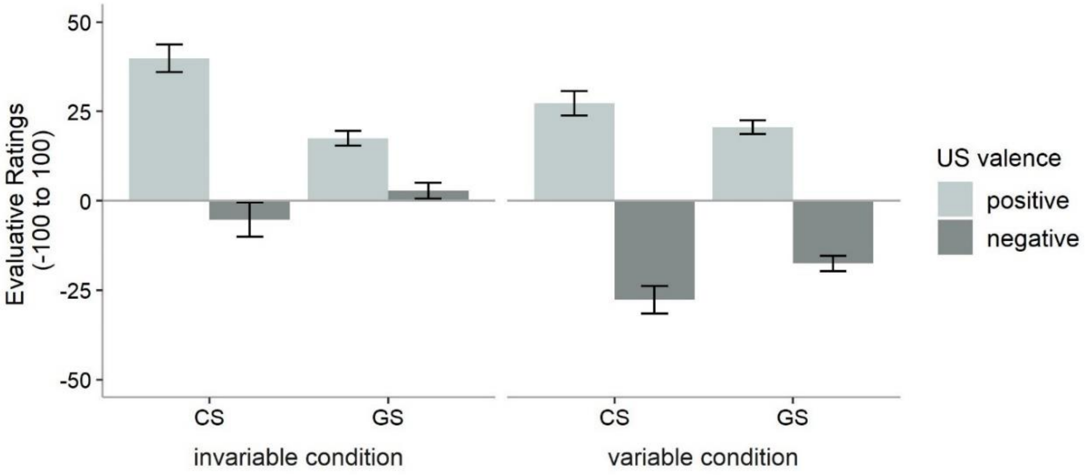
Direct evaluative ratings

Aggregated continuous evaluative ratings are displayed in *Figure 8*. On average, evaluative responses were more positive for stimuli linked to positive than negative US valence, yielding an overall EC effect, B = 38.22, SE = 3.65, t(191.11) = 10.48, p < .001, β = 0.74. The EC effect was reduced for GSs in comparison to CSs, B = -23.64, SE = 3.73, t(2337.98) = -6.34, p < .001, β = -0.46. The three-way interaction of US valence, CS variability, and stimulus type did not reach significance, B = 13.73, SE = 7.45, t(2337.98) = 1.84, p = .066, β = 0.27. However, because the effect was in the expected direction, we further analyzed the three-way interaction by calculating separate interaction effects (US valence × CS variability) for each type of stimulus. For CSs, the interaction did not reach significance, B = 9.78, SE = 9.00, t(430.99) = 1.09, p = .278, β = 0.19. For GSs, the interaction was significant, B = 23.52, SE = 7.29, t(191.11) = 3.23, p = .001, β = 0.46, with a larger EC effect for GSs in the variable than the invariable condition.

To test whether differences in generalization between CS variability conditions were specific for one type of US valence, we calculated simple slopes for CS variability per valence. For GS+, the difference between learning conditions was non-significant, B = 3.14, *SE* = 3.87, *t*(191.60) = 0.81, *p* = .418, β = 0.06. For GS-, ratings were more negative in the variable than the invariable condition, indicating stronger generalization, *B* = -20.37, *SE* = 5.03, *t*(180.23) = -4.05, *p* < .001, β = -0.40. Thus, consistent with the findings of Experiment 1, there was a valence-specific effect of variability on generalization. In particular, negative US valence resulted in higher generalization in the variable relative to the invariable condition.

Figure 8

Mean Evaluative Ratings in Experiment 3



Note. CS = Conditioned Stimuli, GS = Generalization Stimuli. Error bars display standard errors.

AMP

Responses of the AMP collected in Experiment 3 were analyzed using a generalized linear mixed-effect model (0.19% of all trials excluded because they timed out). The model revealed a significant overall EC effect, *B* = 1.90, *SE* = 0.12, *z* = 5.36, *p* < .001. The three-way interaction of US valence, stimulus type and CS variability did not reach significance, *B* = 1.72, *SE* = 0.38, *z* = 1.43, *p* = .154, and neither did any other parameter estimate (smallest *p* = .090).

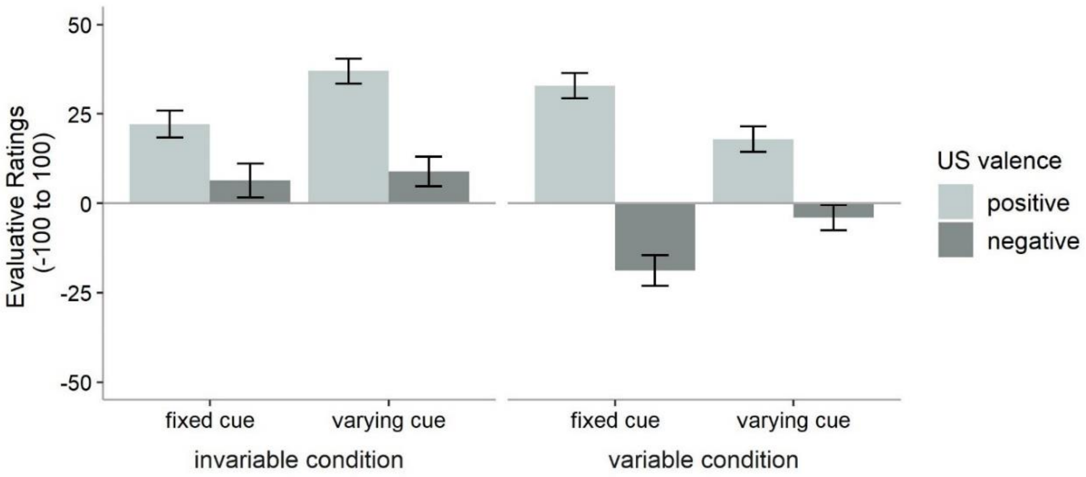
Evaluative ratings of stimulus components

Aggregated continuous evaluative ratings of the CS components are depicted in *Figure 9*. CS components of CSs paired with positive valence were evaluated more positively than CS components of CSs paired with negative valence, *B* = 29.33, *SE* = 3.23, *t*(171.93) = 9.08, *p* < .001, β = 0.55. The size of the EC effect depended on the type of CS component as well as the CS variability, as indicated by a significant three-way interaction, *B* = -42.05, *SE* = 10.13, *t*(1001.99) = -4.15, *p* < .001, β = -0.79. When calculating the two-way interaction of US valence and CS component separately for each CS variability condition, we obtained EC effects of equal size for both CS components in the invariable condition, *B* = 12.26, *SE* = 7.52, *t*(1001.99) = 1.63, *p* = .103, β = 0.23, and a larger EC effect for the fixed than the varying component in the variable condition, *B* = -29.79, *SE* = 6.79, *t*(1001.99) = -4.39, *p* < .001, β = -0.56.

Moreover, testing for valence-specific effects showed that within the invariable condition, evaluative ratings were more extreme for varying than fixed components for positive US valence, *B* = 14.84, *SE* = 5.40, *t*(863.31) = 2.75, *p* = .006, β = 0.28, while the difference was non-significant for negative US valence, *B* = 2.58, *SE* = 5.40, *t*(863.31) = 0.48, *p* = .633, β = 0.05. In the variable condition, the difference in evaluations of the CS components were in the expected direction: Ratings were more extreme for the fixed than the varying component for both positive, *B* = -14.96, *SE* = 4.88, *t*(863.31) = -3.07, *p* = .002, β = -0.28, and negative pairings, *B* = 14.83, *SE* = 4.88, *t*(863.31) = 3.04, *p* = .002, β = 0.28.

Figure 9

Mean Evaluative Ratings of CS Components in Experiment 3



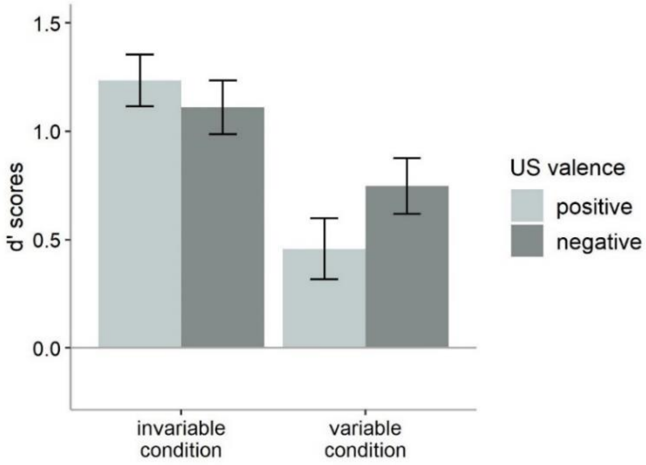
Note. Fixed cue = CS component fixed across characters of a category, varying cue = CS component varying between characters. Error bars display standard errors.

Recognition memory task

As described in Experiment 2, we calculated the signal detection measure *d'* to obtain a sensitivity index for the ability of participants to distinguish between CSs and GSs (*Figure 10).* Trials that timed out (0.15% of all trials) were excluded from data analysis. *d'* was reduced for the variable condition (*M* = 0.60, *SD* = 1.30), as compared to the invariable condition (*M* = 1.17, *SD* = 1.05). This difference was significant, *B* = -0.57, *SE* = 0.13, *t*(166.99) = -4.29, *p* < .001, and did not depend on US valence, *B* = -0.41, *SE* = 0.26, *t*(166.99) = -1.62, *p* = .108. The result indicates that participants were better able to discriminate between CSs and GSs when five rather than one CS per category were presented in the conditioning phase. A test for valence-specific effects revealed that this difference was only significant for positive US valence, *B* = -0.79, *SE* = 0.18, *t*(333.37) = -4.27, *p* < .001, but not for negative US valence, *B* = -0.36, *SE* = 0.18, *t*(333.37) = -1.93, *p* = .054. In a similar vein, an analysis of the raw “old” and “new” responses indicated higher false memory rates for GSs and distractors, and higher correct response rates for CSs (see *Table 1*).

Figure 10

Mean d’ scores in Experiment 3



Note. Error bars denote standard errors.

Discussion

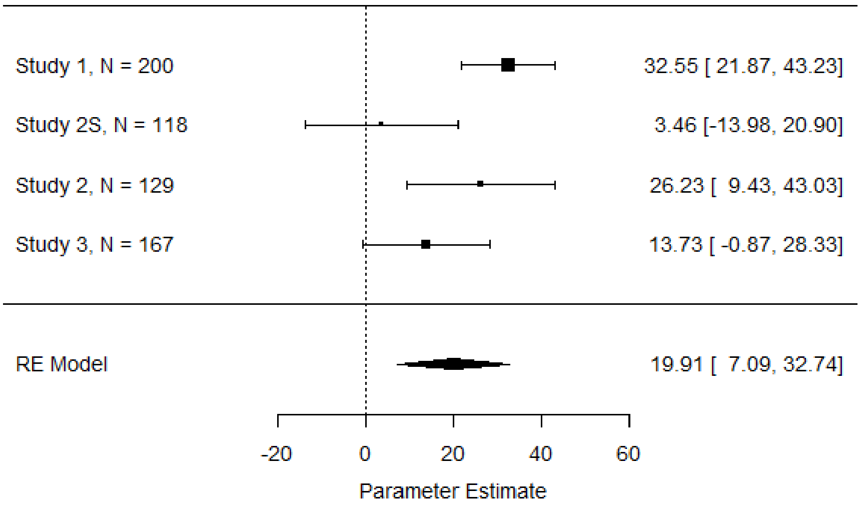
In contrast to the previous experiments, Experiment 3 held the total number of CSs constant across CS variability conditions. Importantly, we replicated the results of the previous experiments on the three different measures. The findings support the notion that the number of CSs per category rather than the total number of CSs is responsible for the reported effects in the previous experiments. It further strengthens the notion that manipulating CS variability does affect abstraction in learning, with consequences for the generalization of likes and dislikes towards novel category exemplars.

Internal Meta-Analysis

To assess the robustness of our findings in light of the varying sensitivity of the experiments to detect an effect, we conducted a maximum-likelihood random-effects meta-analysis using the R package *metafor* (Viechtbauer, 2010). The parameter coefficient for the three-way interaction of US valence, stimulus type and CS variability for direct evaluative ratings in Experiment 1 to 3, plus the additional experiment reported in the supplement ("Study 2S") was significant, *B* = 19.91, *95%CI* = [7.09, 32.74], *SE* = 6.54, *z* = 3.04, *p* = .002, β = .39. *Figure 11* presents the forest plot of this meta-analysis. For indirect evaluative ratings obtained with the AMP, the interaction of US valence, stimulus type, and CS variability did not reach significance (parameter coefficient indicates log-odds), *B* = 0.29, *95%CI* = [-0.07, 0.65], *SE* = 0.19, *z* = 1.56, *p* = .119. The result is depicted in *Figure 12*.

Figure 11

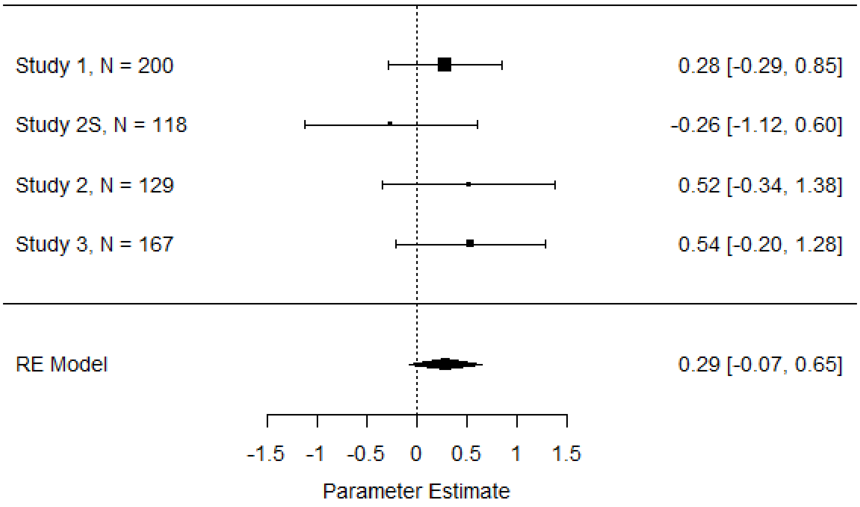
Internal Meta-Analysis on the Effect of CS variability on the Size of the Generalization-Effect for Direct Evaluative Ratings



Note. Internal meta-analysis conducted on the parameter coefficients obtained for the three-way interaction of US valence, stimulus type and CS variability for direct evaluative ratings (-100 to 100) across experiments. Study 2S is reported in the supplemental material.

Figure 12

Internal Meta-Analysis on the Effect of CS variability on the Size of the Generalization-Effect for Indirect Evaluative Ratings obtained in the Affect Misattribution Procedure (AMP)



Note. Internal meta-analysis on the parameter coefficients obtained for the three-way interaction of US valence, stimulus type, and CS variability for AMP results (0 = “unpleasant”, 1 = “pleasant”) across experiments. Study 2S is reported in the supplemental material.

General Discussion

Previously acquired attitudes are often generalized to make judgments and decisions about newly encountered individuals, objects, and situations. Previous research has studied the principles of generalization at the judgment stage. However, until now little has been known about the learning conditions that facilitate or inhibit generalization, even though this question is highly relevant from both a theoretical and an applied perspective. Other domains of learning research have identified variability in training objects as one factor improving the learner's ability to generalize acquired knowledge (Apfelbaum & McMurray, 2011; Estes & Burke, 1953; Hahn et al., 2005; Raviv et al., 2022). The aim of the present work was to examine this factor in the context of attitude acquisition and evaluate its influence on the representation and generalization of likes and dislikes.

We manipulated the variability of training objects via the number of CSs that were included per CS category in an evaluative conditioning procedure. Either one exemplar of a category (invariable condition) or multiple exemplars of a category (variable condition) were presented as CSs during learning. CSs resembled Chinese characters that consisted of two components: one that was fixed across characters of a category, and one that was unique for each character. As a measure of generalization, evaluative responses towards novel stimuli belonging to the same categories (GSs) were collected on both direct and indirect measures of attitudes. In addition, Experiments 2 and 3 included a recognition memory task and evaluations of CS components. Both measures provided further insights into the mechanisms underlying the link between variability and generalization.

The central finding of the studies was stronger generalization towards novel stimuli in the variable as opposed to the invariable condition on direct evaluative ratings. This pattern of results was observable across experiments, and the effect reached significance in an internal meta-analysis that also took an additional experiment conducted in our lab into account (reported in the supplemental materials) that presumably suffered from carry-over effects between tasks (see the method section of Experiment 2).

Evaluations collected on an indirect measure of evaluations (Affect Misattribution Procedure; Payne et al., 2005) did not yield significant differences in generalization between CS variability conditions, although descriptively the effects were in the expected direction. Whereas past work reported high reliability of the AMP (Cameron et al., 2012; Payne & Lundberg, 2014), the reliability of indirect measures is generally under dispute (Cunningham et al., 2001; Dessel et al., 2020; Fazio & Olson, 2003). Moreover, another study that used the AMP as a measure of generalization effects in EC also failed to find statistically significant effects (Spruyt et al., 2014). In the light of these considerations, we refrain from drawing strong conclusions based on the AMP results. Because the effect of variability on generalization was observable at least in descriptive terms, we take the findings as an indicator that the outcomes we observed on the direct measures were not purely driven by demand effects.

To summarize, generalization towards novel stimuli increased as a function of CS variability during learning on a direct evaluative measure (internal meta-analysis: β = .39), in line with current knowledge in other domains of learning demonstrating that variable input facilitates generalization (see Raviv et al., 2022 for a review). Our results thus add to the vast body of research studying the learning conditions that facilitate generalization (Bowman & Zeithamova, 2020; Gentner & Hoyos, 2017), and contribute to our knowledge on the specific procedural aspects that facilitate generalization in evaluative conditioning (Glaser & Kuchenbrandt, 2017; Hütter & Tigges, 2019).

Importantly, the results of the recognition memory task and evaluations of CS components were in line with the hypothesis that the relation between CS variability and generalization is mediated by abstraction. In the following, we first discuss how CS variability could influence abstraction in CS representations (i.e., we address the acquisition stage of evaluative learning that translates evaluative experiences into mental representations), and later turn to the effect of abstraction in CS representations on generalization (i.e., the retrieval stage of evaluative learning that applies mental representations to evaluative judgment; Hütter, 2022; Hütter & Rothermund, 2020).

Acquisition stage: CS variability and abstraction

Manipulating CS variability during evaluative conditioning yielded results that are in line with our proposal that CSs can be represented on varying levels of abstraction. Specifically, recognition memory performance was worse in the variable than the invariable condition across experiments. Moreover, EC effects for CS components were about the same size in the invariable condition (indicating equal weighting of components), but larger for the fixed components than varying components in the variable condition. This corresponds to abstraction as a process of simplification that leads to a loss of detailed knowledge (Burgoon et al., 2013; Taylor et al., 2015), and as a predictive process that emphasizes distinctive attributes in multicomponent stimuli (Reed, 2016; Ramscar et al., 2010). In quantitative terms, evaluative ratings of the CS components were in line with the predictions of error-driven learning models (Rescorla & Wagner, 1972), suggesting that variability in CSs may have helped learners to identify the discriminating features while disregarding non-discriminating ones (Ramscar et al., 2010; Raviv et al., 2022).

Note that cue competition and discriminative learning are generally conceived to be of minor importance for the emergence of EC effects (Beckers et al., 2009; Dwyer et al., 2007; Kattner & Green, 2015), even though there is only limited empirical work on the role of cue competition in EC and its importance might be underestimated (Alves et al., 2020; De Houwer et al., 2001). While discriminative learning is often misconceived as mere associative learning (Hoppe et al., 2022), it goes beyond the latter. In discriminative learning, cues can both be associated and dissociated from outcomes, and thus learning depends on the informativity of specific cues relative to other present cues rather than mere contingencies (Hoppe et al., 2022; Rescorla, 1968). Based on the current findings, we argue that cue competition in EC might serve to specify how CSs are represented in memory.

For example, in the present case cue competition predicts that the category-defining perceptual features of CSs themselves acquire evaluative meaning in the variable condition. Nevertheless, we do not have direct evidence for whether the predictive cue was (always) represented in its original format (i.e., the CS component), or whether it was transformed into another format (e.g., the category-defining component could be translated into verbal descriptions such as “three boxes”). Answering this question would require additional research. Moreover, further insights may be generated by investigating whether cue competition can explain evaluative conditioning based on linguistic labels (e.g., a category label; Glaser & Kuchenbrandt, 2017), relational structures (e.g., an artificial grammar; Jurchiş et al., 2020), or common attributes of CSs (e.g., healthiness; Bui & Fazio, 2016). Such a research program would highlight the format of representations supportive of or required for attaching evaluative meaning to abstract entities.

Retrieval stage: abstraction and generalization

At retrieval, abstract CS representations may influence the generalization of acquired likes and dislikes. In the present study, generalization effects were larger in the variable than the invariable condition, even though they were significant in both conditions. Assuming that evaluative conditioning generally adheres to similarity-based generalization principles (Pearce, 1987; Shepard, 1987), similarity of the concrete CS to the generalization stimulus explains generalization in the invariable condition. In the variable condition, generalization is a function of the encoded features. The better people learn to ignore the uninformative features of a stimulus, the stronger the generalization towards stimuli that share the remaining predictive cues (see *Figure 1*). It is possible that the perceptual similarity between CSs and

GSs in the present research is responsible for relatively high levels of generalization also seen in the invariable condition, which might offer an explanation for why the predicted difference in generalization between variable and invariable conditions was not significant in all experiments. Future research could seek to disentangle generalization driven by abstracted cues from generalization driven by perceptual similarity to concrete CS representations more directly.

An elegant way to do so provides the experimental paradigm introduced by Christie and Gentner (2010) to study relational learning in children. They presented images with objects in specific structural configurations (e.g., three cats, three dogs) together with nonwords. During test, children had to make a forced choice between an object match (e.g., one cat) and a relational match (e.g., three novel animals). The former choice indicates generalization based on perceptual similarity, while the latter choice displays generalization based on abstract properties of the images. In an EC experiment, one could use the images as CSs and replace nonwords with USs. As in Christie and Gentner (2010), a forced choice between an object match and a relational match in terms of evaluations would show whether participants generalize according to perceptual similarity or relational structures. In accordance to the present findings, we would suggest that the proportion of relational matches increases with increasing CS variability, demonstrating generalization via abstract CS representations.

As an alternative to abstraction, one could also argue that broader category knowledge of CS categories drives the link between variability and generalization. Such an account proposes a higher *numerosity* of CS representations in the variable condition, rather than more abstract ones. Consequentially, the likelihood that a new stimulus resembles a known one increases, resulting in stronger generalization (Bowman & Zeithamova, 2020; Hahn et al., 2005; Homa et al., 1981; Nosofsky, 1988; 2011). However, our results propose that differences in the numerosity of CS representations alone are not sufficient to explain differences in generalization. First, because unique components of Chinese characters differed strongly within a category (see *Figure 2*), we consider it unlikely that encountering more characters during learning increases the likelihood that a new character resembles a known one. Second, as mentioned above, recognition memory performance diminished as a function of CS variability, which speaks against the assumption that participants represented all CSs as specific identities in the variable condition. Lastly, the size of the EC effects differed between fixed and variable CS components in the variable condition, which also speaks against a representation of CSs as identities rather than distinctive cues.

Limitations and future directions

The present work is limited to the extent that our experimental set-up does not allow for inferences on how CS variability affects the way USs are represented in memory. In other words, our measures are only informative on the ways attitude object are represented (CSs), but not evaluative meaning itself (USs). According to previous findings of research into EC, variability on the side of the USs can result in a link between the CS and US valence that does not involve the specific US identity (Sweldens et al., 2010). In the present experiments, the manipulation of CS variability is confounded with the number of USs that co-occurred with the CSs of a category (one US per category in the invariable condition, five USs per category in the variable condition). Keeping the number of USs per category constant across learning conditions would make it impossible to judge whether participants group CSs together because of their shared features or because of their shared co-occurrence with one particular US. Thus, to specify how USs got represented in memory one would need to include measures that allow for inferences on US representations (see Sweldens et al., 2010).

A second limitation of our experiments lies in the measurement and interpretation of generalization effects. To test generalization, novel Chinese characters from the respective CS categories were included at test. It is important to note that this dependent measure can be seen as a test of generalization strength (how consistently individuals generalize learning to new items; Raviv et al., 2022) rather than generalization width (how 'distant' training and test stimuli are; Raviv et al., 2022). In the experiments, GSs were approximately equally different from CSs. To provide a direct test of the scope of generalization, one could manipulate the GSs' distance in similarity space from the CSs. This approach should yield a generalization gradient (Shepard, 1987), with larger generalization effects for more similar GSs.

Our experiments demonstrated a general insensitivity of CS evaluations towards manipulations of CS variability (in contrast to evaluations of the GSs), suggesting that the effect on liking of the stimuli used for learning does not depend on that factor. This finding contrasts with results from other learning paradigms showing that reaching a similar level of training takes longer under high- than low-variability conditions (Clopper & Pisoni, 2004; Hahn et al., 2005; Huet et al., 2011; Posner & Keele, 1968; Raviv et al., 2022). Future research has to determine whether the valent nature of the contents or evaluative measures in evaluative conditioning make this a general effect or whether idiosyncrasies of our paradigm (e.g., the nature and number of the CSs used) are responsible for this effect.

Practical implications

As well as contributing to theoretical perspectives and methodological approaches to evaluative conditioning, the present research has important practical implications. Evaluative conditioning is a simple learning paradigm that is thought to constitute a model of many applied phenomena such as social and consumer attitudes (Moran et al., 2023). Clearly, abstraction and generalization play an important role in these domains. For example, attitudes towards social groups differ in their level of abstraction with regard to the ingroup-outgroup dichotomy (Park & Rothbart, 1982). Evaluations of outgroup members are often overgeneralizations, while evaluations of ingroup members rely more strongly on individuating information (Krueger & Rothbart, 1988; Park et al., 1991). In the light of the present findings, variability might be one enabling factor for the occurrence of intergroup biases. When encountering various individuals in negative contexts, their common group membership can acquire negative evaluative meaning that goes beyond the specific individuals. This makes group membership salient and predictive of a particular outcome (Le Pelley et al., 2010). Consequently, the formation of an attitude towards (social) groups is less likely when individuals are perceived as unique and group membership does not predict evaluative outcomes. Moreover, we found that variability had a more pronounced effect on the generalization of negative than positive valence in the majority of studies (Experiments 1 and 3), which suggests that valence asymmetries may also play a role for the way prejudice develops and perpetuates.

In contrast, overgeneralization can be also desirable when it comes to interventions targeting attitude change. Including a variety of positive examples could make interventions more effective, for example those targeting the reduction of intergroup biases (Fitzgerald et al., 2019; Paluck et al., 2021). In particular, interventions that employ evaluative conditioning to induce positive attitudes towards disadvantaged groups (Fitzgerald et al., 2019; French et al., 2013; Olson & Fazio, 2006), might benefit from variability in CSs. These ideas could be tested in future research using a manipulation of an EC phase similar to the one presented here.

Conclusion

The present findings give insight into the mechanisms underlying generalization in EC by providing a cognitive-ecological perspective on generalization effects. Theoretically, our results highlight the relevance of variability in generalization and propose that the learning principle of cue competition serves to specify how CSs come to be represented in memory. Practically, the present results offer insights into learning conditions that lay the ground for the formation of intergroup biases and prejudice and could make the design of interventions targeting attitude change more efficient. The present research illustrates that a focus on the nature of the representations formed as a function of the learning environment can prove fruitful in advancing both theorizing and the practical value of evaluative conditioning (Hütter, 2022).

CHAPTER 2: SOCIAL DISTANCE AND ATTITUDE CHANGE

Are attitudes towards outgroup members more resistant to change? On the role of social categories in attitude change via evaluative conditioning

Kathrin Reichmann[1], I-Ching Lee[2], Mandy Hütter[1]

[1]*Eberhard Karls Universität Tübingen, Tübingen, Germany*

[2]*National Taiwan University, Taipei, Taiwan*

Author Notes

Kathrin Reichmann, Mandy Hütter, Department of Psychology, Eberhard Karls Universität Tübingen, Germany. I-Ching Lee, Department of Psychology, National Taiwan University, Taiwan.

The present research was supported by a grant from the National Science Council (106-2410-H-002 -229 -MY2) awarded to I-Ching Lee, by a grant from the Deutsche Forschungsgemeinschaft DFG (381713393) awarded to Mandy Hütter as part of the research group *Modal and Amodal Cognition* (RU 2718).

We have no conflict of interest to disclose.

Correspondence concerning this article should be addressed to Kathrin Reichmann at the Fachbereich Psychologie, Eberhard Karls Universität Tübingen, Schleichstraße 4, 72076 Tübingen, Germany. Electronic mail may be sent to kathrin.reichmann@uni-tuebingen.de.

Abstract

Humans have a natural tendency to categorize social others as ingroup versus outgroup members. Social categories can have consequences for the way group members are represented in memory (i.e., outgroup members are perceived as more similar to one another than ingroup members). In the present work, we asked the question whether such representational differences affect attitude change. Existing attitudes might be easier to modify for ingroup than outgroup members as past studies showed larger degrees of evaluative learning for distinct than similar stimuli. Two experiments (total *N* = 222) tested this hypothesis using evaluative conditioning (EC) as a means of attitude change. In Experiment 1, participants from a Taiwanese university saw faces labeled as Taiwanese (ingroup) versus Chinese (outgroup) together with images of positive or negative valence during conditioning. Experiment 2 tested a German sample and presented German (ingroup), or Taiwanese (outgroup) faces in the learning phase. EC turned out to be equally effective in changing attitudes towards individual group members, even though the results of a recognition memory task in Experiment 2 indicated outgroup homogeneity. Attitude change measured via an indirect measure was qualified by ingroup identification and yielded contrastive, rather than assimilative learning effects for outgroup CSs. We discuss methodological limitations as well as boundary conditions of attitude change.

Keywords: social categorization, evaluative conditioning, intergroup biases, outgroup homogeneity

Humans have a natural tendency to categorize social others as members of their own group (their ingroup) versus members of another group (an outgroup). Importantly, social categorization can affect the way individual group members are evaluated (Bodenhausen et al., 2012; Kawakami et al., 2017), and represented in memory (Boldry et al., 2007; Judd & Park, 1988; Linville et al., 1989; Park & Rothbart, 1982). For example, outgroup members are often evaluated more negatively than ingroup members, which is also referred to as *intergroup biases* (Hewstone et al., 2002). Intergroup biases can have negative consequences, such as prejudice and discrimination (Gilmour, 2015; Le Pelley et al., 2010). This makes the question of how intergroup biases can be reduced and modified one of the most pressing ones in social psychology (Calanchini et al., 2020; Kurdi & Charlesworth, 2023; Lai et al., 2014; Paluck et al., 2021). In addition to intergroup biases, past studies found that outgroup members are perceived as more similar to one another than ingroup members (*outgroup homogeneity*; Judd & Park, 1988; Linville et al., 1989; Park & Rothbart, 1982). Outgroup homogeneity influences the extent to which group-level versus individual-level information is applied in judgements about an individual (Krueger & Rothbart, 1988; Park et al., 1991), which, in turn, can contribute to the emergence of intergroup biases (Montrey & Schultz, 2019).

In the present work, we examine whether differences in the way group members are represented in memory might not only affect the emergence, but also the modification of intergroup biases. That is, representational differences might influence attitude change on the individual level. For instance, it might be the case that attitudes towards outgroup members are more resistant to change than attitudes towards ingroup members, as outgroup members are represented more homogeneously in memory. Such a finding could help to explain why the effectiveness of interventions targeting intergroup biases is often limited (Paluck et al., 2021), and why intergroup biases can perpetuate over time (Perdue et al., 1990). While previous work touched on the question already (e.g., Bettencourt et al., 1997; Bettencourt et al., 2016; Branscombe et al., 1993), a systematic investigation of the question is still lacking in one of the most frequently employed paradigms of attitude change: evaluative conditioning (EC; De Houwer et al., 2001; Hofmann et al., 2010; Levey & Martin, 1975).

Attitude change via evaluative conditioning

Evaluative conditioning refers to a change in the liking of a stimulus (conditioned stimulus or CS) due to the repeated pairing with a valent other stimulus (unconditioned stimulus or US; De Houwer et al., 2001; see Hofmann et al., 2010, for a meta-analysis, and Moran et al., 2023, for a recent review). Past research has shown that EC can effectively alter existing attitudes towards individuals, thereby reducing intergroup biases (Calanchini et al. 2013, 2020; French et al., 2013; Lai et al., 2014; Olson & Fazio, 2006). For example, Olson and Fazio (2006) paired positive words with images of Black and negative words with images of White faces. White participants' implicit attitudes towards Black faces became more positive after the conditioning phase, indicating that EC successfully changed evaluations towards outgroup members. In a similar study, French and colleagues (2013) employed EC to reduce negative evaluations towards Middle Easterners. They paired positive images with Middle Eastern faces and neutral images with White faces. IAT scores reflecting intergroup biases decreased after conditioning, relative to a control group that did not experience CSs and USs together. Lastly, Calanchini et al. (2013) applied counter-prejudicial training in the form of a modified EC procedure (affirm Black-positive and White-negative picture pairings by pressing a "YES" key and disaffirm Black-negative and White-positive picture pairings by pressing a "NO" key). They reported a decrease in IAT scores for the counter-prejudicial training, compared to a pro-prejudicial training group. To summarize, the studies demonstrated that EC can be effective in modifying existing intergroup biases.

Note that in these previous studies, ingroup versus outgroup membership was confounded with US valence. In most experiments, faces of outgroup members were paired with positive information, and faces of ingroup members were paired with neutral or negative information (e.g., French et al., 2013; Olson & Fazio, 2006). The way intergroup biases changed prior to post conditioning was then measured (e.g., comparison of IAT scores before versus after the learning phase; French et al., 2013). While these experimental procedures allowed for an assessment of the qualities of EC as an intervention, they do not allow for an investigation of potential biases in the way existing attitudes towards ingroup versus outgroup members changed. For instance, pre to post conditioning changes in IAT scores might be mostly driven by changes in attitudes towards ingroup members (with little changes in attitudes towards outgroup members), or changes in attitudes towards outgroup members (with little changes in attitudes towards ingroup members), or both. In the present research, we tested whether attitude change via EC occurs to varying degrees for ingroup versus outgroup members. Differences in the representation of group members might affect how attitudes towards individuals can be modified.

Outgroup homogeneity and attitude change

Representational differences arise as outgroup members are often perceived as more similar to one another than ingroup members (*outgroup homogeneity*; Judd & Park, 1988; Linville et al., 1989; Park & Rothbart, 1982). Past work assessed outgroup homogeneity with different measures of perceived variability (Boldry et al., 2007).[11] For example, recognition memory measures showed that participants had more difficulties in discriminating between seen and novel faces when faces were from the outgroup compared to the ingroup (Ackerman et al., 2006; Brigham & Barkowitz, 1978; Chance et al., 1975; Meissner & Brigham, 2001). Park and Rothbart's (1982) dual-storage model provides an explanation for the findings. According to the model, information about the group and about individual group members is stored in two separate representations in memory. When judging the perceived variability of outgroup members, group-level information is primarily retrieved. When judging the perceived variability of ingroup members, information on individual group members is retrieved as well (Judd & Park, 1988; Park & Rothbart, 1982). As a result, perceived variability is higher for ingroup than outgroup members. Construal Level Theory (CLT; Liberman & Trope, 2008; Trope & Liberman, 2010) provides another account for outgroup homogeneity. The theory suggests that there is an inherent relation between the abstractness of representations and psychological distance (Trope & Liberman, 2010). Psychological distance consists of various dimensions, one of which is social distance. Social distance increases when others are categorized as outgroup as opposed to ingroup members, as members of the outgroup are socially more distal to the self than members of the ingroup (Hess et al., 2018). According to CLT, increasing distance results in more abstract representations (Trope & Liberman, 2010). Consequentially, representations of ingroup members would contain low-level information (i.e., concrete details about a particular individual, such as the way a person looks like), while representations of outgroup members would contain high-level features (i.e., features that are true for all individuals of a group, such as the prototypical look of a group member; Hess et al., 2018). Mental representations of social group members should thus differ in their abstractness.

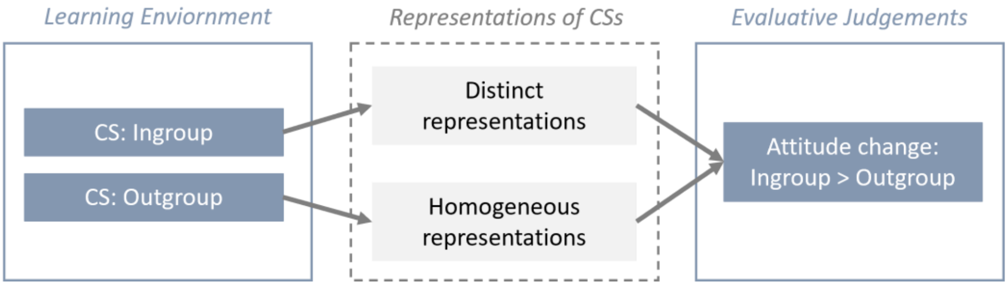
In turn, more abstract representations of outgroup members might make it more difficult to change attitudes on the individual level. Consider a learning scenario where faces from an outgroup versus an ingroup serve as CSs during conditioning (e.g., faces from a foreign country versus one's home country). A representation of ingroup members in terms of their unique attributes and individual features (Park & Rothbart, 1982; Hess et al., 2018) makes ingroup CSs more distinct from one another than outgroup CSs. This should not only lead to better memory for ingroup CSs but might also make evaluative conditioning more effective in changing attitudes towards individual group members. Past work on EC has

[11] Note that there is also research that did not find the outgroup homogeneity effect (e.g., Boldry & Kashy, 1999; Simon, 1992; Simon & Mummendey, 1990).

demonstrated that learning effects increase for more complex and distinct CSs. For example, Hütter et al. (2014; Exp. 2) presented schematic versus naturalistic faces as CSs during conditioning. The EC effect was only significant for naturalistic but not schematic CSs, in line with the idea that easily distinguishable CSs facilitate conditioning effects. Another example stems from Glaser and Kuchenbrandt (2017) who used schematic drawings (Exp. 3; "control" condition) versus more complex pictures of real persons (Exp. 4) as CSs. Effect sizes of EC effects were larger in the latter than the former case, showing that conditioning was more effective for CSs easy to tell apart. Applied to intergroup contexts, this would mean that attitude change via EC should be more pronounced for ingroup than outgroup CSs, when ingroup CSs are represented more distinctively in memory (see *Figure 1*).

Figure 1

Social categorization and attitude change as a function of cognitive representations



Note. Distinct representations of ingroup CSs should increase the degree of attitude change as stimuli can be better discriminated from one another. In turn, more homogeneous representations of outgroup CSs should reduce the degree of attitude change.

The present study

The present study therefore tested the impact of social categories on attitude change via evaluative conditioning. Differences in the way individual group members are represented in memory (i.e., outgroup homogeneity) might make attitudes towards outgroup members more resistant to change.

In the following, we report two EC experiments that investigated this hypothesis in two different intergroup contexts. In Experiment 1, participants from Taiwan saw faces either labeled as Taiwanese (ingroup) or Chinese (outgroup) as CSs together with valent images (USs) during conditioning. Experiment 2 recruited a German sample and presented German (ingroup), or Taiwanese (outgroup) faces as CSs. Ingroup versus outgroup members therefore not only differed in their nationality, but also in their appearance and ethnicity in Experiment 2, which should make representational differences even stronger. At the same time, a stronger intergroup bias was expected in Experiment 1 than Experiment 2, as the respective intergroup context was characterized by both political and military tension between groups (e.g., Gries & Su, 2013; Lee et al., 2018; Lee & Pratto, 2011). Reyling on intergroup contexts that differed in their level of intergroup biases allows one to control for the impact of initial attitudes on attitude change. Because more extreme initial attitudes are generally more difficult to change (Eagly & Chaiken, 1995; Gibson, 2008), more extreme intergroup biases prior to conditioning might reduce the effectiveness of EC to modify existing attitudes in addition to representational differences.

To observe the degree of attitude change, attitudes towards CSs were measured prior and post conditioning. In both experiments, we relied on direct (explicit) and indirect (implicit) measures of attitudes. As opposed to direct measures, indirect measures have the advantage to reduce demand effects and social desirability biases in responding (Gawronski & Brannon, 2018). The present studies are particularly prone to both as participants might think that the study is about racism (or similar) when evaluating faces of different nationalities. The evaluative priming task (Fazio et al., 1986, 1995) was used as an indirect measure of attitudes as it allows for an assessment of attitude change separately for social categories and US valence, instead of providing a relative score (e.g., as it is the case for the IAT; Gawronski & Brannon, 2018). Experiment 2 furthermore employed a recognition memory task as a manipulation check (Fiedler et al., 2021), for the way ingroup as opposed to outgroup members are represented in memory. Participants should have greater difficulties distinguishing between seen and unseen faces of the outgroup than the ingroup (Ackermann et al., 2006; Boldry et al., 2007; Brigham & Barkowitz, 1978; Chance et al., 1975; Meissner & Brigham, 2001), if outgroup members are represented in more abstract terms (Trope & Liberman, 2010). Lastly, we also assessed participants' identification with the ingroup (i.e., identification with Taiwan in Experiment 1 and with Germany in Experiment 2) to control for potential moderating influences on the overall degree of attitude change. Because individuals highly identifying with the ingroup displayed larger intergroup biases in previous studies (Branscombe & Wann, 1994; Hewstone et al., 2002; Leach et al., 2008; Lee et al., 2018), their attitudes might be generally more robust to change.

For all experiments, we report how we determined sample sizes, all data exclusions and all manipulations and measures included in the experiments. Pre-registrations, data files and analysis scripts are publicly available on OSF via https://osf.io/gvq5b/?view_only=313cdab2c4bd459ba7c3f19d7ad29d08. The study material is available upon request.

Experiment 1

The first experiment sought to test our initial hypothesis that EC is more effective in changing attitudes towards ingroup than outgroup members, as a function of the way individual group members are represented in memory. Participants from a Taiwanese university first went through a category learning phase where faces were either labeled as Taiwanese (ingroup) or Chinese (outgroup). The faces were then paired with images of positive or negative valence during conditioning. Attitudes towards the depicted individuals were measured prior and post conditioning with the evaluative priming task (Fazio et al., 1995), as well as on a direct rating scale post conditioning. The study was pre-registered (https://osf.io/pkdcu/?view_only=6fe17c6f07c94c2d8b0ccdf3c536716f).

Method

Participants

We conducted an a-priori power analysis (G*Power; Faul et al., 2007) to estimate the minimum sample size required to detect a small to moderate effect (f = .15) for an interaction of US valence and social category with a power of .80, an alpha-level of .05 and two measurements per cell in the design.[12] The required sample size amounted to N = 90 participants. Data were collected online via a study link that was posted on various social media channels of a Taiwanese university. Only Taiwanese residents living currently in Taiwan were allowed to participate. In total, 94 participants took part in the study. We excluded 10 participants from the data analysis (two participants self-reported that they didn’t follow the instructions or didn’t answer to the question, three failed more than one attention check during the conditioning phase of pressing a button within 5 seconds and five produced more than 20% errors or timeouts in the evaluative priming task).[13] This resulted in a final sample size of N = 84 participants. Participants were between 18 and 58 years old (M = 25.21, SD = 6.14), 54 were female and 30 male. They could sign up for a raffle as a reward for their participation, which took about 15 minutes.

[12] We relied on ANOVAs for the power analysis rather than the multilevel models reported below as estimates for fixed and random effects were not available from prior experiments.

[13] We deviated from the pre-registered exclusion criteria to reduce the number of exclusions by (1) keeping data of participants who self-reported they were distracted during the learning phase but passed the attention checks (5 participants) and (2) keeping participants with ingroup identification scores more than two standard deviations away from the mean scores (4 participants). Reported results did not change substantially when excluding these participants (see analysis script on OSF).

Design

The study followed a 2 (US valence: negative vs. positive) × 2 (time of measurement: pre vs. post) × 2 (social category: ingroup vs. outgroup) within-subjects design.

Materials

15 pleasant and 15 unpleasant pictures from the International Affective Picture System (IAPS; Lang et al., 1997) were used as USs. They differed significantly in valence, $t(28) = 25.12$, $p < .001$, but not in arousal, $t(28) = -0.87$, $p = .390$. A selection of 15 Taiwanese female faces and 15 Taiwanese male faces with neutral facial expressions served as the stimulus pool for CSs. To measure participants' ingroup identification with Taiwan, we used 6 items from Leach et al. (2008; e.g., "I am glad to be Taiwanese") and 4 items from Lee et al. (2018; e.g., "I identify myself as Taiwanese") to operationalize the construct of ingroup identification (Cronbach's $\alpha = .96$ for a joint analysis of both questionnaires). Responses were collected on a 7-point scale with the endpoints "does not apply to me at all" and "applies to me very much". All items are included in the supplementary materials. We also included a short questionnaire about participants' overall attitudes towards Taiwan and Mainland China. Three questions asked how participants generally feel about Taiwan (Mainland China) as a country, how they feel about the Taiwanese (Chinese) government, and the people from Taiwan (Mainland China; Cronbach's $\alpha = .85$ for the joint analysis of ingroup items and $\alpha = .78$ for outgroup items). Responses were recorded on a 7-point scale ranging from "dislike very strongly" to "like very strongly".

Procedure

This and the subsequent experiment were programmed in jsPsych (De Leeuw, 2015). Instructions and target words of the evaluative priming task were presented in Chinese. Participants first provided their demographic data and answered the questions on ingroup identification. They then received the information that the current study investigates how people process images of faces and of negative and positive content (see the supplementary materials for concrete instructions).

Social Categorization Task. Out of the stimulus pool for CSs, four female and four male faces were randomly selected that later also served as CSs during conditioning. In the social categorization task, participants were exposed to two female and two male faces that were presented together with the information that "This person is from Taiwan" (ingroup), and two male and two female faces presented with the statement that "This person is from Mainland China" (outgroup). Every face was shown three times with the statement in a randomized order in the center of the screen, for a duration of 3000ms with a 500ms inter-trial

-interval. We then tested participants' knowledge of the faces' category membership. Participants had to categorize each face as either from Mainland China or Taiwan, by clicking on respective buttons. Feedback was provided for each response. If they categorized more than two faces incorrectly, the categorization test re-started.

Evaluative priming task. To measure participants' attitudes towards CSs before the conditioning phase, we employed the evaluative priming task (EPT; Fazio et al., 1986, 1995). Participants were told to classify presented adjectives as either positive or negative. They should respond as fast and accurately as possible. Twelve positive and 12 negative adjectives were included as target words in the EPT.[14] Negative target words were assigned to the "a"-key on a QWERTY keyboard, and positive target words to the "l"-key. Every trial started with the presentation of a fixation cross for 200ms, followed by the CSs as primes for 200ms, and the presentation of the target words until participants made a response. If they took longer than 1200ms to press a key, the trial would time out and the message "Please respond faster!" would appear for 750ms on the screen. When participants pressed the incorrect key, the feedback "Incorrect response!" was provided. Every CS appeared with six adjectives of negative valence, and six adjectives of positive valence, resulting in a total of 96 trials. The order of the trials was randomized, and two subsequent trials never showed the same CS or the same adjective.

Learning phase. Prior to conditioning, we informed participants that they will now see each face together with a positive or negative image, and that the images should help them form the correct impression of the depicted person. For each CS, a US image was randomly selected from the pool of positive and negative images. One female and one male face per social category were assigned to positive valence, and one female and one male face per category was assigned to negative valence. This resulted in two measurements per cell of the design (US valence × social category), when disregarding the gender of the faces. CSs and USs appeared simultaneously on the screen for 2000ms, followed by an inter-trial-interval of 500ms. Every CS-US pair appeared four times in each of three learning blocks, resulting in a total of 96 stimulus presentations. The order of CS-US pairs was randomized within blocks and left-right assignments of CSs and USs were also randomly determined on a trial-by-trial basis. The same CS-US pair never occurred twice in a row.

Evaluative measures after conditioning. After the conditioning phase, participants again went through the EPT and evaluated each face on a continuous rating scale from -100

[14] Positive target words: "pleasant", "good", "outstanding", "beautiful", "magnificent", "marvelous", "excellent", "appealing", "delightful", "nice", "genuine", "generous"; Negative target words: "unpleasant", "bad", "horrible", "miserable", "hideous", "dreadful", "painful", "repulsive", "awful", "ugly", "phony", "stingy".

(negative) to 100 (positive). They were instructed to indicate whether the depicted person makes a positive or negative impression on them.

Category memory and general attitudes. To see whether participants still remembered the category membership of each face at the end of the experiment, they had to categorize each face according to the assigned nationality (Taiwan vs. Mainland China). Then, participants' general attitudes towards Taiwan and Mainland China were assessed. Participants also indicated whether they paid attention during the learning phase and followed the instructions. Lastly, they had the option to provide a comment on the study.

Results

Data were analyzed with R (R Core Team, 2023), version 4.2.3, using the packages *tidyverse* (Wickham et al., 2019), *lme4* (Bates et al., 2015), *reghelper* (Hughes & Beiner, 2022), and *ggplot2* (Wickham & Chang, 2014).

Ingroup identification and general attitudes

Participants mean ingroup identification scores were significantly above the mean of the scale (M = 5.37, SD = 1.14; on a scale from 1 to 7), t(83) = 6.99, p < .001, d = 0.76. General attitudes towards the ingroup (Taiwan) and outgroup (Mainland China) were assessed by averaging responses across the three items (government, country, people) for each group. One participant was excluded from the analysis who did not provide an answer to the question about the Taiwanese government. Overall, attitudes were more negative towards the outgroup (M = 1.74, SD = 1.04) than the ingroup (M = 4.02, SD = 1.08). This difference was significant, t(82) = 13.75, p < .001, d = 1.51, indicating the presence of a negatively connotated intergroup. General attitudes towards the ingroup strongly correlated with ingroup identification scores, r = .84, t(81) = 14.05, p < .001. This was not the case for attitudes towards the outgroup, r = -.14, t(81) = -1.28, p = .204.

Memory of social categories

Memory for social categories assessed at the end of the experiment and was generally high, with a mean correct response rate of 79.32% significantly above chance level, t(671) = 18.75, p < .001, d = .72. Correct (1) versus incorrect (0) responses were submitted to a generalized linear mixed model with a binomial link function (exponentiated parameters are reported), with a random by-participant intercept and slope for US valence.[15] Chances of correct responses did not differ significantly between social groups, B = 0.80, SE = 0.12, z = -

[15] Fixed effects were effect coded with US valence (-0.5 = negative, 0.5 = positive), and social category (-0.5 = outgroup, 0.5 = ingroup).

1.88, *p* = .061, and did not depend on negative versus positive pairings during conditioning, *B* = 0.75, *SE* = 0.16, *z* = -1.84, *p* = .065, or the interaction of the two variables, *B* = 1.56, *SE* = 0.24, *z* = 1.84, *p* = .066. Descriptively, memory performance was better for negatively (76.19% correct) than positively (77.38% correct) paired ingroup CSs, and better for negatively (87.50% correct) than positively (76.19% correct) paired outgroup CSs.

Evaluative priming task

To analyze the reaction times for positive versus negative target words collected in the evaluative priming task, we calculated evaluative priming scores as recommended by Koppehele-Gossel et al. (2020). Trials that timed out (no response after 1200ms; 1.81%), trials with incorrect responses (4.48%), and trials with reaction times above 1000ms (1.48%) or below 300ms (0.38%) were excluded from data analysis. Mean response latencies were calculated for positive target words and for negative target words per CS and participant for both time points of measurement (pre vs. post conditioning). Those for positive target words were then subtracted from those of negative target words to obtain the priming score. Higher values of the priming score thus indicate more favorable evaluations of CSs, as reaction times were shorter for positive than negative target words following the prime. We submitted EPT scores to a linear mixed-effect model that accounted for inter-individual differences in responding (Judd et al., 2017; Lischetzke et al., 2015). The model included by-subject random intercepts and slopes for time of measurement, US valence, and social categories (without interactions, to reduce model complexity), as well as fixed effects for time of measurement (0 = pre, 1 = post), US valence (-0.5 = negative, 0.5 = positive), and social category (-0.5 = outgroup, 0.5 = ingroup), and the interactions of the fixed effects. *Table 1* reports mean EPT scores and their standard deviations, and *Figure 2* displays the distributions of EPT scores. The supplementary materials include a report of the model with all fixed and random effects.[16]

Prior to conditioning, the main effect of social category did not reach significance, *B* = 4.35, *SE* = 4.96, *t*(384.04) = 0.88, *p* = .381, and EPT scores did not differ between positive and negative US valence, *B* = -0.12, *SE* = 4.91, *t*(904.71) = -0.02, *p* = .981. Moreover, US valence did not interact with time of measurement, *B* = 1.95, *SE* = 6.90, *t*(1090.89) = 0.28, *p* = .778, which indicates that pre-to post changes in EPT scores were not qualified by US valence, as one would expect for the presence of a standard EC effect. Pre- to post

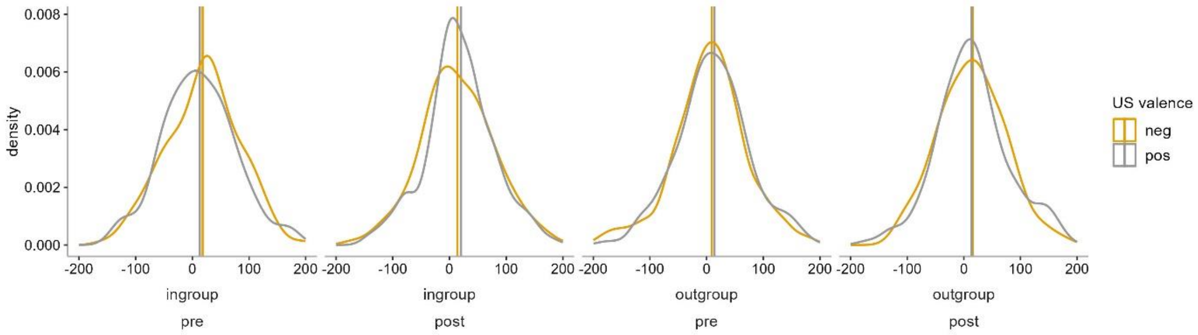
[16] We ran additional analyses that included the gender of the CSs (male vs. female) as a fixed effect in the model. The factor did not result in any main effect or interaction effect on indirect measures (and direct measures) of attitudes and was therefore not included in the report of the analyses.

conditioning differences did also not depend significantly on social categories, *B* = -2.12, *SE* = 6.90, *t*(1090.89) = -0.31, *p* = .758. Lastly, the three-way interaction of US valence, social category and time of measurement was non-significant as well, *B* = 18.61, *SE* = 13.80, *t*(1090.89) =1.35, *p* = .178, in contrast to our initial hypothesis.

Looking at post-conditioning scores only, neither US valence, *B* = 1.83, *SE* = 4.86, *t*(143.85) = 0.38, *p* = .708, nor social category, *B* = 2.22, *SE* = 4.84, *t*(172.37) = 0.46, *p* = .646, or the interaction of the two factors, *B* = 8.83, *SE* = 9.48, *t*(504.00) = 0.93, *p* = .352, elicited a significant influence on the EPT scores.

Figure 2

Density plot of EPT scores per social category, time of measurement and US valence of Experiment 1



Note. Vertical lines indicate mean EPT scores per positive versus negative pairings.

Table 1

Mean EPT scores (and their standard deviations) prior and post conditioning per US valence and social category

	Prior Conditioning		Post Conditioning	
	Positive USs	Negative USs	Positive USs	Negative USs
Experiment 1				
Ingroup	13.10 (65.90)	18.30 (67.80)	20.30 (64.80)	14.10 (64.60)
Outgroup	13.80 (66.50)	9.04 (71.20)	13.70 (65.50)	16.30 (63.90)
Experiment 2				
Ingroup			22.90 (63.00)	14.40 (64.90)
Outgroup			15.30 (69.00)	27.30 (64.80)

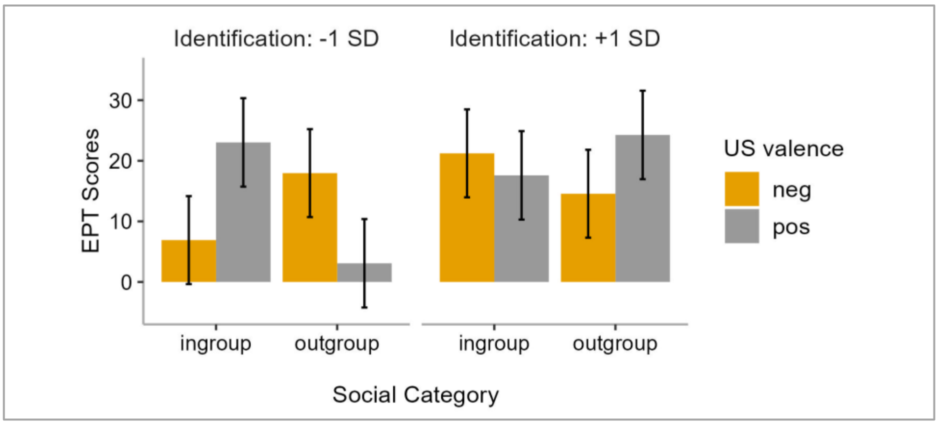
Note. Standard deviations are presented in parentheses. Distributions of the scores are displayed in *Figure 2*.

In an exploratory manner, we added ingroup identification scores (averaged across items per participant, grand-mean centered) to the mixed model. Because higher degrees of ingroup identification were associated with larger degrees of intergroup biases in past research (Lee et al., 2018), ingroup identification could act as a moderating variable in the relation between social category and attitude change. The model yielded a significant four-way interaction between ingroup identification, US valence, social category and time of measurement, *B* = -32.10, *SE* = 12.19, *t*(1090.85) = -2.63, *p* = .009. All the other effects with ingroup identification were non-significant (smallest *p* = .148).

Follow-up analyses (with a Bonferroni-adjusted alpha-level of .013 to account for four exploratory significance tests) showed that for high levels of ingroup identification (+1 *SD*), the three-way interaction of time of measurement, US valence and social category did not reach significance, *B* = -17.65, *SE* = 19.46, *t*(1091.00) = -0.91, *p* = .365. For low levels of ingroup identification (-1 *SD*), differences in EPT scores pre- versus post conditioning depended on the interaction of US valence and social category, *B* = 54.82, *SE* = 19.45, *t*(1090.77) = 2.82, *p* = .005. Here, EPT scores changed in line with US valence for CSs of the ingroup, *B* = 27.00, *SE* = 13.75, *t*(1090.77) = 1.96, *p* = .050. This change occurred in the opposite direction for CSs of the outgroup, *B* = -27.82, *SE* = 13.75, *t*(1090.76) = -2.02, *p* = .043. *Figure 3* displays predicted scores at low (-1 *SD*) and high (+1 *SD*) levels of ingroup identification as a function of social category and US valence post conditioning. To see whether ingroup identification affects EPT scores prior to conditioning, we looked at “pre” measurements separately for each social category. EPT scores for both ingroup, *B* = 4.73, *SE* = 3.89, *t*(84.04) = 1.22, *p* = .227, and outgroup CSs, *B* = 3.22, *SE* = 3.55, *t*(84.00) = 0.91, *p* = .368, were not significantly related to ingroup identification.

Figure 3

Predicted EPT scores for high (+1 SD) and low (-1 SD) levels of ingroup identification as a function of US valence and social category, post conditioning (Experiment 1)



Note. More positive EPT scores indicate more positive evaluations towards CSs. Error bars depict standard errors of predicted scores.

Continuous evaluative ratings

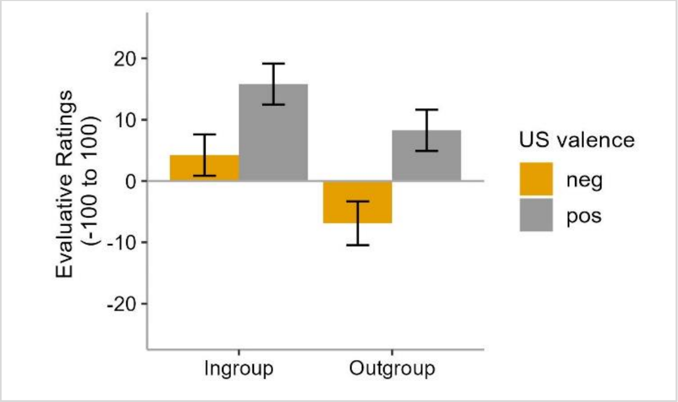
Averaged continuous evaluative ratings (measured on a scale from -100 to 100) are depicted in *Figure 4*, together with their standard errors. Because direct ratings were only collected after conditioning, the calculated mixed model with the factors US valence (-0.5 = negative, 0.5 = positive) and social category (-0.5 = outgroup, 0.5 = ingroup) and random by-subject intercepts.[17] To provide scale-free indicators of parameter coefficients, we report standardized parameter coefficients. They were calculated by fitting the same mixed model to *z*-standardized rather than raw evaluative ratings. Post conditioning, a significant main effect of US valence was observable, *B* = 13.37, *SE* = 3.43, *t*(84) = 3.90, *p* < .001, β = 0.30, indicating the presence of an overall EC effect. On average, evaluative ratings were 13.37 points higher for positively paired CSs than negatively paired ones. Moreover, the main effect of social category was also significant, *B* = 9.33, *SE* = 3.12, *t*(84) = 2.99, *p* = .004, β = 0.21. Thus, ratings were more positive for CSs of the ingroup than the outgroup, indicating the presence of an intergroup bias. The two-way interaction of US valence and social category did not reach significance, *B* = -3.60, *SE* = 5.57, *t*(420) = -0.64, *p* = .519, β = -0.08, which shows that differences in evaluations of negatively versus positively paired CSs did not depend on social categories. Adding ingroup identification scores (averaged across items per participant, grand-mean centered) in an exploratory manner to the mixed model did not reveal a significant interaction with US valence and group membership, *B* = 3.88, *SE* = 4.94, *t*(420) =

[17] Random slopes for US valence and social category were removed due to convergence issues.

0.79, *p* = .433, β = 0.09. All other effects containing self-identification were non-significant as well (smallest *p* = .197).

Figure 4

Mean evaluative ratings in Experiment 1, measured on a scale from -100 to 100



Note. Error bars display standard errors.

Discussion

The goal of the first experiment was to test our initial hypothesis that attitudes towards outgroup members are more resistant to change than attitudes towards ingroup members. Ingroup versus outgroup membership was manipulated by assigning faces either to participants' home country (Taiwan) or to a foreign country (Mainland China). A test of participants' category memory at the end of the experiment showed that category knowledge was sufficiently high, indicating that social category assignments were likely available during learning and test.

Attitudes measured via the evaluative priming task neither displayed an intergroup bias prior to conditioning, nor a standard EC effect post conditioning. In an exploratory manner, we added ingroup identification as a moderator variable to the linear model. For participants low in ingroup identification, a significant EC effect was observable for CSs of the ingroup, and a reversed EC effect (more positive attitudes towards negatively compared to positively paired stimuli) was present for CSs of the outgroup. The former finding was in line with an assimilative learning effect (i.e., evaluations of CSs are informed by USs, such as "CSs paired positively are positive"), while the latter finding corresponded to a contrastive learning effect (i.e., CSs compared to USs, such as "CSs are not as negative as the negative US and therefore positive"; see also Unkelbach & Fiedler, 2016). EC effects were non-significant for participants highly identifying with the ingroup. High ingroup identification might be linked to more extreme initial attitudes towards CSs, which makes attitudes generally more difficult to change (Eagly & Chaiken, 1995; Gibson, 2008). However, EPT scores prior to conditioning were not related to ingroup identification – even though this relation could be observed on an independent measure that assessed participants' general attitudes towards the ingroup (versus the outgroup). Alternatively, it might be the case that ingroup identification influences the way group members are represented in memory, with homogeneous representations of both ingroup and outgroup members for high identifiers (reducing the overall learning effects). In Experiment 2, we added a recognition memory task as an indicator for the way CSs come to be represented in memory. Moreover, we used an intergroup context with highly distinct groups but smaller (expected) intergroup bias to see whether ingroup identification plays a role when initial attitudes are less extreme.

In Experiment 1, attitudes measured directly via a continuous rating scale post conditioning showed a standard EC effect, and evaluations were overall more negative for CSs of the outgroup than the ingroup. This intergroup bias towards Chinese Mainlanders by Taiwanese residents was also observed in previous studies (Gries & Su, 2013; Lee et al., 2018; Lee & Pratto, 2011). On a theoretical level, the finding indicates similar degrees of evaluative learning for ingroup and outgroup members, with an additional impact of group membership on evaluative judgements. Ingroup identification did not influence the size of the EC effect on this measure. To assess attitude change on a direct measure as well, we employed a continuous rating scale both prior and post conditioning in Experiment 2. Additionally, as it is currently unclear whether group-specific differences in evaluative learning were non-existent for attitudes measured via direct ratings, or differences did not emerge because CSs of both social categories were represented in the same way in memory, the recognition memory task in Experiment 2 was also included to disentangle these two alternative explanations.

Experiment 2

Experiment 2 was conducted to replicate and extent the findings of Experiment 1 in a threefold manner. First, we relied on another intergroup context. Data were collected in Germany rather than Taiwan, and CSs from the ingroup were White faces labeled as German, and CSs from the outgroup were Taiwanese faces labeled as Taiwanese. This should make the social category manipulation stronger. Moreover, the intergroup context can be expected to be much more positively connotated than in Experiment 1, leading to a variation in pre-existing biases. Second, we included a recognition memory task as an indicator for the way CSs come to be represented in memory. Better recognition memory performance was expected for ingroup than outgroup members, in line with the outgroup homogeneity effect. Lastly, attitudes towards CSs were assessed on a continuous rating scale prior and post conditioning, to observe attitude change also on a direct measure. The pre-registration for Experiment 2 can be obtained via https://osf.io/dncpy/?view_only=b0e9c059d79c493d8a4633ff96dce5ff.

Method

Participants

We conducted an a-priori power analysis with the package *simr* in R (Green & MacLeod, 2016). To observe a three-way interaction between time of measurement, US valence, and group membership on direct evaluative ratings with a power of .8 (assuming a small effect for the three-way interaction, *B* = 15.00, *SE* = 5.74, with an alpha of .05 and four measurements per cell in the design), the required sample size amounted to *N* = 100.[18] To accommodate for 20% potential data exclusions, we aimed for a sample size of *N* = 120 participants. Data were collected online with a study link distributed via the mailing list of a German university. In total, 128 participants took part in the study.[19] Data sets of 17 participants were excluded according to the pre-registered criteria. Five reported a nationality other than German or did not provide an answer to this question, two self-reported that they did not pay attention during the learning phase, one did not provide answers to the ingroup identification questions and lastly nine participants were excluded because they had ingroup

[18] We based our power analysis on the model that was fitted to the direct evaluative ratings obtained post conditioning in Experiment 1. For the “pre” time point of measurements, responses were sampled from a truncated normal distribution (*M* = 4.15, *SD* = 44.38; assuming no evaluative biases prior to conditioning). The anticipated sample size of *N* = 100 participants provided sufficient power (99.90%, [99.44, 100.00]) to replicate the effect of the three-way interaction between group membership, US valence, and self-identification obtained on evaluative priming scores in Experiment 1 (when only taking the “post” conditioning measurements into account).

[19] Eight additional participants participated because they clicked on the study link directly before it was deactivated. They could thus finish the study even when data collection was already stopped.

identification scores more than two standard deviations away from the mean score. The final sample consisted of 111 participants (89 female, 18 male, 3 diverse, 1 did not answer the question), who were between 18 and 71 years old (*M* = 23.96, *SD* = 8.64). The study took about 15 minutes and participants were reimbursed with course credit or could sign up for a raffle (four vouchers à 25€).

Materials

The same pleasant and unpleasant IAPS pictures (Lang et al., 1997) as in Experiment 1 served as stimulus pools for USs. In addition to the 16 Taiwanese male and 16 female faces, 16 male and 16 female White faces from the Chicago Face Database (Ma et al., 2015) with a neutral expression were added to the CS pool. Four male faces, and four female faces were sampled from Taiwanese and White faces, respectively, resulting in a total number of 16 CSs. Thus, we presented twice as many CSs in this experiment as in Experiment 1. Two of the male, and two of the female faces per social category were assigned to USs of negative, and the other two to USs of positive valence. To measure participants' identification with Germany, the same items as in Experiment 1 were presented in German (Cronbach's α = .81). General attitudes towards Taiwan and Germany were assessed with the three items on the respective country, people, and government (Cronbach's α = .46 for ingroup and α = .66 for outgroup items).

Procedure

The study design and procedure corresponded to Experiment 1 with a few exceptions. Instructions were translated to German. Instead of the category learning phase, participants received the information that they will now see faces from Germany (ingroup CSs) and Taiwan (outgroup CSs). Every CS was presented once on the screen with the sentence "This person is from Taiwan [Germany]" for 3000ms, with a 500ms inter-trial-interval. We did not test category knowledge as strong ceiling effects were expected. Prior to conditioning, participants evaluated each CS on a continuous rating scale from -100 (negative) to 100 (positive). Then, every CS occurred six times with its assigned US of positive or negative valence, resulting in a total number of 96 learning trials as in Experiment 1. Learning trials were divided into three learning blocks, with each CS-US pair occurring in a random order twice per block.

The conditioning phase was followed by a recognition memory task that included the 16 CSs, and 16 previously unseen faces (4 female and 4 male Taiwanese, and 4 female and 4 male White faces) as stimuli. Each stimulus had to be classified as "new" versus "old" by pressing the right or left arrow key on a QWERTY keyboard. Participants were instructed to classify a stimulus as “old” if they recognized the stimulus, and “new” if they didn’t. Every trial started with the presentation of a fixation cross for 1000ms, followed by the target stimulus. If no response was given after 2000ms the trial timed out and the prompt “Please respond faster!” was presented for 750ms. If a response was provided, the screen turned white for 750ms before the next trial started.

Participants then evaluated the CSs on the rating scale and went through the evaluative priming task (EPT). In the EPT, one female and one male face per social category and US valence were randomly selected and presented as primes. In that way, we reduced the required trial number and therefore the overall length of the experiment. The same adjectives (in German) were used as in Experiment 1. The task included four practice trails. The items of the practice trials re-occurred in the actual test phase. Finally, participants answered the questions on their general attitudes towards Germany and Taiwan, and reported whether they paid attention during learning and followed the task instructions.

Results

Ingroup identification and general attitudes

Participants mean ingroup identification scores (*M* = 4.39, *SD* = 0.85) did not differ significantly from the mean of the scale, *t*(110) = -1.39, *p* = .167, *d* = -0.13. For the analysis of general attitudes, data of three participants were excluded who did not provide answers to all three items. Ratings were then averaged across items. Overall, attitudes were more positive towards the ingroup (*M* = 3.73, *SD* = 0.72) than the outgroup (*M* = 3.49, *SD* = 0.64), *t*(107) = 2.95, *p* = .004, *d* = 0.28.[20] As in Experiment 1, general attitudes towards the ingroup were correlated with ingroup identification, *r* = .45, *t*(106) = 5.24, *p* < .001. The correlation was non-significant for attitudes towards the outgroup, *r* = -.01, *t*(106) = -0.12, *p* = .906.

Continuous evaluative ratings

To analyze continuous evaluative ratings (on a scale from -100 to 100), we calculated mixed models with US valence (-0.5 = negative, 0.5 = positive), social category (-0.5 = outgroup, 0.5 = ingroup), and time of measurement (0 = pre, 1 = post) as fixed effects, as well

[20] Due to the low internal consistency of the scales, we also analyzed the items separately. Evaluations of Germany as a country were more positive (*M* = 5.40, *SD* = 0.91) than those of Taiwan (*M* = 4.65, *SD* = 0.87), *t*(107) = 6.40, *p* < .001, *d* = 0.62. Evaluations of the German government (*M* = 4.19, *SD* = 1.25) did not differ significantly from evaluations of the Taiwanese one (*M* = 3.96, *SD* = 0.78), *t*(107) = 1.64, *p* = .104, *d* = 0.16. Lastly, evaluations of Taiwanese were more positive (*M* = 4.85, *SD* = 0.84) than those of Germans (*M* = 4.61, *SD* = 0.96), *t*(107) = -2.49, *p* = .014, *d* = -0.24.

as random by-subject intercepts and slopes for US valence and social category.[21] *Figure 5* depicts the aggregated ratings with their respective standard errors.

Prior to conditioning, the main effect of social category was significant, *B* = -11.26, *SE* = 1.96, *t*(224.53) = -5.75, *p* < .001, β = -0.30. The parameter estimate indicates that evaluations were on average 11.26 points more positive for outgroup than ingroup CSs. Looking at the pre-to post conditioning differences, there was a significant main effect of time, *B* = -2.32, *SE* = 1.07, *t*(3219.02) = -2.17, *p* = .030, β = -0.06, with more negative ratings post compared to prior conditioning. Changes in evaluations were qualified by US valence, as shown by a significant two-way interaction of time and US valence, *B* = 10.20, *SE* = 2.14, *t*(3219.02) = 4.76, *p* < .001, β = 0.27. Central to the present investigation, the three-way interaction of time of measurement, US valence, and social category did not reach significance, *B* = 6.38, *SE* = 428, *t*(3219.02) = 1.49, *p* = .137, β = 0.17. All the other parameter coefficients were non-significant as well (smallest *p* = .167; see the supplementary materials for a full report of the model).

When looking at the "post" conditioning measurements only, it became evident that social category, *B* = -8.29, *SE* = 1.72, *t*(111.00) = -4.82, *p* < .001, β = 0.31, next to US valence, *B* = 8.84, *SE* = 1.89, *t*(111.00) = 4.68, *p* < .001, β = 0.31, influenced the ratings of the CSs. However, the interaction of the two factors was non-significant, *B* = 4.73, *SE* = 2.97, *t*(1442.99) = 1.59, *p* = .112, β = 0.12, in line with the findings of Experiment 1.

Adding ingroup identification scores (averaged across items per participant, grand-mean centered) to the model yielded a two-way interaction between ingroup identification and social category that was present independent of the time point of measurement, *B* = 5.17, *SE* = 1.87, *t*(111.00) = 2.77, *p* = .007, β = 0.14. For ingroup CSs, evaluations were more positive with higher identification scores, *B* = 6.32, *SE* = 2.25, *t*(111.00) = 2.81, *p* = .006, β = 0.17. For outgroup CSs, this was not the case, *B* = 1.15, *SE* = 2.40, *t*(111.00) = 0.47, *p* = .632, β = 0.03. All other effects with ingroup identification were non-significant (smallest *p* = .082).

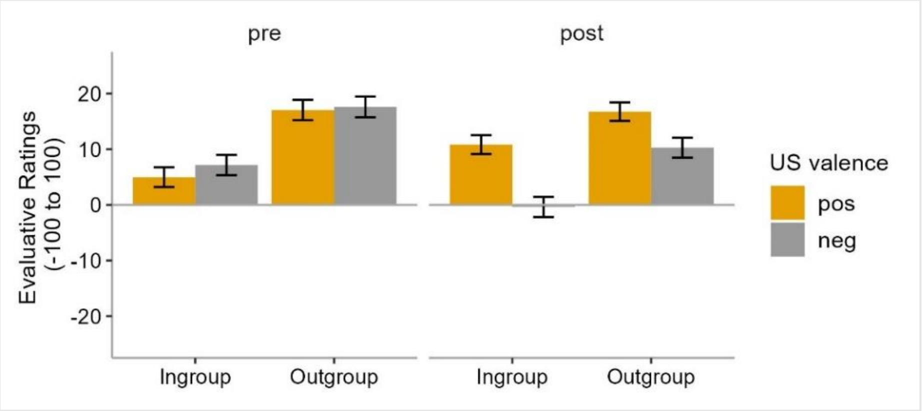
In addition, we tested whether EC effectively reduced initial evaluative biases, looking at positively paired ingroup CSs and negatively paired outgroup CSs only. Prior to conditioning, the main effect of social category was significant, *B* = -12.62, *SE* = 2.68,

[21] The random slope for time of measurement was removed due to convergence issues. Adding gender of the CSs (-0.5 = male vs. 0.5 = female) as an additional factor to the model yielded a significant main effect of CS gender, *B* = 6.85, *SE* = 1.50, *t*(3219.00) = 4.56, *p* < .001, β = 0.18, with more positive evaluations of female than male CSs. The effect depended on social category, *B* = -6.84, *SE* = 3.00, *t*(3219.00) = -2.28, *p* = .023, β = -0.18. The difference between evaluative ratings of male and female CSs was only significant for faces of the outgroup, *B* = 10.27, *SE* = 2.12, *t*(3219.00) = 4.84, *p* < .001, β = 0.27, but not for the ingroup, *B* = 3.43, *SE* = 2.12, *t*(3219.00) = 1.62, *p* = .106, β = 0.09. All the other interactions with CS gender were non-significant (smallest *p* = .478).

t(229.02) = -4.71, *p* < .001, β = -0.33, with more positive evaluations for outgroup than ingroup members. Post conditioning, this effect was no longer significant, *B* = 0.55, *SE* = 2.68, *t*(229.02) = 0.20, *p* = .839, β = 0.01.

Figure 5

Mean evaluative ratings in Experiment 2, pre and post conditioning



Note. Error bars denote standard errors.

Evaluative priming task

We calculated evaluative priming scores as in Experiment 1. Prior to the analysis, seven additional participants were excluded because they had more than 20% errors or timeouts in the EPT (resulting in *N* = 104). We also excluded trials that timed out (1.72%), trials with incorrect responses (6.34%) and trials with reaction times above 1200ms (1.71%) or below 300ms (0.17%). We submitted EPT scores to a mixed-effect model with by-subject random intercepts and slopes for social category and US valence, and fixed effects for US valence (-0.5 = negative, 0.5 = positive), social category (-0.5 = outgroup, 0.5 = ingroup), and their interaction.[22] *Figure 6* displays mean EPT scores per social category and US valence.

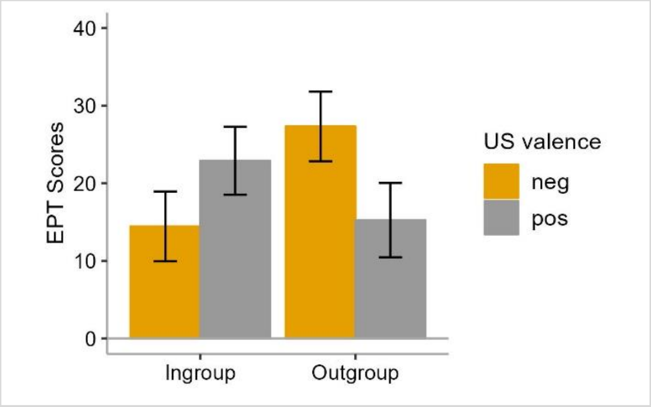
The main effect of US valence did not reach significance, *B* = -1.84, *SE* = 4.18, *t*(669.76) = -0.44, *p* = .660, and neither did the main effect of social category, *B* = -2.65, *SE* = 4.19, *t*(619.38) = -0.63, *p* = .527. However, both factors significantly interacted with each other, *B* = 20.45, *SE* = 8.34, *t*(727.16) = 2.45, *p* = .014. On average, EPT scores for CSs of the outgroup were lower for positive than negative pairings, opposite to what can be expected for a standard EC effect, *B* = -12.07, *SE* = 5.90, *t*(714.35) = -2.05, *p* = .041. EPT scores for CSs of the ingroup did not differ significantly for positive versus negative valence, *B* = 8.39, *SE* =

[22] Adding gender of CSs as an additional factor to the model (-0.5 = male, 0.5 = female) yielded significantly higher EPT scores for female than male faces, *B* = 10.34, *SE* = 4.15, *t*(727.16) = 2.49, *p* = .013. All the other effects with CS gender were non-significant (smallest *p* = .299).

5.91, *t*(714.82) = 1.42, *p* = .156, even though the direction of the effect points towards a standard EC effect. *Table 1* presents mean scores and their standard deviations. Adding ingroup identification scores (grand-mean centered) to the model did not yield any additional significant parameters (smallest *p* = .120).

Figure 6

Mean EPT scores per social category and US valence of Experiment 2



Note. More positive EPT scores indicate more positive evaluations towards CSs. Error bars display standard errors.

Recognition memory task

Trials that timed out in the recognition memory task were excluded from data analysis (2.48%). “Old” (0) and “new” (1) responses were analyzed with a generalized linear mixed-effect model with random by-subject intercepts and slopes for social category and fixed effects for social category (-0.5 = outgroup, 0.5 = ingroup), US valence (-0.5 = negative, 0.5 = positive), and stimulus type (dummy coded, level of interest was set to zero). Exponentiated parameter estimates are reported. For CSs, the odds for responding “old” (correct responses; “hits”) were higher for ingroup than outgroup CSs, *B* = 1.59, *SE* = 0.18, *z* = 2.55, *p* = .011. This effect was not qualified by US valence, *B* = 1.56, *SE* = 0.32, *z* = 1.40, *p* = .161. For lures, odds to provide an “old” response (incorrect response; “false alarms”) were higher for outgroup than ingroup stimuli, *B* = 0.62, *SE* = 0.20, *z* = -2.34, *p* = .019. To summarize, participants were better able to discriminate between seen versus unseen ingroup than outgroup stimuli. Adding ingroup identification as an additional effect to the model did not influence the results neither for CSs, nor for lures (smallest *p* = .246). *Table 2* presents proportions of “old” responses and their standard deviations.

Table 2

Proportion (standard deviations) of 'old' responses for ingroup versus outgroup CSs and lures

	Ingroup	Outgroup
Experiment 2		
CSs	0.91 (0.20)	0.87 (0.21)
Lures	0.07 (0.17)	0.11 (0.22)

Note. 'Old' responses to CSs are correct responses, while 'old' responses to lures are false alarms.

Discussion

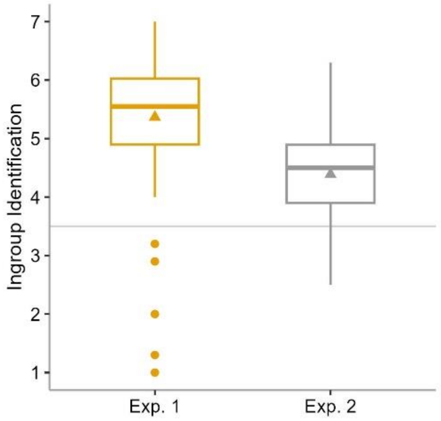
Experiment 2 investigated the influence of social categories on the degree of attitude change in another intergroup context. For direct evaluative ratings, we again observed a main effect of US valence that was not qualified by social category, indicating that EC was equally effective in changing attitudes towards ingroup and outgroup CSs. Social category elicited a main effect on evaluative responding, with more positive evaluations for outgroup than ingroup CSs. Adding ingroup identification as an additional factor to the model showed that this evaluative bias was likely due to the way ingroup CSs were evaluated. Evaluations of ingroup CSs became more positive with increasing ingroup identification, in line with previous work documenting a positive relationship between ingroup identification and ingroup favoritism (e.g., Lee & Pratto, 2011; Lee et al., 2018). Evaluations of outgroup CSs were not associated with ingroup identification. Outcomes of the evaluative priming task corresponded to result patterns obtained for participants low in ingroup identification in Experiment 1. A reversed EC effect was observable on EPT scores for outgroup CSs, while EPT scores pointed towards a standard EC effect for ingroup CSs (albeit this effect was non-significant). In general, ingroup identification was significantly lower in Experiment 2 than Experiment 1 (see *Figure 7*),[23] which would explain why results resembled those of low identifiers in Experiment 1. Lastly, an additional recognition memory task showed that participants were better able to discriminate between ingroup than outgroup stimuli, in line with the results of previous studies (Ackermann et al., 2006; Brigham & Barkowitz, 1978; Chance et al., 1975; Meissner & Brigham, 2001), and the outgroup homogeneity effect

[23] Mean ingroup identification in Experiment 1, $M_{Exp.1}$ = 4.37, $SD_{Exp.1}$ = 1.14, versus Experiment 2, $M_{Exp.2}$ = 3.39, $SD_{Exp.2}$ = 0.85, t(148.49) = 6.61, p < .001, d = .99.

(Boldry et al., 2007). However, in contrast to our initial assumption representational differences did not affect attitude change via EC in the expected manner.

Figure 7

Ingroup identification scores in Experiments 1 and 2



Note. Triangles display mean scores.

General Discussion

Categorizing social others as ingroup versus outgroup members can have consequences for the way individuals are evaluated (i.e., intergroup biases) and represented in memory (i.e., outgroup homogeneity). In the present study, we tested whether differences in memory representations of ingroup versus outgroup members affect attitude change. We hypothesized that initial attitudes are easier to modify for ingroup than outgroup members, as past studies showed larger degrees of evaluative learning for distinct than similar stimuli (e.g., Hütter et al., 2014). Using evaluative conditioning (EC; Hofmann et al., 2010) as a means of attitude change, two experiments presented faces as conditioned stimuli (CSs) with images of positive or negative valence (unconditioned stimuli; USs) during learning. Participants were informed that each face was either from their home country (ingroup) or from a foreign country (outgroup). In Experiment 1, CSs were perceptually similar but differed in their assigned group membership. In Experiment 2, ingroup versus outgroup CSs also differed in their appearance and ethnicity, making the social category manipulation stronger. Both direct and indirect measures of attitude change were employed in the experiments, and a recognition memory task was added to Experiment 2 to test the way CSs were encoded in memory.

Contrary to our initial hypothesis, attitudes measured on a continuous rating scale changed in the direction of US valence independent of the social category of CSs. Post conditioning, evaluations reflected the presence of an intergroup bias in Experiment 1, and more positive evaluations of outgroup than ingroup members in Experiment 2. Evaluative judgements of CSs were thus determined both by individuating information (the conditioned attitudes) and group-level information (attitudes towards the ingroup versus the outgroup) across experiments. The results are in line with Park and Rothbarts' (1982) dual storage model, suggesting that people store both abstract information about groups and individual information about group members in memory, and retrieve both in judgements. They also correspond to connectionist models of impression formation that postulate a joint contribution of stereotypes and individuating information to evaluations (Kunda & Thagard, 1996; Labiouse & French, 2001). Experiment 2 measured attitudes directly both prior and post conditioning, which allowed us to test whether pre-existing biases in evaluations were reduced via EC. A preference of outgroup over ingroup CSs was no longer present post conditioning (looking at positively paired ingroup CSs and negatively paired outgroup CSs only). This adds to a vast body of research demonstrating that EC can effectively reduce biases in initial evaluations (e.g., Calachini et al. 2013; French et al., 2013; Lai et al., 2014; Olson & Fazio, 2006).

Attitudes measured indirectly via the evaluative priming task (EPT; Fazio et al., 1986, 1995) were also not in line with the initial hypothesis. In Experiment 1, attitudes changed in the direction of US valence only for CSs of the ingroup, for participants low in ingroup identification. Attitudes towards CSs of the outgroup changed in the opposite direction of US valence (higher EPT scores for negatively than positively paired CSs). Significant attitude change did not occur for participants high in ingroup identification. Results of Experiment 2 resembled those of low identifiers in Experiment 1. A reversed EC effect was present for outgroup CSs, whereas a standard EC effect was observable (but did not reach significance) for ingroup CSs. The results of the two studies are comparable to the extent that ingroup identification was generally lower in Experiment 2 than Experiment 1. From these initial results, two implications can be drawn that should be tested in future research. First, ingroup identification seems to be an important variable in attitude change on indirect measures. It might be the case that high ingroup identification makes attitudes towards individual group members generally more difficult to change, as high identifiers hold strong beliefs towards social groups (e.g., Branscombe & Wann, 1994; Hewstone et al., 2002; Leach et al., 2008; Lee et al., 2018). Second, attitudes on indirect measures did not only change in the direction of US valence (for ingroup CSs and low identifiers in Exp. 1, and ingroup CSs in Exp. 2), but reversed EC effects were observable as well (for outgroup CSs and low identifiers in Exp. 1, and outgroup CSs in Exp.2). Several EC studies reported reversed EC effects, for example when relational qualifiers were introduced for CS-US pairs via instructions (CS as a “friend” vs. “enemy” of the US, Fiedler & Unkelbach, 2011; see also Förderer & Unkelbach, 2012), or via an additional task (e.g., select the more likable face out of two; Unkelbach & Fiedler, 2016). Unkelbach and Fiedler (2016) argued that participants encoded contrastive CS-US relations when having to compare stimuli in the forced-choice task. Applied to the present case, this would mean that participants compared outgroup CSs with valent USs (i.e., the CS “is different from” the US), while informing ingroup CSs by valent USs (i.e., the CS “is similar to” the US). Open questions remain about the functional relevance of the distinct learning strategies as well as the role of ingroup identification in the encoding of contrastive CS-US relations. For example, low identifiers might switch strategies when learning about ingroup versus outgroup members because they perceive the US valence as too extreme for unfamiliar outgroup CSs (e.g., “this person cannot be that bad”). One way to test this hypothesis would be to manipulate ingroup identification experimentally prior to conditioning and ask participants to indicate whether they view the CS as “similar to” versus “different form” the US.

In addition to evaluative judgements, Experiment 2 tested participants' recognition memory for presented CSs to see whether ingroup CSs were indeed represented more distinctively in memory. Recognition memory was indeed better for ingroup than outgroup CSs, in line with the outgroup homogeneity effect (Boldry et al., 2007; Judd & Park, 1988; Park & Rothbart, 1982). This finding is an important one as it shows that similar degrees of evaluative learning observed on the direct measure were not due to similar memory representations of ingroup versus outgroup CSs, but independent from differences in the way CSs are represented.

Limitations and future directions

The present work can be seen as a starting point of studying the role of social categories in attitude change via EC. The findings are limited to the extent that attitude change was assessed on the individual rather than the group level. In the conditioning procedures, category membership was not diagnostic of the valence of USs (because both ingroup, and outgroup CSs co-occurred with positive or negative USs). This differed from other EC experiments that used CSs from different categories. For example, Reichmann and colleagues (2023) presented categories of CSs that were characterized by one common element. CSs of a category were then either paired positively or negatively during conditioning, and attitudes towards novel stimuli of the same category were measured. Similarly, Glaser and Kuchenbrandt (2017) presented CSs from two fictitious groups with USs of either positive or negative valence. They assessed the generalization of acquired attitudes by presenting novel stimuli from the fictitious groups during test. The degree of generalization would thus be one way to infer learning on the group level (Glaser & Kuchenbrandt, 2017; Reichmann et al., 2023; see also Jurchiș et al., 2020; Luck et al., 2020). As of now, it is unclear whether social categorization would affect attitude change towards categories per se. Manipulating US valence between-subjects, rather than within-subjects (e.g., pairing ingroup CSs positively and outgroup CSs negatively, as well as the other way around) would provide the chance to measure generalization (e.g., presenting novel ingroup and outgroup stimuli). For instance, it might be the case that attitude change towards individual group members occurs to the same degree for ingroup and outgroup members, but attitude change towards categories is more pronounced for the outgroup as group membership should be more salient for outgroup members.

Secondly, to ensure the generalizability of our findings, measures other than the ones employed in the present experiments should be used in future studies. This includes alternative indirect measures of attitudes (e.g., the affect misattribution procedure; Payne et al., 2005; Payne & Lundberg, 2014), especially because the EPT was repeatedly criticized for its low reliability (e.g., Cunningham et al., 2001; Gawronski & De Houwer, 2014; Koppehele-Gossel et al., 2020). Moreover, ingroup identification could be assessed indirectly, for example by asking for the perceived similarity of group members with the self (Stephan et al., 2011), or via tasks like the dictator game that tests whether participants allocate more resources to the ingroup than the outgroup (Forsythe et al., 1994; Guala & Mittone, 2010). Lastly, measures beyond recognition memory may be employed to test for outgroup homogeneity, such as categorization tasks (e.g., the "subgroup generation measure"; Boldry et al., 2007; Park et al., 1992) or range tasks (e.g., perceived distribution of members along a quantitative dimension; Simon & Brown, 1987).

Boundary conditions of attitude change

One of the most intensely studied topics in social psychology constitutes interventions to reduce and modify intergroup biases (Calanchini et al., 2020; Kurdi & Charlesworth, 2023; Lai et al., 2014; Paluck et al., 2021). While much work generally tested the effectiveness of interventions, little is known about the enabling factors of attitude change. This highlights the importance to study determinants of the learning stage (i.e., the acquisition and modification of attitudes) beyond determinants of the judgement stage (i.e., the retrieval and application of attitudes in evaluations; Hütter & Rothermund, 2020). The present research studied the social categorization of attitude objects as one potential factor. We initially hypothesized that more distinct representations of ingroup members does not only result in better memory for ingroup CSs but also increases EC effects towards ingroup compared to outgroup members. This would make the distinctiveness of attitude objects an important boundary condition of attitude change.

Even though we did not obtain evidence for a direct impact of social categories on the effectiveness of evaluative conditioning, aspects of the learning environment that are indirectly linked to social categories might still influence attitude change. For instance, in real-world contexts people are exposed to more information about the ingroup than the outgroup (Blau, 1994; Denrell, 2005; Konovalova & Le Mens, 2020). For example, Konovalova and Le Mens (2020) argued that more frequent exposure to information about the ingroup than outgroup leads to more variable perceptions of the ingroup, which can explain outgroup homogeneity. In the present experiment, such a bias in the distribution of information in the environment was not reflected in the conditioning procedure. Every CS-US pair occurred with the same frequency in the learning phase. Being exposed to a higher number of CS-US pairings for CSs of the ingroup than the outgroup offers more chances to revise one’s attitudes towards ingroup members. As a result, attitudes towards outgroup members would remain stable, while attitudes towards ingroup members would flexibly adapt to novel information.

Relatedly, a similar result would occur if people could freely choose to receive information on ingroup versus outgroup members. More negative initial attitudes towards outgroup members would make it likely that people choose members of the ingroup over outgroup members, given that people approach things they expect to be positive and avoid things expected to be negative (Denrell, 2005; Fazio et al., 2004; Hütter et al., 2022). However, only through approach behavior initial attitudes can get updated and revised (Fazio et al., 2004). In an evaluative conditioning paradigm where participants can freely choose between CSs that would then occur together with USs of positive or negative valence (as introduced by Hütter et al., 2022), this would mean that only CSs of the ingroup would change in the direction of US valence. Attitudes towards avoided outgroup CSs would remain stable over time.[24] This example illustrates how sampling biases due to pre-existing attitudes towards ingroup versus outgroup members would make attitude change more likely towards ingroup members. In turn, introducing specific sampling goals (e.g., to learn more about outgroup members specifically) could help to update attitudes on outgroup members as well (see Niese & Hütter, 2023). To conclude, interventions targeting intergroup biases might benefit from taking characteristics of the learning environment into account (e.g., the amount of information available or sampled about ingroup versus outgroup members).

Conclusion

The present research focused on the potential consequences of social categories for attitude change, based on the observation that outgroup members are represented more homogeneously in memory than ingroup members. Whereas social categories did not influence the degree of evaluative learning in the predicted manner, future research could further test the role of ingroup identification, contrastive versus assimilative learning effects, and characteristics of the information ecology associated with social categories as boundary conditions of attitude change. A more nuanced understanding of such moderating factors would help to explain why intergroup biases perpetuate over time and could offer promising ways to develop more effective interventions to reduce intergroup biases.

[24] Note that in such a learning scenario, people would have equal opportunity to interact with members of each group, holding the likelihood of interactions constant across social categories (as opposed to the influence of social proximity or similarity on the likelihood of interactions as suggested by Denrell, 2005).

CHAPTER 3: ECOLOGICAL CONDITIONS OF US REVALUATION

Abstract representations of attitudes: Do they make evaluative conditioning resistant to US revaluation? A study of ecological conditions

Kathrin Reichmann & Mandy Hütter

Eberhard Karls Universität Tübingen

Author Notes

Kathrin Reichmann, Mandy Hütter, Department of Psychology, Eberhard Karls Universität Tübingen, Germany.

The present research was supported by a grant from the Deutsche Forschungsgemeinschaft DFG (381713393) awarded to Mandy Hütter, Barbara Kaup, and Michael Ramscar as part of the research group *Modal and Amodal Cognition* (RU 2718).

We have no conflict of interest to disclose.

Correspondence concerning this article should be addressed to Kathrin Reichmann or Mandy Hütter at the Fachbereich Psychologie, Eberhard Karls Universität Tübingen, Schleichstraße 4, 72076 Tübingen, Germany. Electronic mail may be sent to kathrin.reichmann@uni-tuebingen.de or mandy.huetter@uni-tuebingen.de.

We thank Marcus Dietlein and Julia Bogner for their help in collecting the data of the second experiment.

Abstract

Past research suggests that attitudes acquired via evaluative conditioning (EC) are sensitive to a post-conditioning change of the valence of unconditioned stimuli (USs). In the present research, we propose that EC can result in abstract representations of US valence (stimulus-valence learning) that should make EC effects resistant to US revaluation. In a first experiment, we showed that pairing one conditioned stimulus (CS) with multiple USs of the same valence (“one-to-many” pairings) reduces US revaluation effects and diminishes memory for specific CS-US pairs. Experiments 2 and 3 then focused on the presentation sequence as another constraint of stimulus-valence learning. One-to-many pairings were presented either in a forward (CSs before USs) or backward (USs before CSs) manner. Presentation sequence should influence US revaluation if prediction drives abstraction in stimulus-valence learning. As this was not the case, we suggest comparison, rather than prediction, as a candidate mechanism underlying stimulus-valence learning in EC.

Keywords: evaluative conditioning, attitude acquisition, abstraction, US revaluation, presentation sequence

Intuitively, once acquired attitudes and preferences are difficult to change in the light of novel information. Learning that a company falsely advertised its environmental friendliness might still lead to a positive perception of the company (e.g., “greenwashing”; de Freitas Netto et al., 2020). A bad experience with a friend might still leave a bad impression, even when the friend later puts their behavior into context. Counter to this intuition, research on attitude acquisition and change suggested that likes and dislikes are sensitive to revaluations of initial evaluative experiences (Baeyens et al., 1992; Jensen-Fielding et al., 2018; Sweldens et al., 2010; Walther et al., 2009). These past studies relied on one of the most straightforward ways to acquire attitudes: Evaluative conditioning (EC), which refers to attitude acquisition via the pairing of stimuli (De Houwer et al., 2001). In the present research, we aim to study the boundary conditions that make attitudes acquired via EC susceptible to a revaluation of initial learning experiences.

Evaluative conditioning and US revaluation

In evaluative conditioning, neutral stimuli (“conditioned stimuli”, CSs) co-occur with stimuli of positive or negative valence (“unconditioned stimuli”, USs) in spatiotemporal proximity on the screen. The typical result is a shift in evaluations of the CSs in the direction of the US valence (*EC effect;* see De Houwer et al., 2001; Hofmann et al., 2010; Moran et al., 2023, for reviews). Several studies documented the sensitivity of EC effects to changes of initial evaluative experiences (Baeyens et al., 1992; Jensen-Fielding et al., 2018; Sweldens et al., 2010; Walther et al., 2009). For example, Jensen-Fielding and colleagues (2018) presented geometric shapes (CSs) with angry versus happy faces (USs) during conditioning. In a subsequent revaluation phase, participants received information about the USs that was either congruent to the US valence (e.g., positive information about a happy face) or incongruent to the initial valence (e.g., negative information about a happy face). Incongruent information reduced the overall size of the EC effect relative to congruent information, demonstrating the presence of a *US revaluation effect*. Another example provides the study of Walther et al. (2009). They showed that pictures of faces (CSs) that co-occurred with liked others (USs) were evaluated positively first, but evaluated negatively when the liked others were put into a bad light (Walther et al., 2009). Other work documented US revaluation effects in EC as well (Baeyens et al., 1992; Gast & Rothermund, 2011; Walther et al., 2009).

Sensitivity of EC to US revaluation not only demonstrates how likes and dislikes can change in the light of novel information but has also theoretical implications for the way attitudes are represented in memory. More specifically, US revaluation effects indicate that the identity of the US must have been stored together with the CS (see also Dayan &

Berridge, 2014). Because the CS never occurs directly with the revaluated US, the specific US still must play a role for the evaluations of the CS. In other words, US revaluation effects imply that links between a particular CS (e.g., a geometric shape) and a particular US (e.g., a happy face) were acquired during conditioning. Evaluative representations that contain both stimulus identities can also be described as stimulus-stimulus (S-S) learning (Fazio, 2001). In S-S learning, stimuli are encoded as episodic representations of specific events (Gawronski & Bodenhausen, 2018), such as a pictorial code of an image, or an olfactory trace of a pleasant smell.

However, there is also evidence that S-S learning alone might not be sufficient to explain the outcomes of EC. There are studies that did not obtain US revaluation effects in certain conditioning procedures (Baeyens et al., 1998; Gast & Rothermund, 2011; Sweldens et al., 2010). For example, Gast and Rothermund (2011) asked participants to evaluate CS-US pairs during conditioning, and obtained EC effects that were not affected by subsequent US revaluation. In addition, Sweldens and colleagues (2010; Experiment 1) paired one CS with multiple USs of the same valence ("one-to-many" pairings) and reported that evaluations did not change after USs were revaluated. The examples suggest that a systematic relation might exist between specifics of the EC procedure and the degree of US revaluation observable. In the present work, we propose that learning conditions that highlight the US valence while discounting other, irrelevant features of the USs might lead to representations that are more abstract than S-S links. In turn, abstract representations of US valence would reduce the impact of US revaluation on EC effects.

Stimulus-Valence learning in EC

Abstract representations of US valence might no longer entail specific US identities, but rather the overall valence of USs (which we will refer to as stimulus-valence or S-V learning in the following). For example, conditioning procedures with one-to-many pairings may enable participants to identify the common attribute of USs (i.e., their valence), while discounting specific details of USs. As a result, abstract representations of the inferred valence are less specific than representations of concrete stimuli. The inferred valence of USs could be encoded in different ways, for example in the form of linguistic labels (e.g., "positive" or "negative"), numerical codes (e.g., -1 or 1), or evaluative responses (e.g., positive or negative affective responses). [25] Independent of their specific content, abstract representations of USs would make EC resistant to US revaluation, as the details of US

[25] We thus consider stimulus-response (or S-R learning; e.g., Gast & Rothermund, 2011) a subtype of S-V learning. See the general discussion.

identities are not retained during the abstraction process (in line with a definition of abstraction as a process that omits irrelevant features and retains relevant ones; Trope & Liberman, 2010; see also Burgoon et al., 2013; Gilead et al., 2020; Reed, 2016). Additionally, memory for CS-US pairings should get “fuzzy” in stimulus-valence learning, as representations concentrate on US valence rather than specific USs.

Under what kind of learning conditions does stimulus-valence learning become particularly likely? As mentioned before, one relevant factor might be the pairing schedule of CSs and USs (Sweldens et al., 2010; Experiment 1). Pairing one CS with various different USs in one-to-many pairings, as opposed to pairing the same CS repeatedly with a US (“one-to-one” pairings) might facilitate abstraction as the variability in USs helps learners to identify what USs have in common. A similar argument was raised before in other contexts, for example regarding the way attitude objects are encoded in memory (more abstract representations for variable attitude objects; Reichmann et al., 2023), or the way relational concepts are acquired (multiple learning exemplars facilitate the extraction of relations; Christie & Gentner, 2010). Whereas Sweldens et al. (2010) showed that one-to-many pairings resulted in smaller effects of US revaluation than one-to-one pairings, a conceptual replication of this effect as well as a more direct measure of the acquired representations is still missing. The first aim of the present paper was thus to compare US revaluation effects in one-to-one versus one-to-many pairings. A memory measure for CS-US pairings should indicate whether representations indeed become more abstract in one-to-many pairings, leading to reduced memory performance.

Next to the pairing schedule, a second factor of relevance might be the presentation sequence of stimuli. Some authors argue that abstraction is inherently related to prediction (e.g., Gilead et al., 2020; Tenenbaum et al., 2011). Learning conditions that allow learners to predict and then correct their expectations facilitate the extraction of relevant features across stimuli (Hoppe et al., 2022; Ramscar, 2021; Ramscar et al., 2010). Applied to evaluative conditioning, one-to-many pairings might only lead to abstraction if they are presented in a forward rather than backward manner. In forward conditioning, CSs are presented before USs, while the reversed presentation sequence is employed in backward conditioning (e.g., Kim et al., 2016). In forward presentations, CSs can correctly predict the valence of USs across trials (i.e., because the same CS is followed by different USs of the same valence), but not specific US identities. In backward presentations, CSs cannot obtain this predictive function as they occur after each US. Accordingly, if such a predictive function of CSs plays a role in stimulus-valence learning, effects of US revaluation should be smaller after forward than backward conditioning. This would also mean that CSs would obtain a signaling function in EC similar to the postulated role of CSs in classical conditioning (Baeyens et al., 1992; Bouton, 1994; De Houwer et al., 2001; McSweeney & Bierley, 1984). The second aim of the present work therefore was to test the role of prediction in stimulus-valence learning, by presenting one-to-many pairings either in a forward or a backward manner. US revaluation, as well as memory for CS-US pairings, should depend on the presentation sequence if abstraction is indeed facilitated by the accurate prediction of US valence across trials.

The present research

In the following, we report three evaluative conditioning experiments that studied the boundary conditions of US revaluation in EC. In the first experiment, the pairing schedule of CSs and USs was manipulated (one-to-one vs. one-to-many pairings), and the degree of US revaluation as well as participants' memory for CS-US pairings were assessed after conditioning. The second and third experiments varied the presentation sequence of one-to-many pairings (forward vs. backward presentations). Again, US revaluation and memory for CS-US pairs were measured as indicators for the way USs were represented in memory.

The experiments were pre-registered and approved by the ethics committee of the authors' home institution. Materials, analysis scripts and data files can be obtained via (https://osf.io/b5yvr/?view_only=3de7c37c057e4877af938075927aa2d2). For all experiments, we report a-priori power analyses, all data exclusions and all manipulations and measures included in the experiments. We conducted Bayesian analysis in addition to frequentist analysis to test whether null effects indicate data insensitivity or provide evidence for the null hypothesis over the alternative one (Dienes, 2014; Lakens et al., 2020).

Experiment 1

The first experiment tested the hypothesis that one-to-many pairings in EC facilitate the extraction of US valence across learning trials, leading to a representation of the overall valence rather than specific US identities (stimulus-valence learning). In line with the findings of Sweldens et al. (2010; Experiment 1), we expected reduced effects of US revaluation after one-to-many than one-to-one pairings. In addition to a conceptual replication of this known effect, we assessed participants memory for CS-US pairings to inform our abstraction hypothesis. Memory performance should be diminished after one-to-many compared to one-to-one pairings if participants acquired more abstract US representations in the former learning condition. The present experiment held the total number of USs presented during conditioning constant across learning conditions, thereby controlling for cognitive load. The pre-registration for Experiment 1 is available under https://osf.io/p9zuv/?view_only=fc5087728bfd4b9ca259668b70a0e2ff.[26]

Method

Participants

We conducted an a-priori power analysis based on the effect of pairings (one-to-one vs. one-to-many) on the degree of US revaluation reported by Sweldens et al. (2010, Experiment 1), using a mixed model that included by-subject and by-CS random intercepts.[27] To have reasonable power to observe a three-way interaction of US valence, US revaluation, and pairing procedure that would confirm our initial hypothesis (with an alpha-level of .05, one measure per cell in the design, and a power of .80), the sample size was set to *N* = 230 participants. The study was conducted online via Prolific (www.prolific.co). The participant pool was restricted to participants from Germany, with German as a first or fluent language. Of the final data set (*N* = 242),[28] one participant was excluded who reported to not have paid attention during the learning phase, and two were excluded because they did not pass the attention check during conditioning (button press within 7 seconds between learning blocks).

[26] Data for Experiment 1 were collected after Experiments 2 and 3. The experiment included an additional task that measured participants' reaction times for pressing the left arrow key for "negative" and right arrow key for "positive", indicating whether a CS makes a positive or negative impression on them. The data of this additional task are available on OSF.

[27] Sweldens et al. (2010, p. 479) reported a small effect for the difference in revaluation for one-to-one (*M* = 4.70 and *M_reval* = 3.94) versus one-to-many pairings (*M* = 4.02 and *M_reval* = 3.89) for positively paired CSs. For the power analysis, we transformed the means to a 0 to 100-point scale (positive valence) and -100 to 0 scale (negative valence), assuming similar effects for positive and negative valence. We based the power calculations on a mixed model with variances of the random intercepts for CSs (*SD* = 17.98) and participants (*SD* = 6.34) taken from prior studies, and outcomes sampled from normal distributions with *SD* = 25.00.

[28] Data of 12 additional participants were collected, as they timed-out on prolific but finished the study nevertheless, which resulted in a higher number of data sets than anticipated.

The sample consisted of university students (*n* = 53), people currently employed (*n* = 157) or looking for a job (*n* = 16; *n* = 12 did not provide an answer to this question), all aged between 18 and 72 years (*M* = 33.68, *SD* = 11.19). The study took about 10 minutes and participants received 1.50 GBP as reimbursement.

Design

The experiment employed a 2 (US valence: positive vs. negative, within- subjects) × 2 (US revaluation: congruent vs. incongruent, within- subjects) × 2 (pairing schedule: one-to-one vs. one-to-many, between- subjects) mixed design.

Materials

USs were 20 smiling faces and 20 angry faces from the Chicago Face Database (Ma et al., 2015). A pool of 56 unknown logos served as CSs. For US revaluation, 32 statements describing positive behaviors (e.g., “goes shopping for a sick neighbor”) and 32 statements describing negative behaviors (e.g., “tells lies about colleagues”) were created in German.

Procedure

The experiment was programmed in jsPsych (de Leeuw, 2015), and executed on the participants’ private laptop or PC. Participants first completed the conditioning phase, followed by the US revaluation phase and the evaluative ratings as well as the memory task.

Conditioning phase. For the conditioning phase, 4 logos were randomly selected as CSs from the pool of 48 CSs. Two CSs were paired with USs of positive valence, and two CSs with USs of negative valence. Our operationalization of the pairing schedule consisted of either one US per CS (one-to-one pairings) or five different USs of the same valence per CS (one-to-many pairings). In one-to-one pairings, each CS-US pair was presented five times during the conditioning phase, while each CS-US pair was presented once in one-to-many pairings. In that way, each CS appeared equally often in both conditions. In addition, we added ten positive and ten negative USs (one-to-one pairings) or two positive and two negative USs (one-to-many pairings) as filler stimuli to the conditioning procedure. In that way, we made sure that participants saw an equal number of USs in both conditions (24 USs). Filler USs were presented once for one-to-one pairings, and five times for one-to-many pairings. The total number of learning trials (40 trials) was therefore held constant across learning conditions. The 40 trials were divided into two learning blocks with 20 trails each. CS-US pairs were presented in a forward manner. Specifically, CSs were presented for 1500ms, followed by an inter-stimulus-interval of 400ms and the presentation of the US for 1500ms. Filler USs appeared for 1500ms without a CS. After an inter-trial-interval of

3000ms, the next stimulus was presented. CSs and USs were presented at the center of the screen.

Prior to the conditioning phase, participants were informed that they will see logos that were designed in a graphical workshop by teams of multiple people (one-to-many pairings) or a single person (one-to-one pairings). Every logo would be followed by the person who contributed to the design of the logo (one-to-many pairings) or designed the logo (one-to-one pairings). In addition, they also received the information that they will see faces of people who did not participate in the workshop and thus appear on their own. See the supplementary materials for the complete task instructions presented in the experiment.

US revaluation phase. After the conditioning phase, participants were informed that they will now see every individual just encountered again with additional information, to test how first impressions are revised. They should remember the additional information about every individual as best as they could. In the US revaluation phase, each US (filler USs and USs paired with CSs) was then presented together with statements that were either congruent (positive [negative] statements for positive [negative] USs) or incongruent (positive [negative] statements for negative [positive] USs) with the initial valence of the US. For example, in the congruent condition a smiling face could occur with the statement “is often in a good mood”, while in the incongruent condition a smiling face could occur with the statement “makes fun of others”. Filler USs were also presented during US revaluation, half of them paired with congruent, and half of them with incongruent statements. The statements appeared below each US. Participants could take as long as they needed to read the statements and continued via button press. Statements were presented in two learning blocks, with every US appearing once per block. This resulted in a total number of 48 learning trials in the US revaluation phase.

Evaluative ratings. Participants rated each CS on a scale from -100 (unpleasant) to 100 (pleasant) by moving a slider on the screen, being instructed to indicate whether the CS seems pleasant or unpleasant to them.

US identity memory. Next, we tested participants’ memory for individual CS-US pairs. For each US that previously occurred with a CS, participants were asked to select the CS out of the four CSs that the US occurred with. This procedure allowed us to test participants memory for every CS-US pair while keeping the response options (the four CSs) the same across pairing schedules.

Lastly, they answered two questions on their task compliance (whether they followed the instructions and paid attention during the learning phase) and were then thanked and dismissed.

Results

We performed all analyses in R (R Core Team, 2023), version 4.2.3, using the packages *tidyverse* (Wickham et al., 2019), *lme4* (Bates et al., 2015), *ggplot2* (Wickham & Chang, 2014) and *brms* (Bürkner, 2017; 2018), which is based on *Stan* (Stan Development Team, 2017; Carpenter et al., 2017), for this and the subsequent experiments.

Direct evaluative ratings

Evaluative ratings were submitted to a mixed-effect model that accounted for inter-individual and inter-item differences in responding (Judd et al., 2017). The model included fixed effects for US valence, US revaluation and presentation sequence (and their interactions), and random by-subject and by-CS intercepts. Fixed effects were effect coded.[29] To provide scale-free indicators of effect sizes, we also report standardized regression coefficients. To do so, we fitted the mixed model to z-standardized evaluative ratings. Because non-significant findings in a frequentist framework are ambiguous as they can indicate a lack of adequate power or support of a null hypothesis over the alternative (Dienes, 2014), we report Bayes factors together with the three-way interaction of US valence, US revaluation, and presentation sequence (which reflects our main hypothesis). We fitted Bayesian multilevel models with the same fixed and random effects as specified for the model described above to standardized evaluative ratings. We used the *brms* package's default priors for the intercept (Student's t-distribution with $\nu = 3$, $\mu = 24$ and $\sigma = 54.9$) and for the standard deviations of the random effects as well as sigma (Student's t-distribution with $\nu = 3$, $\mu = 0$ and $\sigma = 54.9$). Flat priors that render any logically possible parameter value equally likely were employed for parameter coefficients. For every parameter, we also report the range of "most likely" estimates with 95% credible intervals (95% CI). To calculate the Bayes factor, we compared the model with the three-way interaction of US valence, US revaluation, and pairing schedule to a model without the interaction. The resulting Bayes Factor indicates the extent to which data are more likely under the full model over the simplified model. Thus, a Bayes Factor of $BF_{10} > 1$ indicates that the full model is favored over the model without the effect given the data. We interpret a $BF_{10} > 3$ as substantial evidence for the alternative

[29] Specified as lmer(ratings ~ USvalence * revaluation * pairing schedule+ (1| subject) + (1|CS)) in R, using the lme4 package (Bates et al., 2015). Effect coded: US valence (-0.5 neg, 0.5 pos), US revaluation (-0.5 congruent, 0.5 incongruent), pairings (-0.5 one-to-one, 0.5 one-to-many).

relative to the null hypothesis, and a BF_{10} < 0.33 as substantial evidence for the null relative to the alternative hypothesis (Dienes, 2014; Wagenmakers et al., 2018).[30]

Mean evaluative responses as a function of US valence, US revaluation, and pairing schedule are displayed in *Figure 1*. Overall, there was a main effect of US valence on evaluative ratings, *B* = 23.73, *SE* = 2.79, *t*(697.04) = 8.50, *p* < .001, β = 0.49. On average, evaluative ratings were 23.73 points higher for positive than negative pairings (on a scale from -100 to 100), demonstrating the presence of a standard EC effect. The size of this main effect was larger with one-to-many than one-to-one pairings, as indicated by a significant two-way interaction of US valence and pairing schedule, *B* = 18.11, *SE* = 5.66, *t*(714.24) = 3.20, *p* = .001, β = 0.37. Moreover, US revaluation moderated the size of the EC effect, *B* = -23.25, *SE* = 5.61, *t*(703.61) = -4.14, *p* < .001, β = -0.48. The negative parameter estimate shows that the EC effect was, on average, smaller for CSs with incongruently paired USs than congruently paired ones. The three-way interaction of US valence, US revaluation and pairing schedule was significant, *B* = 23.83, *SE* = 11.24, *t*(706.43) = 2.12, *p* = .034, β = 0.49. Additional Bayesian analysis confirmed that the present data presented substantial evidence for the alternative hypothesis over the null hypothesis, B_{Bayes} = 0.49, 95% CI [0.03, 0.95], BF_{10} = 5.33. All other parameters of the model did not differ significantly from zero (smallest *p* = .064). See Supplement B for a full report of the model and remaining parameter coefficients, in this and the subsequent experiments.

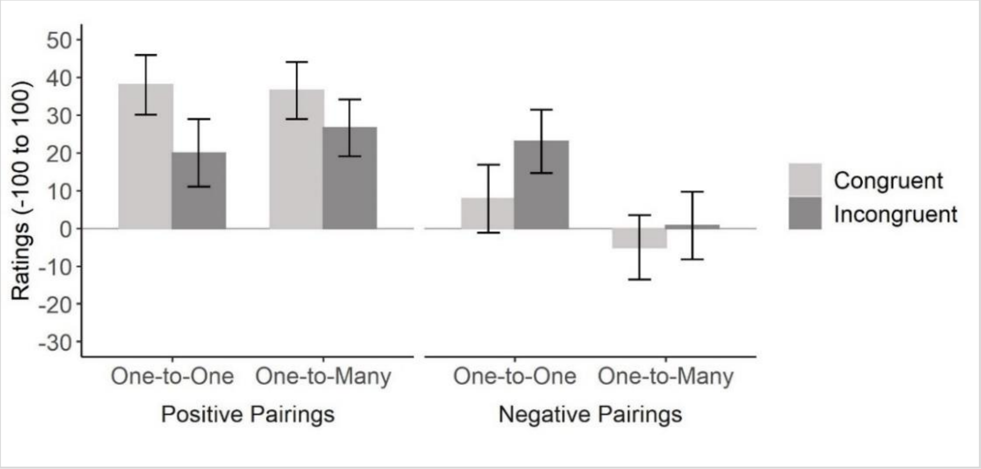
We further analyzed the three-way interaction by calculating the US revaluation effect (US valence × US revaluation) separately for each pairing schedule, by setting the condition of interest to zero and estimating the two-way interaction of US valence and US revaluation. For one-to-one pairings, the interaction was significant, *B* = -35.17, *SE* = 7.84, *t*(697.40) = -4.49, *p* < .001, β = -0.72. Evaluative ratings of CSs were higher after congruent than incongruent pairings for positive US valence, *B* = -19.07, *SE* = 5.55, *t*(700.21) = -3.43, *p* < .001, β = -0.39. Ratings were lower for congruent than incongruent pairings for negative US valence, *B* = 16.09, *SE* = 5.56, *t*(702.28) = 2.89, *p* = .004, β = 0.33. For one-to-many pairings, the US revaluation effect did not reach significance, *B* = -11.34, *SE* = 8.05, *t*(711.91) = -1.41, *p* = .159, β = -0.23, even though the effect was in the expected direction.

We also compared the size of standard EC effects by looking at congruent pairings only. Pairing schedule did not qualify the difference between positive and negative pairings, *B* = 6.19, *SE* = 8.00, *t*(714.77) = 0.77, *p* = .439, β = 0.13, indicating that standard EC effects did not differ between pairing schedules.

[30] All Bayesian models conducted included 20000 iterations, 4 chains, and 2000 warm-ups per chain.

Figure 1

Mean Evaluative Ratings in Experiment 1



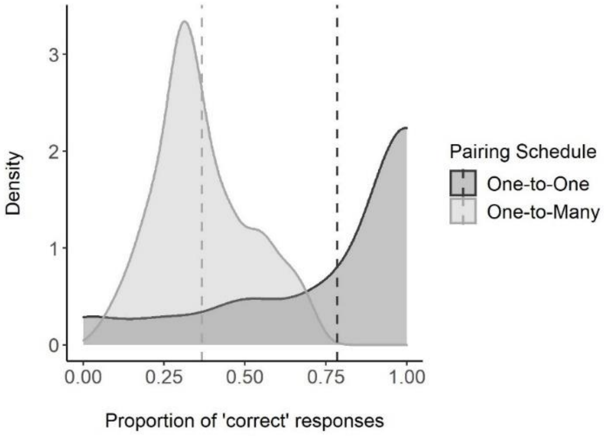
Note. Error bars display 95% confidence intervals.

Memory task

Next, we analyzed participants' memory performance for CS-US pairings by submitting incorrect (0) and correct (1) responses to a generalized mixed effect model with random by-subject and by-CS intercepts, and a fixed effect for pairing schedule, US valence, and US revaluation. We report exponentiated parameter coefficients. The odds of providing a correct response were lower for one-to-many than one-to-one pairings, B = 0.13, SE = 0.16, z = 12.86, p < .001. *Figure 2* depicts proportions of correct responses for each pairing schedule condition in a density plot. All the other parameter coefficients were non-significant (smallest p = .249). They are included in a table in Supplement C.

Figure 2

Density plot of correct responses of the memory task in Experiment 1



Note. Dashed lines display mean proportions of correct responses for each pairing schedule condition.

Discussion

The aim of Experiment 1 was to investigate whether one-to-many pairings as opposed to one-to-one pairings facilitate stimulus-valence learning in evaluative conditioning. In line with the findings of Sweldens et al. (2010; Experiment 1), the effect of US revaluation on evaluative responses was only significant for one-to-one, but not one-to-many pairings. When taking only congruent pairings into account, the EC effect did not differ between pairing conditions. Thus, while both pairing schedules resulted in standard EC effects, presenting various USs per CS made evaluations more resistant to US revaluation. In addition, memory performance for CS-US pairs was reduced after one-to-many compared to one-to-one pairings, even though participants saw an equal number of USs in both learning conditions. The outcome corresponds to the idea that one-to-many pairings facilitate the formation of abstract representations of US valence that no longer conserve specific US identities. With the next two studies, we turned to the role of the presentation sequence in stimulus-valence learning, testing whether prediction plays a role in the abstraction of US valence.

Experiment 2

Experiment 2 manipulated the presentation sequence of stimuli (forward vs. backward conditioning), using a one-to-many pairing schedule. Presentations sequence might act as another environmental constraint of stimulus-valence learning if prediction plays a role in the abstraction of US valence. In forward conditioning (CSs presented before the USs), CSs correctly predict the US valence across trails, but not specific US identities. This might facilitate the extraction of relevant features across trials (i.e., US valence), while neglecting irrelevant ones (i.e., US identities). In backward conditioning (USs presented before the CSs), CSs cannot obtain this predictive function. As a result, smaller effects of US revaluation should be observable for forward compared to backward conditioning. The pre-registration for the experiment is available under https://osf.io/pvhtd/?view_only=ffba4923d0784e7fb5d147222e4fd8f9.

Method

Participants

We aimed to collect N = 200 data sets based on an a-priori power analysis in G*Power (Faul et al., 2007; to obtain a small effect of f = .1 for a within-between interaction in an ANOVA, with a minimum power of .8, an alpha-level of .05, 2 groups, and 2 repeated measures per group),[31] plus an additional 20% to account for potential data exclusions (resulting in N = 240 participants). Data collection was conducted online via the university mailing list. The final sample consisted of 209 university students,[32] of which 30 had to be excluded because they reported that they had not paid attention during the learning phase or failed more than one of three attention checks (press a button within 5 seconds in between learning blocks). Thus, data sets of 179 university students (130 female, 47 male, one diverse, one did not respond) of different majors, aged between 18 and 59 years (M = 23.42, SD = 6.47) were included in the data analysis. Participants could sign up for a raffle (20 x 20€ vouchers for a local bookstore) or receive course credit as reimbursement. The experiment took about 15 minutes.

[31] Because this experiment was the first one conducted in the present line of research, estimates for fixed and random parameters were not available for an a-priori power analysis. We therefore relied on ANOVAs to approximate the required sample size. We also pre-registered ANOVAs for the data analysis but conducted multilevel models as they can be considered to be the more appropriate approach (DeBruine & Barr, 2021; Judd et al., 2017).

[32] Data collection was terminated after N = 209 participants because no more students clicked on the study link within a week.

Design

The study employed a 2 (US valence: positive vs. negative) × 2 (US revaluation: congruent vs. incongruent) × 2 (presentation sequence: forward vs. backward) mixed design with repeated measures on the first two factors.

Procedure

We used the same sets of stimuli as described for Experiment 1. For the conditioning phase, 8 logos were randomly selected as CSs from the pool of logos. 4 CSs were paired positively, and 4 were paired negatively. For each CS, 3 different USs of the same valence were randomly chosen from the pool of smiling (US+) or angry faces (US-). This resulted in a total of 24 USs for 8 CSs. Every CS-US pair was presented once per learning block, with a total of 96 learning trials in four learning blocks. Thus, every CS-US pair occurred four times during the conditioning phase. *Table 1* presents an overview of the experimental parameters of this and the other experiments. Participants could take a short break after each learning block. In the forward condition, CSs were presented for 1500ms, followed by an inter-stimulus-interval of 500ms and the presentation of the US for 1500ms. After an inter-trial-interval of 2000ms, the next stimulus pair was presented. In the backward condition, the sequence of CS and US presentations was reversed. CS-US pairs occurred in a random order in each learning block. As in Experiment 1, participants were informed that they will see logos designed by teams of multiple people, and every logo will appear with the person who contributed to the design of the logo. During US revaluation, USs of two positively (negatively) paired CSs appeared with congruent statements, and USs of two positively (negatively) paired CSs were paired with incongruent statements. At the end of the experiment, participants evaluated each CS on its pleasantness on a rating scale ranging from -100 (unpleasant) to 100 (pleasant).

Table 1

Overview of procedural parameters for "one-to-many" pairings across experiments.

	Exp. 1	Exp. 2	Exp. 3
Number of USs per CS	5	3	6
Total Number of CSs	4	8	4
Total Number of USs	20 (+4)	24	24
Number of Repetitions	1	4	1
Total Number of Learning trails	20 (+ 20)	96	24

Note. CS = conditioned stimulus. US = unconditioned stimulus. In Experiment 1, 4 filler USs were included and presented in 20 additional learning trials without CSs (in parentheses).

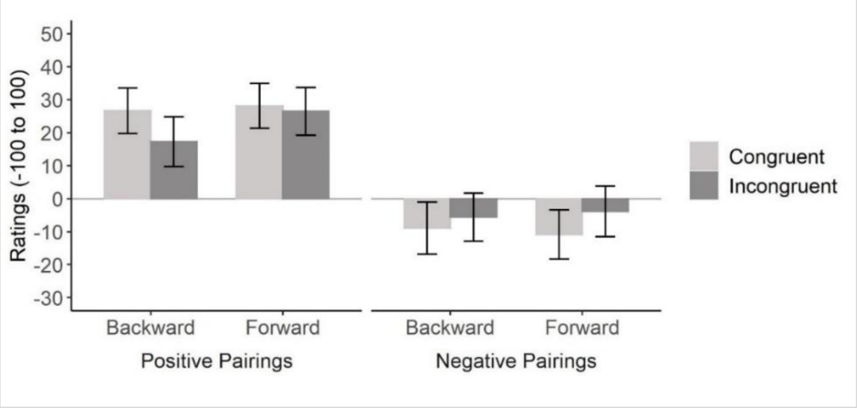
Results

Direct evaluative ratings

Evaluative ratings were submitted to the similar mixed effect model as conducted for Experiment 1, in which the factor pairing schedule was replaced with the factor presentation sequence.[33] Aggregated evaluative ratings are displayed in *Figure 3*. There was a main effect of US valence on evaluative ratings, *B* = 31.50, *SE* = 2.48, *t*(1211.54) = 12.68, *p* < .001, β = 0.60, demonstrating a standard EC effect across sequence and revaluation conditions. The size of the EC effect did not depend on the presentation sequence, as the two-way interaction of US valence and presentation sequence was non-significant, *B* = 6.79, *SE* = 4.97, *t*(1212.50) = 1.36, *p* = .173, β = 0.13. The result demonstrates similar amounts of evaluative learning in backward and forward conditioning. However, US revaluation significantly reduced the EC effect for incongruent compared to congruent pairings, *B* = -12.35, *SE* = 5.00, *t*(1221.56) = -2.47, *p* = .038, β = -0.24. Lastly, the parameter estimate for the three-way interaction of US valence, US revaluation and sequencing did not reach significance, *B* = 5.39, *SE* = 9.91, *t*(1207.82) = 0.54, *p* = .587, β = 0.10. All other parameters were non-significant (smallest *p* = .362). Additional Bayesian analysis indicated that the outcome provided "anecdotal" evidence for the null hypothesis over the alternative hypothesis, B_{Bayes} = 0.10, 95% CI [-0.27, 0.48], BF_{10} = 0.55 (Wagenmakers et al., 2018). However, considering that $0.33 \leq BF_{10} \leq 3$, the data can be seen as insensitive in distinguishing between the alternative and the null hypothesis (Dienes, 2014).

Figure 3

Mean Evaluative Ratings in Experiment 2



Note. Error bars depict 95% confidence intervals.

[33] Effect coded: US valence (-0.5 negative, 0.5 positive), US revaluation (-0.5 congruent, 0.5 incongruent), presentation sequence (-0.5 backward, 0.5 forward).

Discussion

Experiment 2 presented one-to-many pairings either in a forward or backward manner, to test the role of prediction in stimulus-valence learning. We found that presentation sequence did not influence the overall size of the EC effect, showing that similar degrees of evaluative learning were observable for backward and forward conditioning (in line with the findings of Kim et al., 2016). Importantly, the results indicated that the degree of US revaluation was not influenced by the presentation sequence of stimuli, even though the effect was overall significant. Thus, the results did not confirm the hypothesis that forward conditioning would facilitate the abstraction of US valence. However, the finding should be interpreted with caution as additional Bayesian analysis indicated data insensitivity (Dienes, 2014). We therefore aimed to strengthen the role of predictive learning in forward conditioning in Experiment 3.

Experiment 3

Experiment 3 sought to test the same hypothesis as Experiment 2, but employed an experimental procedure that should further strengthen the potential role of prediction in stimulus-valence learning. First, we paired each CS with six different USs of the same valence instead of three. Second, we presented each CS-US pair only once during conditioning instead of four times. Under these adaptations, CSs correctly predict the US valence across learning trials in 100%, but the specific US identity only in 16.67% of the trails in forward conditioning. This should facilitate the extraction of US valence across trials, leading to abstract representations of the overall US valence. In backward conditioning, such a predictive process is not possible as USs occur before CSs. We therefore again expected smaller degrees of US revaluation for forward as opposed to backward conditioning. The experiment also included a memory task on CS-US pairings as an additional indicator for the way USs are represented in memory. Here, backward conditioning should lead to better memory performance than forward conditioning if prediction plays a role in stimulus-valence learning. The pre-registration for Experiment 3 is available under https://osf.io/y4swv/?view_only=32fc3b243475494d947a57e3a2a3cea3.

Method

Participants

We conducted an a-priori power analysis with the package *simr* in R (Green & MacLeod, 2016) based on the multilevel model fitted to the data of Experiment 2. To obtain a significant fixed effect for the three-way interaction (assuming an effect size of B = 30 instead of B = 5.39, SE = 8.87, and with one instead of two measurements per cell in the design), a minimum sample size of 320 participants was required to achieve sufficient power of .8 (α = .05). Data were collected online via the university's mailing list. After the exclusion of 11 participants who reported to not have paid attention during learning, a total of N = 309 data sets were included in the analysis. The resulting sample consisted of university students of different majors (241 female, 61 male, 2 diverse, 1 did not answer), who were between 18 and 70 years old (M = 22.72, SD = 5.42). The study took about 12 minutes and participants received course credit or could sign up for a raffle (20 x 20€) for a local bookstore.

Procedure

Study design and procedure matched Experiment 2, with two modifications of the conditioning procedure. First, we changed the number of USs per CS from three to six, and simultaneously reduced the total number of CSs from eight to four. As a result, 24 stimuli were included as USs (as in Experiment 2), with six USs assigned to each CS. One CS was assigned to the positive-congruent, one to the positive-incongruent, one to the negative-congruent, and one to the negative-incongruent condition. CSs and USs were either presented in a forward, or a backward manner (between-subjects). As a second adaptation, CS-US pairs were shown only once in a single learning block, instead of four times across four learning blocks. Thus, the total number of learning trails amounted to 24 (instead of 96 as in Experiment 2; see *Table 1*).

In addition to evaluative ratings on the continuous rating scale (-100 to 100), we appended the same memory task for the CS-US pairings as employed in Experiment 1. Every US was displayed in the center of the screen and the four CSs of the conditioning phase were shown in the bottom. Participants were instructed to select the CS that occurred with the respective US during the learning phase. The next trial started after a response was made.

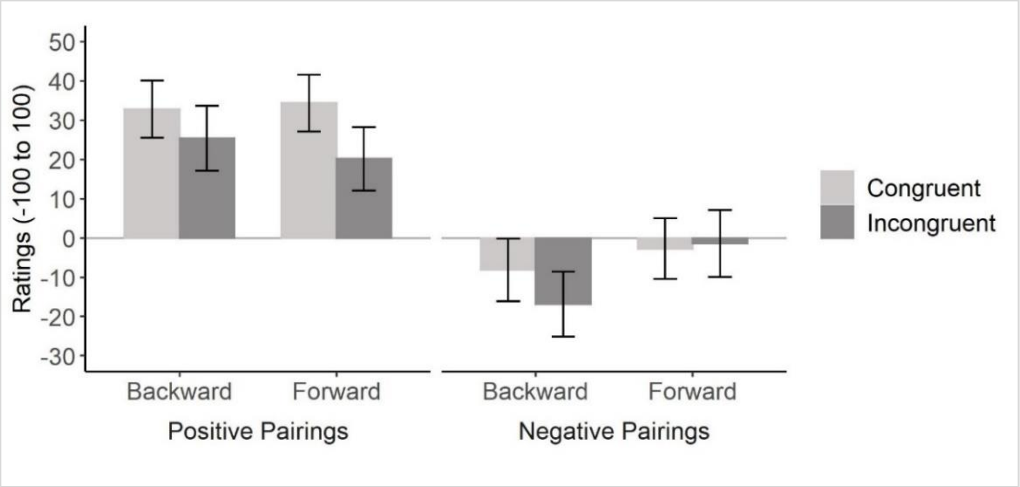
Results

Direct evaluative ratings

Direct evaluative ratings were submitted to the same mixed-effects model as specified for Experiment 2. *Figure 4* displays the aggregated evaluative ratings per US valence, US revaluation, and presentation sequence. Overall, the effect of US valence on evaluative ratings was significant, B = 35.54, SE = 2.69, t(1199.73) = 13.19, p < .001, β = 0.66, demonstrating the presence of a standard EC effect across experimental conditions. The EC effect was moderated by the presentation sequence, B = -12.02, SE = 5.41, t(1203.26) = -2.22, p = .026, β = -0.22, with a reduced EC effect after forward than backward conditioning. There was no US revaluation effect, as US revaluation did not interact significantly with US valence, B = -6.99, SE = 5.42, t(1206.34) = -1.29, p = .198, β = -0.13. In addition, the main effect of US revaluation did not reach significance, B = -5.31, SE = 2.71, t(1204.90) = -1.96, p = .050, β = -0.10. Lastly, the three-way interaction of US valence, US revaluation and presentation sequence was also non-significant, B = -15.22, SE = 10.77, t(1199.32) = -1.41, p = .158, β = -0.28. All the other parameter coefficients did not reach significance (smallest p = .216). The Bayes factor for the hypothesis that there is a three-way interaction over the hypothesis that there is none indicated “anecdotal” evidence for the null hypothesis, B_{Bayes} = -0.28, 95% CI [-0.68, 0.11], BF_{10} = 1.36. Again, because the Bayes factor was situated between 0.33 ≤ BF_{10} ≤ 3, the evidence can be seen as rather indecisive (Dienes, 2014).

Figure 4

Mean Evaluative Ratings in Experiment 3



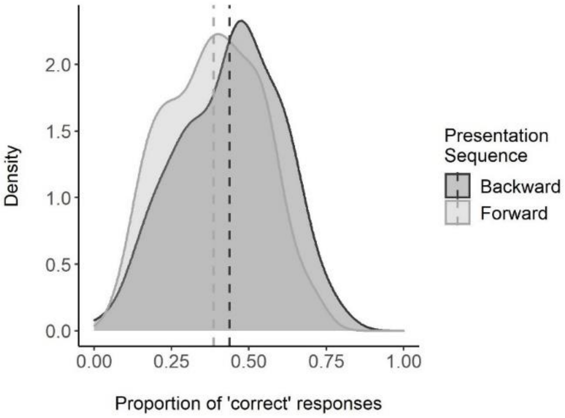
Note. Error bars display 95% confidence intervals.

Memory task

We submitted incorrect (0) and correct (1) responses collected in the memory task to a generalized mixed effect model with fixed effects for US valence, US revaluation, and presentation sequence, with random by-subject and by-CS intercepts. We report exponentiated parameter coefficients. The main effect of presentation sequence was significant, *B* = 0.80, *SE* = 0.08, *z* = -2.89, *p* = .004, with lower chances to give a correct response after forward than backward conditioning. *Figure 5* shows the proportions of correct responses as a function of the presentation sequence in a density plot. The main effect of US revaluation was also significant, *B* = 0.88, *SE* = 0.05, *z* = -2.52, *p* = .012. The chance for a correct response was lower for incongruent than congruent pairings. This effect was qualified by US valence, *B* = 0.73, *SE* = 0.10, *z* = -3.16, *p* = .002. Memory performance diminished for incongruent compared to congruent pairings only for positive pairings, *B* = 0.76, *SE* = 0.07, *z* = -4.02, *p* < .001, but not for negative pairings, *B* = 1.03, *SE* = 0.07, *z* = 0.45, *p* = .652. The other coefficients of the model did not reach significance (smallest *p* = .249).

Figure 5

Mean proportions of correct responses of the memory task in Experiment 3



Note. Dashed lines display mean proportions of correct responses for forward versus backward presentations.

Discussion

Experiment 3 implemented a learning procedure that should further strengthen the role of prediction in stimulus-valence learning. Despite the modifications, we again did not observe differences in US revaluation depending on the presentation sequence. However, the findings are limited to the extent that the absence of the effect did not simultaneously imply evidence for the (null) hypothesis that US revaluation effects did not differ between presentation sequences. In contrast to Experiment 2, the main effect of US revaluation was not significant, and evaluative conditioning lead to more pronounced EC effects in forward than backward conditioning. The finding is interesting as memory for CS-US pairings was reduced in forward as compared to backward conditioning. In other words, even though evaluative learning effects were stronger for forward than backward conditioning, better memory for CS-US pairings did not seem to underlie this difference. While this speaks for a representation of USs where US identities play only a minor role, presentation sequence overall does not seem to act as an environmental constraint of stimulus-valence learning in EC.

General Discussion

Past research found that evaluations acquired via evaluative conditioning (EC) are sensitive to a post-conditioning change in the valence of unconditioned stimuli (Baeyens et al., 1992; Jensen-Fielding et al., 2018; Sweldens et al., 2010; Walther et al., 2009). At the same time, there are also a few studies that reported insensitivity to US revaluation under certain conditioning procedures (Baeyens et al., 1998; Gast & Rothermund, 2011; Sweldens et al., 2010). In the present research, we argued that the way USs come to be represented in memory as a function of the learning environment influences the degree of US revaluation in EC. More specifically, learning conditions that highlight the US valence while discounting other, irrelevant features of the USs might lead to evaluative representations that hold the overall US valence (stimulus-valence learning) rather than specific US identities (stimulus-stimulus learning). Three experiments tested the pairing schedule (the number of USs presented with each CS) and the presentation sequence (forward versus backward presentations) as two candidate constraints of stimulus-valence learning.

The first experiment presented either one US (one-to-one pairings), or multiple USs (one-to-many pairings) together with CSs during conditioning. In line with the findings of Sweldens and colleagues (2010), we observed smaller degrees of US revaluation after one-to-many than one-to-one pairings. An additional memory measure further indicated that participants memory for specific CS-US pairs was better after one-to-one than one-to-many pairings. In the task, they should select the correct CS for each US. Similar results were obtained in studies from Stahl and Unkelbach (2009), for a memory measure where participants should first select the correct valence (valence memory), and then the correct US for each CS (identity memory). Both valence and identity memory were better after pairing only one than multiple USs per CS. In Sweldens et al. (2010), identity memory was also better after one-to-many than one-to-one pairings. In contrast to the assumptions of Stahl and Unkelbach (2009) that the outcome is due to differences in contingency awareness, and of Sweldens et al. (2010) that one-to-many pairings elicit misattribution of affect (Jones et al., 2009), we would argue that one-to-many pairings facilitate the extraction of US valence while omitting US identities during an abstraction process (stimulus-valence learning).

This abstraction hypothesis implies that EC is not restricted to stimulus-stimulus learning, but also involves the consolidation of evaluative experiences in less detailed, but widely applicable representations. As a result, both the degree of US revaluation and memory for CS-US pairs would diminish. Note that a similar argument was recently substantiated regarding the way attitude objects (CSs in EC) can be represented in memory. Reichmann and colleagues (2023) presented either one exemplar, or multiple exemplars of a category as CSs during conditioning. They not only observed stronger generalization of acquired attitudes to novel exemplars in the latter condition, but also reduced memory performance in a recognition memory task. The authors argued that the variability in category exemplars facilitated the formation of attitudes towards common elements across stimuli, leading to an abstract representation of CSs. Moreover, they highlighted a strong role of prediction in abstraction. In the learning condition that presented different exemplars per category, common features across exemplars can become predictive of US valence beyond individual features of CSs.

The present research also refines the work of Sweldens and colleagues (2010). In their experiments two and three, pairing schedule and simultaneity of the pairings were confounded. That is, in the one-to-many condition USs were always shown simultaneously with the CS. In the one-to-one condition, the CS-US pairings were always shown in a forward manner. The results of the present Experiment 1 show that simultaneous pairings are not required for abstraction. Moreover, our findings imply that abstraction processes offer a more parsimonious account of evaluative learning. To explain the different patterns in their two paring schedule conditions, Sweldens and colleagues (2010) argued that implicit misattribution of affect was responsible for learning effects in the one-to-many condition, whereas propositional learning was present in the one-to-one condition. Implicit misattribution of affect as it is currently conceptualized (Jones et al., 2010) does not apply to the present setting, because sequential pairings minimize the potential of source confusion.

The potential role of prediction in abstraction on the side of the evaluative meaning (USs) was tested in Experiments 2 and 3 of the present study. One-to-many pairings were either displayed in a forward (CSs before USs) or backward (USs before CSs) manner during conditioning. The initial hypothesis was that forward conditioning would facilitate the extraction of US valence, as CSs can predict US valence more accurately than specific USs across learning trials. On the other hand, CSs cannot attain such a predictive function in backward conditioning. Contrary to the hypothesis, the presentation sequence of stimuli did not influence the degree of US revaluation. US revaluation overall influenced evaluative judgements only in Experiment 2 (where three USs were paired with one CS), but not in Experiment 3 (six USs were paired with one CS). The results speak for the pairing schedule as an important environmental constraint of US revaluation in EC, while the presentation sequence seems to have little impact on the nature of representations. Similar results were obtained for one-to-one pairings, where the presentation sequence neither influenced the size of the EC effect, nor the memory for CS-US pairings (Gast et al., 2016; Kim et al., 2016). Evaluative conditioning might thus differ from other learning paradigms. Research in linguistic concept learning (Hoppe et al., 2020; Ramscar et al., 2010), Pavlovian conditioning (McSweeney & Bierley, 1984; Rescorla, 1968), and visual statistical learning (Tummeltshammer et al., 2017) identified presentation sequence as a relevant factor for what was learned. In EC, it might be the case that the CS cannot attain a predictive function, as suggested by referential accounts of evaluative conditioning (Baeyens et al., 1992; De Houwer et al., 2001), but a bidirectional link is established between the CS and US (valence) (Kim et al., 2016; Stahl & Unkelbach, 2009).

Environmental constraints of stimulus-valence learning

Given that stimulus-stimulus versus stimulus-valence learning do not seem to differ qualitatively but rather circumscribe the relative abstractness of US representations, it is an intriguing question for future research to ask what kind of learning conditions (besides the pairing schedule) might influence the relative contribution of either type of representation in learning outcomes.

Next to prediction via abstraction, one relevant aspect to consider here is the organizing function of abstraction in the storage of information in memory (Peters et al., 2017; Rosch, 1988; Taylor et al., 2015; Tenenbaum et al., 2011). Encoding different USs in terms of their shared valence reduces the amount of information that needs to be stored, while adding additional information on how the stimuli are related to one another. In learning environments where cognitive resources are scarce, stimulus-valence learning could prove particularly beneficial as learners might not be able to store individual stimulus identities. Examples are learning environments where many distractors are present (e.g., Olson & Fazio, 2001), multiple tasks need to be executed at the same time (e.g., Dedonder et al. 2010; Pleyers et al. 2009), or the number of exposures to individual CS-US pairs is low (e.g., Kattner, 2014; Kurdi & Banaji, 2019; Stuart et al., 1987). Especially the number of exposures might be a factor of interest – also keeping in mind that the manipulation of the pairing schedule in Experiment 1 was confounded with the number of CS-US presentations (repeated presentations in one-to-one, single presentations in one-to-many pairings).

Secondly, the results of the present experiments can be neatly explained by an account that considers *abstraction via comparison* (Christie & Gentner, 2010; Christie, 2022; Kurtz et al., 2013). Comparison promotes the extraction of relational structures across exemplars by making their common relational content salient (Kurtz et al., 2013). Relational structures can generally refer to non-observable, abstract attributes of stimuli (e.g., valence), beyond their surface elements (e.g., their color; Gentner, 2005). Consequently, factors that promote comparison should also promote abstraction (Gentner & Hoyos, 2017; Kurtz et al., 2013). One of these factors are common labels that co-occur with exemplars (Namy & Gentner, 2002). A common label signals learners that the exemplars share important attributes, thereby inviting comparison (Gentner & Hoyos, 2017). In evaluative conditioning, such a situation arises in one-to-many pairings, where the CS can act as a common label across USs. As a result, comparison across USs should make their common relational structure (i.e., shared valence) salient, leading to abstract representations of USs. Importantly, the results of Experiments 2 and 3 align with the notion of comparison, as comparison should yield abstract representations of USs independent of the presentation sequence of stimuli.

Another example of a factor facilitating comparison is the within-category similarity of stimuli (Gentner & Hoyos, 2017). Similarity facilitates comparison by providing a common ground for comparison via surface elements. It has been shown for EC that high similarity between CSs and USs induce comparison processes and consequently contrast effects (Alves & Imhoff, 2023). Applied to the universe of the USs, highly similar USs that vary only in their relational structure but align in their surface elements (e.g., smiling versus angry faces) should facilitate abstraction, compared to USs low in similarity. In the latter case, instructing participants to focus on similarities of stimuli (versus their differences) may enhance abstraction accordingly. For example, Corneille et al. (2009) introduced a task prior to conditioning that required participants to either focus on the similarity versus differences in the comparison of stimuli. They obtained larger EC effects in the similarity-focus than the difference-focus condition. It would be interesting to see whether the manipulation affects the degree of US revaluation and the memory for specific CS-US pairs as well, indicating higher levels of abstraction on the level of the USs. In sum, we currently consider the comparison account the most parsimonious explanation of abstraction in evaluative learning. As this account was not tested explicitly in the present experiments, future research should illuminate the cognitive processes underlying abstraction in this paradigm.

Practical implications

In many real-world contexts, it can have both desirable and harmful consequences when likes and dislikes do not change in the light of novel information. For example, marketers might have an interest in stable preferences towards a brand that are not affected when brand-associated people fall into disgrace (e.g., negative press about top managers; scandals of brand ambassadors). On the other hand, in therapeutic contexts it can be helpful to change existing attitudes by revaluating the initial learning experience. An example are negative attitudes towards the self that were acquired in a social situation negatively perceived by the individual but perceived positively by the interaction partner. Here, the revaluation of the evaluative experience would provide one way to correct for the maladaptive attitude towards the self. Implementing learning conditions that likely result in successful revaluations would be advantageous. For example, therapeutic settings could profit from concentrating on the revaluation of a single negative learning experience (in line with “one-to-one” pairings). On the other hand, considering multiple positive experiences (as in “one-to-many” pairings) should elicit more robust positive attitudes towards the self.

Conclusion

In past research, evaluative conditioning effects were found to be insensitive to US revaluation under specific learning procedures. We suggest that this is the case when aspects of the conditioning phase facilitate the extraction of US valence during learning (stimulus-valence learning), leading to abstract representations of USs. We found that abstraction is a function of the variability of USs, but not necessarily of the sequence of the pairings. The present research thereby highlights the role of cognitive-ecological factors in abstraction. It also illustrates how theorizing on abstraction in other domains can inform research into evaluative conditioning and how therewith a better integration with other learning paradigms can be achieved.

GENERAL DISCUSSION

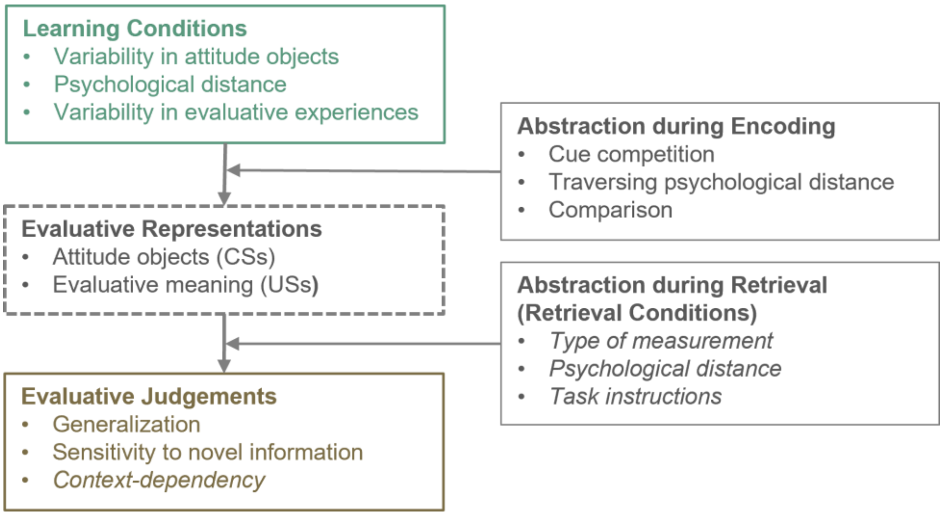
One of the central questions in cognitive psychology refers to the way information is stored and represented in memory. Mental representations can vary in their level of abstractness (Burgoon et al., 2013; Gilead et al., 2020; Reed, 2016; Tenenbaum et al., 2011; Trope & Liberman, 2010), depending on generative conditions of the learning environment (Smith, 2014). Abstraction has consequences for learning outcomes, such as the generalization and updating of knowledge (Dayan & Berridge, 2014; Ledgerwood, 2014; Ramscar et al., 2010). The present thesis studied abstraction in the domain of attitude acquisition to predict the generalization and robustness of likes and dislikes as a function of the learning environment. Three empirical projects applied this cognitive-ecological perspective to attitude acquisition via evaluative conditioning (EC), an experimental paradigm referring to the formation of likes and dislikes via the pairing of stimuli. The projects demonstrated that the abstractness of representations of attitude objects (CSs), and evaluative meaning itself (USs) can vary depending on specific aspects of the conditioning procedure. They assessed the consequences of abstraction for the generalization of acquired attitudes (Chapter 1), the degree of attitude change (Chapter 2), and the sensitivity of attitudes to changes in the environment (Chapter 3). Next to specific implications of the individual projects, they overall demonstrate the importance of considering abstraction in evaluative learning (Section 1, General Discussion). Studying the format and content of evaluative representations also adds to our understanding of attitudes in general, even though many open questions remain about the way attitudes are encoded in human memory (Section 2, General Discussion). Lastly, the theoretical framework presented here allows for implications in practical settings, namely regarding the design of interventions trying to modify (maladaptive) attitudes towards the self and others (Section 3, General Discussion). All in all, the present work shows that studying the way the human mind represents the internal and external world is of central importance also in the domain of social cognition, highlighting the necessity to consider representational formats in an integrative manner across domains of psychology (Kaup et al., 2023).

(1) Abstraction in evaluative learning

Evaluative learning, in general, can be described as a two-stage process that leads to a “change in evaluative mental representations that is due to experience” (Hütter & Rothermund, 2020, p. 2). Evaluative experiences are translated into mental representations of evaluations in a first step (the acquisition stage). Evaluative representations are then activated, retrieved, and applied during evaluative responding in a second step (the retrieval stage), resulting in biased perception, thought and action (Hütter & Rothermund, 2020). As of now, both processing steps were mainly considered in the context of automatic versus non-automatic evaluative learning (Hütter & Rothermund, 2020). However, just like automaticity, abstraction could also act both upon the acquisition and the retrieval of evaluative representations (*Figure 1*). During encoding, specific learning experiences can trigger abstraction (e.g., variability in attitude objects, variability in evaluative experiences, psychological distance). During retrieval, the nature of the task (e.g., whether it introduces a direct or indirect measure of attitudes) and the context (e.g., the specific task instructions) could trigger abstraction as well. The empirical projects presented here mainly focused on abstraction during encoding (1.1). Abstraction during retrieval was touched on by past research on attitudes, but still lacks a systematic investigation (1.2). A cognitive-ecological perspective that considers abstraction in evaluative learning jointly with the information ecology offers both heuristic and predictive value for understanding evaluative judgements, but also faces methodological as well as theoretical limitations (1.3).

Figure 1

Two-stage model of evaluative learning (Hütter & Rothermund, 2020), adapted to abstraction processes


Learning Conditions
• Variability in attitude objects
• Psychological distance
• Variability in evaluative experiences
Abstraction during Encoding
• Cue competition
• Traversing psychological distance
• Comparison
Evaluative Representations
• Attitude objects (CSs)
• Evaluative meaning (USs)
Abstraction during Retrieval (Retrieval Conditions)
• *Type of measurement*
• *Psychological distance*
• *Task instructions*
Evaluative Judgements
• Generalization
• Sensitivity to novel information
• *Context-dependency*

Note. The model describes two steps of evaluative learning, including an encoding stage that translates evaluative experiences into mental representations, and a retrieval stage that transforms mental representations in evaluative judgements (Hütter & Rothermund, 2020). From a theoretical point of view, abstraction can occur both during the encoding, and retrieval of evaluative representations. Bullet points in italics highlight a lack of research activity.

(1.1) Abstraction during encoding

During encoding, evaluative experiences are translated into mental representations of evaluations (Hütter & Rothermund, 2020). Abstraction facilitated by learning conditions acting upon encoding can influence the format and content of evaluative representations, both regarding the way attitude objects and their evaluative meaning are stored in memory. Importantly, this has consequences for evaluative judgements, as was shown in three empirical projects employing evaluative conditioning as a means of attitude acquisition.

Chapter 1 (Reichmann et al., 2023) tested the hypothesis that variability in attitude objects facilitates the generalization of likes and dislikes via the formation of abstract representations. Variability might foster the discriminative learning of cues, which can be seen as one way to acquire abstract representations that generalize widely (Ramscar et al., 2010). To manipulate the variability of attitude objects, either one exemplar of a category was presented as a CS during conditioning (*invariable* condition), or multiple exemplars were included per category (*variable* condition). CSs were Chinese characters with one component in common for all characters of a category. Our main finding was that variability increased the generalization of acquired likes and dislikes to novel stimuli participants had never seen before. Moreover, outcomes of two additional measures employed as manipulation checks indicated more abstract representations of CSs in the high variability condition. First, recognition memory performance was reduced in the variable as opposed to the invariable condition, even when controlling for the total number of CSs presented during learning. The outcome is in line with the idea that abstraction decreases the specificity of representations (Gentner & Hoyos, 2017; Liberman & Förster, 2009). Secondly, evaluations of individual components of CSs displayed a shift towards the category-defining feature in the variable condition, showing an emphasis of features invariant across instances. Again, this corresponds to a notion of abstract representations to highlight predictive attributes (Ramscar et al., 2010; Reed, 2016). Importantly, the findings are limited to the extent that the measures only allow for inferences regarding the way CSs are encoded in memory (i.e., the attitude objects), but not for the way USs are represented as a function of learning experiences. In addition, future research should test whether similar results can be obtained for categories that are defined via non-observable characteristics (e.g., such as the common social group membership of individuals) to test generalization based on abstracted features rather than perceptual similarity. Nevertheless, the project showed the theoretical value of connecting learning conditions (here: the variability of CSs) to learning outcomes (here: the generalization of likes and dislikes) by considering abstraction during learning.

Chapter 2 tested whether psychological distance, another cognitive-ecological factor of abstraction, influences the degree of attitude change. According to Construal Level Theory (Trope & Liberman, 2010), psychologically distal objects are represented in more abstract terms than proximal ones. An example provides the outgroup homogeneity effect, which shows that outgroup members are perceived as more similar to one another than ingroup members (Boldry et al. 2007; Linville et al., 1996; Park & Judd, 1990; Park & Rothbart, 1982). The outgroup homogeneity effect is in line with the idea that outgroup members are represented more abstractly in memory, as they are socially more distant than ingroup members (Hess et al., 2018). Two experiments tested whether representational differences of ingroup versus outgroup members have consequences for attitude change. Because past studies showed larger degrees of evaluative learning for distinct than similar stimuli (e.g., Glaser & Kuchenbrandt, 2017; Hütter et al., 2014), we expected larger EC effects for faces as CSs that were previously assigned to the ingroup, compared to faces assigned to the outgroup. In contrast to this initial hypothesis, EC turned out to be equally effective in changing attitudes towards individual group members, as measured on a continuous rating scale. At the same time, group membership still contributed to evaluative judgements, as faces of the outgroup were generally evaluated more negatively (positively) than faces of the ingroup in Experiment 1 (Experiment 2). Interestingly, results of a recognition memory task indicated outgroup homogeneity, as participants made more errors discriminating between seen and unseen faces when they were from the outgroup compared to the ingroup in Experiment 2. The results of the direct attitude measure therefore did not reflect a failure to manipulate the abstractness of representations via psychological distance. Attitude change measured indirectly, via the evaluative priming task (Fazio et al., 1986, 1995), was mediated by ingroup identification in Experiment 1. While attitude change did not occur for high ingroup identifiers, assimilative learning effects (standard EC effects) were observable for ingroup, and contrastive learning effects (reversed EC effects) were observable for outgroup CSs for those low in ingroup identification. This latter result was also obtained in Experiment 2. One way to interpret these findings would be to assume that participants applied distinct strategies for ingroup versus outgroup CSs, for example by comparing outgroup CSs with valent USs (i.e., the CS “is different from” the US), while informing ingroup CSs by valent USs (i.e., the CS “is similar to” the US; see also Unkelbach & Fiedler, 2016). It would be an interesting question for future research to study the role of ingroup identification, as well as the applied learning strategy in evaluative learning in intergroup contexts. Whereas social distance per se did not turn out to qualify attitude change directly, it might nevertheless prove fruitful to consider how social categories influence the information ecology (e.g., information on ingroup members is often more readily available than information on outgroup members; Denrell, 2005; Konovalova & Le Mens, 2020). Moreover, it might also be possible that psychological distance affects abstraction and thus evaluative judgement also at retrieval stage of evaluative learning (see 1.2.).

Lastly, a third line of experiments (**Chapter 3**) studied the ecological conditions of US revaluation effects in evaluative conditioning, by considering abstraction on the side of the evaluative meaning. Contrary to the general assumption that EC resembles an instance of stimulus-stimulus learning (S-S learning, e.g., Walther et al., 2009), unconditioned stimuli might also be represented in abstract ways, namely in terms of their overall valence (stimulus-valence learning). Importantly, the latter case should reduce the impact of postconditioning changes in the valence of the US on the evaluation of the CS. In a first experiment, different USs of the same valence co-occurred with one CS (“one-to-many” pairings), or the same US was presented repeatedly with one CS (“one-to-one” pairings). “One-to-many” pairings reduced US revaluation effects compared to “one-to-one” pairings, conceptually replicating the findings of Sweldens and colleagues (2010; Experiment 1). Moreover, “one-to-many” pairings diminished memory for specific CS-US pairs, in line with the idea that abstract representations of US valence no longer entail specific US identities. The findings of this experiment are compatible both with abstraction via prediction (i.e., CSs predict the US valence with a higher accuracy than specific US identities in “one-to-many” pairings) and abstraction via comparison (i.e., via structural alignment; the US valence becomes the salient element across USs in “one-to-many” pairings). Two additional experiments (Experiment 2; Experiment 3) then tested the presentation sequence as another ecological constraint of stimulus-valence learning. “One-to-many” pairings were either presented in a forward (CSs before USs) or backward (USs before CSs) manner. Presentation sequence did not influence the degree of US revaluation, showing that the pairing schedule seems to be the primary means of stimulus-valence learning. Moreover, the outcome speaks for abstraction via comparison, as abstraction via prediction should depend on the presentation sequence of stimuli (only allowing for abstraction in “forward” conditioning). Importantly, this project highlights that procedural factors in evaluative conditioning can influence the functional properties of EC effects, such as their resistance to US revaluation procedures (see also Gawronski et al., 2020; Walther et al., 2018). Thus, several hallmark characteristics of EC making it fundamentally different from other learning paradigms might actually depend on the type of representation invoked by the implemented learning procedure. For example, the finding that evaluative conditioning is not susceptible to blocking (Beckers et al., 2009; Kattner & Green, 2015), in contrast to classical conditioning (Kamin, 1969), might be attributed to the standard procedure used to study blocking in EC. Here, CSs co-occur with either positive, or negative USs during learning, and US valence is often manipulated within-subjects. This creates a clear contrast between positively and negatively paired stimuli (e.g., Kattner & Green, 2015). In a modified learning procedure that presented each CS with both negative and positive USs that were either similar or distinct for different CSs (Alves et al., 2020), blocking-like learning outcomes occurred – USs that co-occurred with other CSs were less effective in producing conditioned attitudes (McSweeney & Bierley, 1984). Another example refers to the context-dependency of attitudes acquired via evaluative conditioning versus impression formation (Gawronski et al., 2010). As further discussed below, impression formation tasks might induce more abstract representations of evaluative meaning than EC procedures, with consequences for the context-dependency of acquired preferences.

To summarize, the three empirical projects presented here provide a cognitive-ecological perspective on the generalization and robustness of likes and dislikes acquired via evaluative conditioning. By considering abstraction during learning, they allow for predictions on how learning conditions relate to evaluative judgements. Importantly, the projects provide evidence against the premise that evaluative conditioning constitutes a passive learning phenomenon that leads to the acquisition of links between stimulus identities. Instead, EC seems to involve generative processes applied by the learner (see also Corneille et al., 2009; Fiedler & Unkelbach, 2011; Sperlich & Unkelbach, 2022), and can lead to abstract representations of both the CSs, and the USs (see also Gawronski & Bodenhausen, 2018; Hütter & Tigges, 2019). Especially the latter aspect illustrates that EC should not be treated as an isolated learning phenomenon but rather considered in an integrative manner with other domains of research on learning and abstraction (e.g., relational learning; Gentner & Hoyos, 2017; concept learning; Ramscar et al., 2010; reinforcement learning; Dayan & Niv, 2008). In addition, evaluative learning paradigms apart from evaluative conditioning might benefit from taking on a similar theoretical perspective, also shedding light on how EC might differ from other ways to acquire attitudes.

One interesting example to consider here lies in the findings of Gawronski and colleagues (2020) who compared the context-dependency of attitude change via EC with the “impression formation paradigm” (Rydell & Gawronski, 2009). In the impression formation paradigm, different verbal statements of either positive or negative valence are presented with an unknown person. The paradigm thus resembles a “one-to-many” evaluative conditioning procedure, with valent verbal statements instead of valent images. Contextualized attitude change was first demonstrated with this paradigm by Rydell and Gawronski (2009). It refers to the finding that mental representations of counterattitudinal information about an attitude object are often bound to the context the information occurred in. The result are "dual" representations, with one being context-free and representing the initial attitude, and the second one being context-dependent and representing the novel attitude (Gawronski et al., 2018). The existence of these two distinct representations becomes evident in *renewal effects* (Bouton, 2004; Gawronski et al., 2010). When initial attitude acquisition takes place in context A, and counterattitudinal information is learned in context B, the initial attitude still determines evaluations in context A (*ABA renewal*) or in a novel context C (*ABC renewal*). At the same time, in context B evaluations correspond to the counterattitudinal information, demonstrating that the renewal effects were not due to a lack of learning in context B (see Gawronski et al., 2015, for a meta-analysis).

Interestingly, renewal effects were only obtained for the impression formation task, but not for evaluative conditioning. Gawronski and colleagues (2020) manipulated the context by presenting CS-US pairs against different background colors in the conditioning, counterconditioning, and testing phase. Evaluations obtained in a speeded evaluation task showed that counterconditioning reversed initially conditioned attitudes regardless of the context. However, in the impression formation task renewal effects occurred and the initial attitude continued to dominate evaluations in a novel context C (Gawronski et al., 2020; Experiment 5). Note that the two evaluative learning paradigms did not only differ in the modality of the valent material (images vs. verbal statements), but also in the pairing schedule of attitude objects and valent stimuli ("one-to-one" pairings in EC vs. "one-to-many" pairings in the impression formation task). An alternative explanation to the expectancy-violation account put forward by Gawronski and colleagues (2020) is that evaluative meaning is represented in more abstract terms in the impression formation task, making initial attitudes context-independent, and counter attitudes context-dependent.[34] However, this would mean that impression formation and evaluative conditioning differ in the pairing schedule of stimuli (see Chapter 3) instead of being functionally different. As a working hypothesis, renewal effects might be obtained in EC as well if one CS would occur with different USs of the same valence ("one-to-many" pairings), leading to an abstract representation of evaluative meaning

[34] For example, it might be the case that counterattitudinal information is stored in a context-dependent fashion for abstract representations of evaluative meaning to facilitate the integration of seemingly contradicting information (e.g., something that is overall positive can be negative in a certain context) and thus avoids the emergence of ambivalence.

and context-dependent learning during counterconditioning. The example illustrates how the investigation of abstraction processes in evaluative learning can be expanded to other evaluative learning paradigms and learning outcomes (i.e., the context-dependency of attitudes; see *Figure 1*). Moreover, it offers an interesting angle on seemingly ambiguous findings of different evaluative learning paradigms.

(1.2) Abstraction during retrieval

Next to abstraction during the acquisition of attitudes, it is also feasible that abstraction occurs during retrieval, namely when evaluative representations are translated into evaluative responses (the second stage in the two-sage model of evaluative learning; Hütter & Rothermund, 2020). Retrieval can occur via various processes, including reflective processes such as the construction of meaning or the anticipation of optimal response strategies to maximize rewards, and less deliberate processes such as spreading activation and conflict adaptation (Hütter & Rothermund, 2020). Just like abstraction at the encoding stage, abstraction during retrieval might depend in central ways on environmental aspects of the retrieval situation. Immediate consequences for evaluative responding can be expected, for example for the degree of generalization or the context-dependency of attitudes. This would make it necessary to consider abstraction also in concert with conditions during retrieval (see *Figure 1*).

An example is the type of measurement employed to assess attitudes. Attitudes can be measured directly, via self-reported evaluations of attitude objects, or indirectly, by inferring attitudes from objective performance indicators, such as participants' speed and accuracy in responding to attitude objects (Corneille & Hütter, 2020; Gawronski & Brannon, 2018). First, direct measures of attitudes can be arranged on a continuum of abstraction both regarding the way they refer to attitude objects, and the way they refer to the evaluative meaning of an attitude object. For example, measures that focus on an evaluation of a category (e.g., a social group) refer to a more abstract notion of attitude objects than measures focusing on an individual exemplar of the category (e.g., a member of a social group; Ledgerwood et al., 2020). Similarly, measures can ask for evaluations of an overall attribute (e.g., "intelligence" in romantic partners) or for an attribute expressed to a certain degree in a specific individual (e.g., liking of a specific intelligent romantic partner). The former refers to a more abstract notion of an attribute (e.g., "intelligence") than the latter one (Ledgerwood et al., 2018, 2020). In addition, self-reported evaluations of attitude objects can reflect both affective and semantic components of evaluative meaning. In "knowledge-focused" self-reports that ask participants to evaluate how positive or negative a stimulus is, evaluations are likely based on semantic knowledge, including general factual knowledge, cultural norms and stereotypes (Hamzani et al., 2019; Itkes & Kron, 2019). In contrast, in “feelings-focused” self-reports participants should report their internal feelings (e.g., how positive or negative their feelings are), which likely taps into introspective feeling states or the episodic memory of feelings one felt during a specific event (Itkes & Kron, 2019). One way to describe the difference between the two self-reports lies the abstractness of the evaluative meaning they refer to, with “feelings-focused” self-reports being more likely to reflect evaluative meaning bound to a specific event, while “knowledge-focused” self-reports likely reflect valence abstracted across instances and situations.[35]

Along direct measures, also indirect measures of attitudes can elicit abstraction to differential degrees. Most performance-based measures involve evaluative responses to exemplars but require speeded categorizations in positive versus negative valence categories (Ledgerwood et al., 2020). For example, in the evaluative priming task (EPT; Fazio et al., 1986, 1995), attitude objects serve as primes and are followed by positive or negative target words that should be categorized. The task thus likely leads to the retrieval of concrete representations of attitude objects (i.e., specific exemplars as primes), and to the retrieval of abstract representations of evaluative meaning (i.e., via the classification of target words into positive versus negative valence). In line with a notion of evaluative representations that both the attitude object and its evaluative meaning can vary in their level of abstractness independently of one another, this should not pose a problem per se. However, it might explain divergent findings between direct and indirect measures and the rather low correlations between the measures (Cameron et al., 2012; Hofmann et al., 2005). As noted by Gawronski (2019), the attitude objects a measurement refers to are often confounded with the type of measurement. For example, direct measures such as the Modern Racism Scale (McConahay, 1986) refer to attitude objects as categories, while indirect measures such as the EPT would employ Black or White faces as primes (Fazio et al., 1995). Consequentially, the conceptual correspondence between the two measures decreases and so does the correlation between them (Gawronski, 2019; Ledgerwood et al., 2020).

A second example for a potential role of abstraction during retrieval may be the mindset taken on during judgement. That is, instructions presented before evaluative judgements might lead participants to adapt a more “concrete” or “abstract” mindset. With an

[35] However, this does not imply that semantic knowledge of valence cannot be concrete (e.g., knowledge about valence can stem from a specific learning instance), or affective valence expressed in internal feeling states cannot be abstract (e.g., an affective response can occur independent of the original evaluative experience).

"abstract" mindset, participants might be more likely to retrieve representations that contain invariant features across instances, rather than specific exemplars (in line with the definition of abstraction presented by Burgoon et al., 2013). For instance, when they are instructed to imagine a hypothetical scenario (e.g., "Imagine you are a recruiter and have to evaluate applicants for a vacant position") their thinking might be more abstract than when making judgements for the here-and-now (e.g., "Evaluate the person sitting next to you."). This would be in line with the assumption of Construal Level Theory that abstraction increases with hypotheticality, one dimension of psychological distance (Trope & Liberman, 2010). Consequentially, evaluative judgements might not only be influenced by individual-level information but also reflect attitudes towards the social category an individual belongs to (Chapter 2; Kunda & Thagard, 1996; Labiouse & French, 2001; Park & Rothbart, 1982). An example provides the study of Milkman and colleagues (2012), who showed that race- and gender-based discrimination increased with increasing temporal distance.

Next to psychological distance, instructions could influence the retrieved representation via additional information on the attitude objects to be evaluated. For example, one interesting question would be whether the effects of CS variability on generalization reported in Chapter 1 can also be obtained when instructing participants before the test phase how much variability can be expected amongst CSs (see Ram et al., 2023, for an implementation of this procedure in predictive learning). Instructions alone might be sufficient for abstraction on the side of the CS that widens the applicability of learning experiences, leading to an increase in the generalization of likes and dislikes to other instances of a category. Lastly, tasks presented between attitude acquisition and evaluative responding might also manipulate the abstractness of retrieved representations. Corneille et al. (2008) had participants list as many similarities or differences between two drawings as possible before an evaluative conditioning phase, to introduce a similarity versus differences focus between stimuli. EC effects were more pronounced in the similarity- than the difference-focus condition. The task might influence evaluative judgements also when it is included after the conditioning phase. For instance, focusing on similarities could increase the generalization to novel stimuli that are similar to conditioned stimuli. Importantly, this would not only apply to perceptual similarities but also non-observable characteristics of the stimuli (e.g., their superordinate category membership), leading to abstract representations of the attitude objects.

To summarize, the cognitive-ecological perspective on evaluative judgements can also be applied to the testing situation, by jointly considering abstraction with retrieval conditions

(*Figure 1*). Examples of such retrieval conditions include the specific measurement of attitudes, instructions provided before evaluations, and tasks included after the learning phase. As of now, a systematic investigation of the ecological conditions influencing abstraction during retrieval is still missing, especially in the domain of evaluative learning. However, the theoretical approach provides the potential to shed light onto divergent findings in attitude research (e.g., for direct and indirect measures of attitudes; Gawronski, 2019; Ledgerwood et al., 2020), and again fosters the specification of boundary conditions for the generalization and updating of attitudes.

(1.3) Strengths and limitations of the present framework

The present work aimed to follow methodological recommendations as close as possible that were defined as a response to the replication crisis in psychology (Koch et al., 2018; Marsman et al., 2017; Open Science Collaboration, 2015).[36] In addition to calls for open science, pre-registrations and advanced statistical methods, the crisis was also addressed in terms of the way researchers in psychology theorize (the “theory crisis” next to the replication crisis; Fiedler et al., 2021; Oberauer & Lewandowsky, 2019; Szollosi & Donkin, 2021). This makes it necessary to evaluate the underlying theoretical framework beyond the empirical outcomes of research projects.

Here, it is important to note that the considerations of different theoretical accounts of abstraction (i.e., discriminative learning, traversing psychological distance, comparison) in the present thesis resembles discovery-oriented rather than theory-testing research (Oberauer & Lewandowsky, 2019). That is, as failures to find the predicted results do not question the theory per se, the falsifiability of the theoretical perspective is restricted. For example, in Chapter 3 we expected similar degrees of stimulus-valence learning for forward versus backward conditioning according to abstraction via prediction, but then considered abstraction via comparison to explain the observed results. Nevertheless, a theory can be considered valuable as long as it leads to new answers to old questions and generates new predictions. It is this heuristic and predictive value of theories that make them worthwhile to consider (De Houwer, 2018). The different theoretical accounts of abstraction mentioned here provide a fresh perspective on the way stimuli presented during evaluative conditioning are processed and encoded in memory. By specifying which features get retained and which ones get omitted during abstraction, they allow for predictions on the generative conditions of the

[36] For most experiments, required sample sizes were estimated via an a-priori power analyses with a minimum power of .8, hypotheses and analyses plans were pre-registered, data were analyzed with multilevel linear-mixed models or Bayesian analyses and results were replicated in additional experiments. For all studies, data sets and analysis scripts are publicly available on OSF.

learning environment that should facilitate abstraction. The joint consideration of abstraction processes and the information ecology allows one to move from discovery-oriented to theory-testing research.

Theory-testing research generates strong hypotheses that can be rejected based on empirical evidence (Oberauer & Lewandowsky, 2019). The theory leads to predictions for conditions under which a certain outcome should be observed, which means that strong evidence both in favor and against a theory can be obtained (Oberauer & Lewandowsky, 2019). For example, Chapter 1 considered abstraction via discriminative learning to derive the prediction that the variability of the attitude objects facilitates the generalization of acquired attitudes. Here, a strong link between theory and hypotheses was achieved as cue competition principles are formally specified in the Rescorla-Wagner model (Rescorla & Wagner, 1972). Participants evaluated not only novel stimuli, but also components of the CSs. The Rescorla-Wagner model makes the prediction that variability in attitude objects should result in more extreme evaluations of stimulus components that are fixed across different CSs (see Supplement A, Chapter 1). Evaluations of stimulus components were in line with the prediction of the model. Results contradicting the hypothesis would lead to a rejection of the assumption that cue competition shifts representations to the most predictive CS component. The example illustrates how a specification of the learning conditions that should facilitate abstraction makes predictions falsifiable and leads to precise theorizing about the expected learning outcomes.

Connecting intrapsychic processes to aspects of the information ecology furthermore has the advantage that the *explanas* becomes sufficiently distinct from the *explanandum* (here: evaluative judgements). According to Fiedler (2014), environmentally anchored theories of social cognition can provide both objective (i.e., precisely defined) and theoretically distal (i.e., falsifiable) explanations of social phenomena. While intrapsychic factors may contribute to learning outcomes, Fiedler (2014) suggests that one should assess and control for the superordinate influence of the environmental input first. An example are judgement biases that can arise as a natural consequence of the statistical properties of the learning environment, in interaction with a lack of metacognitive abilities of individuals to control for such sampling constraints (Fiedler, 2000, 2014). In a similar vein, one could argue that abstraction is restricted to the information ecology imposed by the experimenter. For example, while evaluative conditioning effects are generally described as sensitive to US revaluation (Baeyens et al., 1992; Jensen-Fielding et al., 2018; Sweldens et al., 2010; Walther et al., 2009), this outcome seems to depend on the EC procedure most frequently implemented (i.e., repeated presentations of the same CS-US pair). Accordingly, modifications of the conditioning procedure can reduce US revaluation effects, as evaluative representations of different format and content can be acquired (Chapter 3; Sweldens et al., 2010). This highlights the importance to consider environmental constraints of intrapsychic processes.[37]

Next to methodological considerations and strong theorizing, it is also of central importance to ensure that the variables of theoretical interest are operationalized with high divergent and discriminant validity (Fiedler et al., 2021). Here, the implementation of appropriate manipulation checks (MCs) can help to ensure the intended purpose of a manipulation. Ideally, MCs should be operationally independent of the dependent variable in the experiment and should help to rule out unwanted effects of alternative variables (Fiedler et al., 2021). Because the current work concentrated on the format and content of evaluative representations that cannot be observed directly, a methodological challenge lies in finding measures that validate the theoretical construct.

The present experiments mostly relied on memory measures to infer the content of acquired representations. Memory measures test whether details from the learning episode are retained after learning, a hallmark for the distinction between more concrete or abstract representations (Burgoon et al., 2013; Reed, 2016).[38] For example, participants should have greater difficulties distinguishing between “seen” and “unseen” stimuli when stimuli are represented more abstractly (Chapter 1, Chapter 2). Moreover, their memory for specific CS-US pairs should diminish when USs are represented in terms of their overall valence rather than specific identities (Chapter 3). At the same time, the inferences that can be gained from memory measures are limited. Memory tasks were included only after the learning phase and even after evaluative judgements, making it impossible to distinguish between abstraction during encoding versus retrieval (see *Figure 1*). The specific type of memory task likely influences the retrieved representation as well (e.g., memory tasks that require participants to select the correct US identity versus the correct US valence for each CS; Stahl & Unkelbach, 2009). In addition, the representation retrieved in evaluative judgements might differ from the

[37] Generally speaking, environmental constraints can be seen as inherent and stable properties of the information ecology (e.g., positive information is overall more similar and less diverse than negative information; positive information is more frequent than negative information; Unkelbach et al., 2019). At the same time, environmental constraints can also refer to flexible properties of an information ecology (e.g., the variability of training material, see Chapter 1; what kind of study material children in school are exposed to, Glenberg et al., 2012). While the present thesis concentrates on the latter definition, it would also be interesting to consider the former one in the context of abstraction.

[38] Note that evaluative judgements and memory performance are related to one another but still can be seen as operationally independent, as past research has shown that memory is not a necessary precondition for the emergence of EC effects (Hütter et al., 2012; Mierop et al., 2019; Walther & Nagengast, 2006; but also see Bar-Anan et al., 2010; Gast, 2018; Gast et al., 2012).

representation retrieved in the memory task. Because the experiments presented here describe a pathway of abstraction from specific instances to abstract representations, memory for specific instances might still be present after learning even though evaluative judgements could depend on abstract entities. Thus, concrete and abstract representations of learning episodes can co-exist in memory, making it hard to know which one is retrieved in the memory task or during evaluative judgements.

An important task of future research thus lies in the introduction of measures besides memory tasks that allow for inferences on the abstractness of representations (Burgoon et al., 2013). For example, one could measure the breadth or inclusiveness of categorization by asking participants to rate the belongingness of atypical exemplars to a category (Isen & Daubman, 1984). Higher ratings of belongingness indicate more abstract representations (Burgoon et al., 2013; Smith & Trope, 2006). Another example are measures that infer abstraction form language use (Maass, 1999; Semin & Fiedler, 1988; Vallacher & Wegner, 1989). For instance, the Behavior Identification Form (Vallacher & Wegner, 1989) requires participants to identify actions on either a lower, concrete level (e.g., *reading* as "following lines of print") or a higher, abstract level (e.g., *reading* as "gaining knowledge"). An abstraction score can be calculated based on the selection of participants, counting the number of higher-level actions identified (e.g., Fujita et al., 2006; Luguri et al., 2012; Smith & Trope, 2006). Alternatively, abstraction can also be inferred by analyzing participants' language use via Natural Language Processing (NLP). With NLP tools, one can estimate the change of latent variables (e.g., the valence or abstractness of language) by training a model on large language corpora and then fitting the model to text produced by participants (Charlesworth et al., 2021; Jackson et al., 2022; Kurdi & Charlesworth, 2023). Such an approach even offers a way to investigate thinking not only after, but also during learning. For instance, one conceivable application in evaluative conditioning would be to let participants describe the conditioned stimuli multiple times during the learning phase (either verbally by "thinking out loud" or by writing a short description). Via NLP, one could detect how the frequency of words referring to attitudes or emotions on different levels of abstraction changes over time (Jackson et al., 2022). To summarize, future research should employ alternative measures to increase the confidence that differences in evaluative judgements can be attributed to differences in the abstractness of acquired representations as hypothesized.

(1.4) Interim Summary

The two-stage model of evaluative learning presented by Hütter and Rothermund (2020) suggests that evaluative learning involves both an encoding, and a retrieval stage.

Accordingly, the study of abstraction in attitude acquisition should take both processing stages into account. While the empirical projects presented here mainly focused on the encoding stage, future research should also investigate the consequences of abstraction during retrieval. A cognitive-ecological perspective that considers the ecological conditions of abstraction to understand learning outcomes has the advantage to (1) offer a way to explain functional differences between evaluative learning paradigms by considering the type of representation invoked by the learning procedure, (2) inspire future research by providing a novel way of thinking about divergent findings in the literature and (3) allow for precise theorizing and the specification of falsifiable predictions. However, because evaluative representations are latent constructs than cannot be observed directly, more attention should be devoted to the development of measures that can serve as indicators for the abstractness of acquired mental representations.

(2) Theoretical implications for theories of attitudes

Research on attitudes ranges back almost a century (e.g., Allport, 1935; Bem, 1972). This long history of scientific endeavor illustrates the importance of studying the construct, but also suggests that there has been continued disagreement (Ferguson & Fukukura, 2012). In the following, I will discuss abstraction in attitude acquisition in the light of two prominent questions that were discussed frequently in the attitude literature during the last couple of years. The first question refers to the *operating principles of attitude acquisition*. Operating principles describe the quality of cognitive processes and distinguish between processes that operate in an associative versus propositional manner (De Houwer, 2018; Gawronski & Bodenhausen, 2018; Hütter, 2022). Here, it becomes evident that both accounts allow for abstraction and that a study of the format and content of evaluative representations might not require a confinement to a particular operating principle.[39] The second question refers more directly to the way *attitudes are encoded in memory,* asking whether evaluative representations mostly rely on semantic memory or also involve affective systems and instrumental memory systems that encode feedback-based reward learning (Amodio & Berg, 2018). The role of abstraction in representations stored in affective systems or instrumental memory systems remains yet to be explored. While the present work offers some theoretical implications on both questions, it also becomes clear that the study of attitudes remains a challenging but promising field of research that could foster our understanding of human learning beyond attitude acquisition.

(2.1) Associative versus propositional learning

Operating principles, in general, specify how observed stimuli are translated into mental representations, and how existing representations determine judgements and behavior (Gawronski & Bodenhausen, 2018). In the past decades, different process theories were introduced that tried to specify the operating principles of evaluative conditioning. Here, an important distinction was made between associative and propositional learning principles of attitude acquisition. Associative learning results in the formation of mental links between co-occurring stimuli, whereas propositional learning also encodes the particular relation as well

[39] Related to the distinction between associative versus propositional learning principles is the question whether attitudes are acquired in an automatic or non-automatic way (Corneille & Stahl, 2019). Whereas some theoretical models equate automatic and associative, and non-automatic and propositional processing (e.g., Gawronski & Bodenhausen, 2006; Mitchell et al., 2009), it is indeed an empirical question how both distinctions are related to one another (Hütter, 2022; Hütter & Rothermund, 2020; Gawronski & Bodenhausen, 2009). Secondly, associative versus propositional principles were also mapped onto the distinction of direct versus indirect measures of attitudes (Gawronski & Bodenhausen, 2006). Again, the way the two distinctions are related might not be as clear-cut as initially thought (Gawronski & Brannon, 2018). The following discussion will thus concentrate on propositional versus associative learning principles without making additional claims about the automaticity of processing or the type of measurement.

as the perceived truth value of the relation between stimuli (Gawronski & Brannon, 2009). The relative importance of the two operating principles varies between different theoretical models of EC. For example, the associative-propositional evaluation model (APE model; Gawronski & Bodenhausen, 2018) suggests that both processes operate simultaneously, in a functionally independent manner. Other models, like the integrated propositional model (De Houwer, 2018), propose that propositional processes alone are sufficient to explain outcomes of EC studies. In more recent years, some authors have argued that most empirical evidence in EC can be accounted for by propositional processes only (Corneille & Stahl, 2019).

Interestingly, the distinction between associative and propositional learning principles in the context of EC can be related to the long-standing debate in psychology on whether learning can be better described by association-based or cognitive (symbolic) theories (Hummel, 2010; Hummel & Holyoak, 2003; Ramscar et al., 2010; Shanks, 2010). Association-based theories assume that knowledge is represented as connections between nodes in large-scale networks that are activated via spreading activation during retrieval (Collins & Loftus, 1975; McClelland & Rumelhart, 1985; Rescorla & Wagner, 1972; see Pearce & Bouton, 2001, for a review). Associative linkages can be both excitatory or inhibitory in nature, which allows associative networks to represent rich and complex informational structures (McClelland & Rumelhart, 1985). For example, they allow for the encoding of abstract representations via the up-weighting of relevant, and the down-weighting irrelevant features (e.g., Apfelbaum & McMurray, 2011; Ramscar et al., 2010), and thus mostly align with the theoretical perspective on abstraction that takes discriminative learning into account (Chapter 1). Cognitive (symbolic) theories of learning suggest that relationships between stimuli are represented as propositional beliefs in connectionist systems (Hummel & Holyoak, 2003). Propositional beliefs in connectionist systems include explicit representations of the relations between stimuli (De Houwer, 2018; Doumas et al., 2008; Hummel & Holyoak, 2003). These explicit representations are also referred to as *predicates* that can be flexibly bound to objects or features of objects (e.g., “A *above* B”, “D *above* C”, Doumas et al., 2008). According to Doumas and colleagues (2008), the acquisition of predicates involves multiple processing steps: First, features that are fixed across instances must be detected via comparison and then need to be isolated from other properties of the objects, in order to be represented explicitly as a predicate (Doumas et al., 2008). In other words, one could say that the encoding of predicates closely resembles abstraction via comparison (Chapter 3). Thus, also cognitive theories of learning allow for representations that can be more abstract or concrete in nature.

Just like association-based or cognitive (symbolic) theories of learning, associative versus propositional learning principles considered in evaluative conditioning allow for representations of varying abstractness as well. For instance, representations that encode the co-occurrence of CSs and USs during conditioning can either be seen as an association (i.e., an excitatory link between two stimuli), or a proposition (i.e., two stimuli that are linked by the predicate "*co-occur*"; Hütter, 2022). Similarly, a representation that encodes relational information between stimuli (e.g., the CS is "an enemy", or "a friend" of the US; Fiedler & Unkelbach, 2011) might result from an associative learning process (e.g., inhibitory links between stimuli) or a propositional learning process (e.g., the encoding of the predicate "*enemy of*" or "*friend of*"). Thus, representations resulting from associative versus propositional learning processes can hardly be distinguished on any methodological basis (Hütter, 2022). This is problematic, as it implies that knowledge about operating principles in EC does not necessarily foster our understanding of the learning conditions that lead to the encoding of relational qualifiers versus mere co-occurrences in memory. For example, even though evaluative conditioning is nowadays often referred to as an instance of propositional learning (Corneille & Stahl, 2019; De Houwer, 2018), it is yet unclear under what kind of learning conditions evaluative judgements would reflect relational qualifiers, or mere co-occurrences between stimuli (but see Moran et al., 2015, for a notable exception). Yet, this knowledge would be an important one as relational qualifier can reverse an evaluation (e.g., the notion that a vaccine *prevents* a disease makes the vaccine a positive rather than negative attitude object; Hu et al., 2017).

Focusing on the format and content of evaluative representations rather than operating principles might offer a fruitful way to generate insights in this regard. For example, it would be an interesting question to ask whether an increase in the psychological distance of attitude objects to the self increases the impact of relational qualifiers, as representations of the attitude objects become more abstract. In a hypothetical scenario where a company advertises the environmental friendliness of its products that is then declared as false (e.g., in "greenwashing"; de Freitas et al., 2020), a consumer might evaluate the company positively in the here and now (thinking about a specific product of the company that co-occurred with positive information in the advertisement) and negatively when thinking about the consumption of the products in the future (taking the negative brand image of the company into account). Similar questions could be investigated in future research that focus on the environmental constraints of the encoding of relational qualifiers and truth values by considering the abstractness of evaluative representations. As discussed above, these questions can be independent of whether associative learning principles, or propositional learning principles are better suitable to explain the outcomes. A shift towards studying what information is retained or omitted throughout learning might therefore provide a fruitful alternative to the endeavor of investigating the operating principles of evaluative learning.

(2.2) Multiple memory systems of attitudes

Next to operating principles, another open yet related question refers to the specific memory systems that could underlie evaluative representations. For instance, according to the Memory Systems Model of Attitudes (Amodio & Berg, 2018), attitudes might rely on multiple memory systems that can act independently, in concert, or in competition. These include semantic and declarative memory systems (encoding conceptual associations of stimuli in brain areas such as the hippocampus and the temporal cortex), non-declarative, instrumental systems (governing feedback-based reward learning, subserved by dopaminergic activity in the striatum), and affect-based, emotional systems (underlying fear conditioning supported by the amygdala). Thus, according to Amodio and Berg (2018) attitudes are constituted of multiple components such as semantic associations as well as affective and reward-based responses. Neuropsychological studies supported the assumption, showing that patients with bilateral damage of the amygdala are still able to accurately judge rewarding versus aversive stimuli without showing a physiological response, whereas patients with damage to the hippocampus showed the opposite pattern (Bechara et al., 1995). Moreover, they demonstrated the involvement of feedback-based reward learning in the formation of attitudes in a neuroimaging study (Hackel et al., 2015). More specifically, both reward prediction errors (i.e., expected versus received outcomes of a social interaction) and generosity prediction errors (i.e., expected versus received proportion of resources shared by the social other) contributed to impression formation, leading to activation in the ventral striatum that is also involved in reward learning. Importantly, Amodio and Berg (2018; see also Hackel et al., 2015) highlighted the importance to consider affective and instrumental memory systems for the storage of evaluative representations beyond semantic memory systems. The measures typically involved in studies of attitude acquisition (i.e., self-report measures; indirect measures of attitudes) are often not informative in this regard. Thus, the research projects presented here most likely tap into attitudes that rely on semantic and declarative memory systems, rather than systems involved in the storage of affective and reward-based components of attitudes. Nevertheless, it would be interesting to consider the role of abstraction also in the context of affective and instrumental memory systems.[40]

A first example provides the question whether affective responses vary systematically with the abstractness of evaluative representations, especially regarding the way evaluative meaning is represented in memory. For instance, it seems feasible that the abstractness of representations acquired via evaluative conditioning influences the affective quality of the representation. One would need to implement physiological measures (e.g., measures of movements in facial muscles via electromyography) to study such a research question. Intuitively, concrete representations of evaluative meaning (e.g., pictorial codes of valent images co-occurring with attitude objects) could be characterized by stronger affective responses than abstract representations. However, it could also be the other way around, considering that abstract words were found to be emotionally more valanced than concrete ones (Kousta et al., 2011; Vigliocco et al., 2014). Nevertheless, affective responses as one component of attitudes might be inherently related to the modality-specificity, rather than the abstractness, of representations. Sensory qualities can range from modality-specific to modality-general (Barsalou, 2008; Kaup et al., 2023). Modality-specific representations would concern only one sensory modality (e.g., visual, auditory, tactile etc.), and are thus grounded in bodily experiences (Kaup et al. 2023). For example, fear conditioning might result in representations that are grounded in one sensory modality and therefore also elicit strong affective responses, while evaluative conditioning could lead to representations that are modality-general and elicit less physiological arousal. This could be independent of abstraction. Fear conditioning might still lead to abstract representations of the original experience, explaining why fear responses generalize relatively quickly and can result in phobias or the generalized anxiety disorder.

Next to the question of how affective responses relate to the format and content of evaluative representations, abstraction might also play a role in the context of reward learning (also referred to as reinforcement learning; Dayan & Berridge, 2014). Here, a notion of abstraction can be found in the distinction between model-free and model-based forms of reinforcement learning (Dayan & Niv, 2008; Doll et al., 2012). Both forms are formally specified and differ in their neural underpinnings (Dayan & Berridge, 2014). Model-free

[40] Even though Amodio and Berg (2018) postulate that different memory systems underly the different components of attitudes, it is important to note that this is not a theoretical necessity. Rather, it could also be possible that evaluative representations rely on a single memory system, but different neural circuits are activated during learning or retrieval. The different components of attitudes might thus help to define the specific format and content of evaluative representations, both in terms of the attitude object and the evaluative meaning linked to the attitude object.

reinforcement learning results in estimates of long-run values of actions. Model-based reinforcement learning operates on representations of the environment to make predictions about future outcome values (Dayan & Berridge, 2014; Dayan & Niv, 2008). With that, model-free versus model-based forms of reinforcement learning reflect the concrete-to-abstract dimension of mental representations (Gilead et al., 2020). One path of future research lies in studying both forms of learning in the context of attitude acquisition (Hackel et al., 2015, 2019; Kurdi et al., 2019). For example, one could investigate the learning conditions that influence the differential involvement of model-based versus model-free forms in attitude acquisition. Such a research program would not only demonstrate the role of instrumental memory systems in evaluative learning, but also offer a way to model abstraction in attitude acquisition in computational terms.

To summarize, Amodio and Berg (2019) highlighted that attitudes are not restricted to semantic associations but also involve affective responses as well as reward learning. The present research can only make a little contribution regarding the latter two components of attitudes, as appropriate measures were not included in the experiments. Nevertheless, it would be interesting to see how abstraction is related to physiological arousal, and to test the relative contribution of model-free (more concrete) and model-based (more abstract) forms of reinforcement learning depending on the learning environment. In the terms of Marr (1982), this would foster our understanding of attitudes also on an algorithmic level (in a computational specification of the theoretical models), and an implementational level (how the computational specifications are physically realized in the brain).

(3) Practical implications for interventions targeting attitude change

The present work concentrated on the generalization and robustness of likes and dislikes as the main learning outcomes of interest. Both aspects play a role in practical settings, for example for the design of interventions targeting attitude change. These interventions aim to modify (maladaptive) attitudes in order to change behavior associated with the attitude. Examples are interventions trying to facilitate health behavior (e.g., by inducing negative attitudes towards smoking; Sherman et al., 2003; or by inducing positive attitudes towards healthy food items; Bui & Fazio, 2016). Other interventions rely on attitude change to reduce prejudice and discrimination (Kurdi & Charlesworth, 2023; Paluck et al., 2021), or to improve interpersonal relationships (Li et al., 2021). Such interventions would prove most effective if existing attitudes would be sensitive to novel information and if the learning outcomes generalize to novel situations and instances of a category. For example, negative attitudes acquired towards one brand of cigarettes should generalize to cigarettes of other brands as well if the intervention should have relevant impact. Learning conditions such as the variability of training objects could be one way to facilitate the generalizability of learning outcomes by leading to an acquisition of attitudes on the category-level (Chapter 1; Glaser & Kuchenbrandt, 2017). Moreover, especially negative experiences seem to widen generalization (see Chapter 1; Schechtman et al., 2010), which implies that it might be more effective to induce negative attitudes (e.g., towards unhealthy food items) than positive attitudes (e.g., towards healthy food items). The effectiveness of interventions also depends on the observable degree of attitude change. Here, it would be particularly important to identify characteristics in the learning environment that facilitate or diminish attitude change. One factor could be the psychological distance of attitude objects to the learner, with greater degrees of attitude change when an attitude object is psychologically proximal to oneself. However, the present empirical investigation did not confirm this hypothesis (Chapter 2).

At this point, it is important to note that the practical implications of the present work are limited to the extent that the empirical findings were obtained in strongly controlled experimental environments. The evaluative conditioning paradigm, in general, is often criticized for a lack of external validity (Moran et al., 2023). Likewise, learning outcomes were observed on the level of attitudes, but without taking behavioral outcomes into account – and attitudes do not necessarily translate into behavior as one would intuitively expect (Fazio, 1986; Wicker, 1969). That is, the behavioral expression of evaluative representations might differ substantially from their application in evaluative judgements. In addition, attitudes were only measured directly after conditioning phases, making it hard to tell how learning outcomes would develop after time has passed. Yet, the investigations can be seen as a starting point for understanding cognitive-ecological conditions that might influence the effectiveness of interventions. Considering such cognitive-ecological conditions might also help to understand divergent findings of the outcomes of interventions. For example, in their comparative study of interventions to reduce implicit racial preferences, Lai and colleagues (2014) identified interventions of the category “vivid counterstereotypic scenario” as more effective than interventions of the category “appeals to egalitarian values”. One major difference between the interventions lies in their focus on concrete exemplars (i.e., counterstereotypical exemplars are more concrete than the induction of an egalitarian mindset), one aspect that might facilitate the effect of the intervention. For example, concrete exemplars may help to reduce negative overgeneralizations.[41]

Next to the choice of procedural parameters in the design of interventions, one might also ask whether it could be more effective to change how attitude objects and their evaluative meaning are encoded in memory, rather than changing evaluative meaning per se. A prominent example are attitudes towards the self. Different mental disorders are characterized by maladaptive attitudes towards the self that emerge from overgeneralizations of negative experiences (Beck, 1963; Ganellen, 1988; Raes et al., 2023; Thew et al., 2017; Van Den Heuvel et al., 2012). Beck’s cognitive model of depression postulates that the overgeneralization of negative events to negative evaluations of the self is common in depression (Beck et al., 1979). Overgeneralization also plays a role in eating disorders (Thew et al., 2017). Next to inducing positive attitudes towards the self in depression (e.g., Grumm et al., 2009) and eating disorders (e.g., Aspen et al., 2015), one way to circumvent overgeneralization could be to modify the representation of attitude objects and their evaluative meaning in a way that they become more concrete. In other words, the intervention would aim to attach (negative) evaluative meaning to concrete behavioral instances of the self rather than the self as a whole. Promising results were reported for a “concreteness training” in dysphoric individuals that introduced training of concrete processing via a series of questions (e.g., asking participants what they could see, or what was specific about the context of an event; Watkins et al., 2009).

Taking on a cognitive-ecological perspective as presented here, an avenue of future research lies in studying how overgeneralization could be reduced by training procedures that

[41] An alternative explanation for the differential effectiveness of the approaches lies in the correspondence between training stimuli and stimuli appearing during testing (i.e., the race-IAT in Lai et al., 2014).

alter the abstractness of evaluative representations. For example, reducing overgeneralized negative attitudes towards the self might require representations of evaluative meaning that are again tied to specific negative evaluative experiences. A therapeutic intervention could thus target the retrieval of one specific negative evaluative experience that might explain one's negative feeling towards the self (e.g., one situation in which one's own behavior co-occurred with negative feedback), while asking for the retrieval of numerous positive evaluative experiences (e.g., multiple situations in which one's behavior co-occurred with positive feedback). Such a procedure would resemble the pairing schedule of attitude objects and evaluative experiences as investigated in Chapter 3, and should lead to a more concrete representation of evaluative meaning in terms of negative valence (reducing overgeneralization), and a more abstract representation of evaluative meaning in terms of positive valence (increasing the overgeneralization of positive experiences).

To summarize, practical implications for the design of interventions can be derived from the studies presented here. Nevertheless, the applicability of the findings to real-world contexts still needs to be tested empirically. However, in the light of the importance to develop effective interventions both in social and therapeutic contexts, the theoretical approach presented here might still inspire future research also outside the domain of social cognition.

CONCLUSION

The present thesis focused on abstraction in attitude acquisition, to provide a cognitive-ecological perspective on the generalization and robustness of likes and dislikes. Three empirical projects studied learning conditions that facilitate abstraction and thereby influence evaluative judgements. They illustrate how the theoretical framework presented here can be used to specify boundary conditions of learning outcomes and can lead to an integration of findings from research on abstraction with research on attitude acquisition. The framework can be extended to other evaluative learning paradigms beyond evaluative conditioning, as well as to learning outcomes besides the generalization and robustness of preferences. Furthermore, an interesting avenue for future research lies in the study of abstraction during retrieval besides abstraction during encoding. It thus provides a flexible yet fruitful approach to understand attitude acquisition and learning more generally. On a broader level, studying the format and content of mental representations in different domains of psychology can deepen our understanding of human behavior and the human mind (Kaup et al., 2023).

REFERENCES

Ackerman, J. M., Shapiro, J. R., Neuberg, S. L., Kenrick, D. T., Becker, D. V., Griskevicius, V., Maner, J. K., & Schaller, M. (2006). They all look the same to me (unless they're angry): From out-group homogeneity to out-group heterogeneity. *Psychological Science*, *17*(10), 836–840. https://doi.org/10.1111/j.1467-9280.2006.01790.x

Allport, G. W. (1935). Attitudes. In C. Murchison (Eds.), *Handbook of social psychology* (pp. 798–844). Worcester, MA: Clark University Press.

Alves, H., Högden, F., Gast, A., Aust, F., & Unkelbach, C. (2020). Attitudes from mere co-occurrences are guided by differentiation. *Journal of Personality and Social Psychology*, *119*(3), 560–581. https://doi.org/10.1037/pspa0000193

Alves, H., & Imhoff, R. (2023). Evaluative Context and Conditioning Effects Among Same and Different Objects. *Journal of Personality and Social Psychology*, *124*(4), 735–753. https://doi.org/10.1037/pspa0000323

Amodio, D. M., & Berg, J. J. (2018). Toward a Multiple Memory Systems Model of Attitudes and Social Cognition. *Psychological Inquiry*, *29*(1), 14–19. https://doi.org/10.1080/1047840X.2018.1435620

Apfelbaum, K. S., & McMurray, B. (2011). Using variability to guide dimensional weighting: Associative mechanisms in early word learning. *Cognitive Science*, *35*(6), 1105–1138. https://doi.org/10.1111/j.1551-6709.2011.01181.x

Aspen, V., Martijn, C., Alleva, J. M., Nagel, J., Perret, C., Purvis, C., Saekow, J., Lock, J., & Taylor, C. B. (2015). Decreasing body dissatisfaction using a brief conditioning intervention. *Behaviour Research and Therapy*, *69*, 93–99. https://doi.org/10.1016/j.brat.2015.04.003

Baeyens, F., Eelen, P., Crombez, G., & van den Bergh, O. (1992). Human evaluative conditioning: Acquisition trials, presentation schedule, evaluative style and contingency awareness. *Behaviour Research and Therapy*, *30*(2), 133–142. https://doi.org/10.1016/0005-7967(92)90136-5

Baeyens, F., Eelen, P., Van den Bergh, O., & Crombez, G. (1992). The content of learning in human evaluative conditioning: Acquired valence is sensitive to US-revaluation. *Learning and Motivation*, *23*(2), 200–224.

https://doi.org/10.1016/0023-9690(92)90018-H

Baeyens, F., Hermans, D., & Eelen, P. (1993). The role of CS-US contingency in human evaluative conditioning. *Behaviour Research and Therapy*, *31*(8), 731-737. https://doi.org/10.1016/0005-7967(93)90003-D

Bar-Anan, Y., De Houwer, J., & Nosek, B. A. (2010). Evaluative conditioning and conscious knowledge of contingencies: A correlational investigation with large samples. *Quarterly Journal of Experimental Psychology, 63*(12), 2313–2335. https://doi.org/10.1080/17470211003802442

Barsalou, L. W. (2003). Abstraction in perceptual symbol systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *358*(1435), 1177–1187. https://doi.org/10.1098/rstb.2003.1319

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59*, 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bechara, A., Tranel, D., Damasio, H., Adolphs, R., Rockland, C., Damasio, A. R., & Keith, L. B. (1995). Double dissociation of conditioning and declarative knowledge relative to the amygdala and hippocampus in humans. *Science, 269*(5227), 1115–1118. https://www.jstor.org/stable/2888057

Beck, A. T. (1963). Thinking and depression I: Idiosyncratic content and cognitive distortions. *Archives of General Psychiatry, 9*, 324–333. https://doi.org/10.1001/archpsyc.1963.01720160014002

Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy of depression.* Guilford Press.

Beckers, T., De Vicq, P., & Baeyens, F. (2009). Evaluative conditioning is insensitive to blocking. *Psychologica Belgica*, *49*(1), 41–57. https://doi.org/10.5334/pb-49-1-41

Behm, J. E., Edmonds, D. A., Harmon, J. P., & Ives, A. R. (2013). Multilevel statistical models and the analysis of experimental data. *Ecology*, *94*(7), 1479-1486.

Bem, D. J. (1972). Self-perception theory. *Advances in Experimental Social Psychology*, *6*, 2–

57. https://doi.org/10.5040/9781492595779.ch-006

Bettencourt, B. A., Dill, K. E., Greathouse, S. A., Charlton, K., & Mulholland, A. (1997). Evaluations of ingroup and outgroup members: The role of category-based expectancy violation. *Journal of Experimental Social Psychology, 33*(3), 244–275. https://doi.org/10.1006/jesp.1996.1323

Bettencourt, B. A., Manning, M., Molix, L., Schlegel, R., Eidelman, S., & Biernat, M. (2016). Explaining extremity in evaluation of group members: Meta-analytic tests of three theories. *Personality and Social Psychology Review, 20*(1), 49–74. https://doi.org/10.1177/1088868315574461

Bodenhausen, G. V., Kang, S. K., & Peery, D. (2012). Social categorization and the perception of social groups. In S. T. Fiske, & C. N. Macrae (Eds.), *The SAGE handbook of social cognition* (pp. 311-329). SAGE Publications. https://doi.org/10.4135/9781446247631.n16

Boldry, J. G., Gaertner, L., & Quinn, J. (2007). Measuring the measures: A meta-analytic investigation of the measures of outgroup homogeneity. *Group Processes and Intergroup Relations*, *10*(2), 157–178. https://doi.org/10.1177/1368430207075153

Boldry, J. G., & Kashy, D. A. (1999). Intergroup perception in naturally occurring groups of differential status: A social relations perspective. *Journal of Personality and Social Psychology*, *77*(6), 1200–1212. https://doi.org/10.1037//0022-3514.77.6.1200

Bouton, M. E. (1994). Context, ambiguity, and classical conditioning. *Current Directions in Psychological Science*, *3*(2), 49–53. https://doi.org/10.1111/1467-8721.ep10769943

Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learning and Memory, 11*(5), 485–494. https://doi.org/10.1101/lm.78804

Bowman, C. R., & Zeithamova, D. (2020). Training set coherence and set size effects on concept generalization and recognition. *Journal of Experimental Psychology: Learning Memory and Cognition, 46*(8), 1442–1464. https://doi.org/10.1037/xlm0000824

Branscombe, N. R., Wann, D. L., Noel, J. G. N., & Coleman, J. (1993). In-group or out-group extremity: Importance of the threatened social identity. *Personality and Social Psychology Bulletin, 19*(4), 381–388. https://doi.org/10.1177/0146167293194003

Branscombe, N., & Wann, D. L. (1994). Collective self-esteem consequences of outgroup derogation when a valued social identity is on trial. *European Journal of Social Psychology, 24,* 641–657. http://doi.wiley.com/10.1002/ejsp.2420240603

Brigham, J. C., & Barkowitz, P. (1978). Do “They all look alike?” The effect of race, sex, experience, and attitudes on the ability to recognize faces. *Journal of Applied Social Psychology, 8*(4), 306–318. https://doi.org/10.1111/j.1559-1816.1978.tb00786.x

Bui, E. T., & Fazio, R. H. (2016). Generalization of evaluative conditioning toward foods: Increasing sensitivity to health in eating intentions. *Health Psychology, 35*(8), 852–855. https://doi.org/10.1037/hea0000339

Burgoon, E. M., Henderson, M. D., & Markman, A. B. (2013). There are many ways to see the forest for the trees: A tour guide for abstraction. *Perspectives on Psychological Science, 8*(5), 501–520. https://doi.org/10.1177/1745691613497964

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Bürkner, P. C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *R Journal*, *10*(1), 395–411. https://doi.org/10.32614/rj-2018-017

Calanchini, J., Gonsalkorale, K., Sherman, J. W., & Klauer, K. C. (2013). Counter-prejudicial training reduces activation of biased associations and enhances response monitoring. *European Journal of Social Psychology*, *43*(5), 321–325. https://doi.org/10.1002/ejsp.1941

Calanchini, J., Lai, C. K., & Klauer, K. C. (2020). Reducing implicit racial preferences: III. A process-level examination of changes in implicit preferences. *Journal of Personality and Social Psychology, 5*(1). https://doi.org/http://dx.doi.org/10.1037/pspi0000339

Cameron, C. D., Brown-Iannuzzi, J. L., & Payne, B. K. (2012). Sequential priming measures of implicit social cognition: A meta-analysis of associations with behavior and explicit attitudes. *Personality and Social Psychology Review, 16*(4), 330–350. https://doi.org/10.1177/1088868312440047

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1). https://doi.org/10.18637/jss.v076.i01

Chance, J., Goldstein, A. G., & McBride, L. (1975). Differential Experience And Recognition

Memory For Faces. *The Journal of Social Psychology, 6*(7), 65–94. https://doi.org/10.1080/00224545.1975.9923344

Charlesworth, T. E., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science*, *32*(2), 218-240. https://doi.org/10.1177/0956797620963619

Christie, S. (2022). Why play equals learning: Comparison as a learning mechanism in play. *Infant and Child Development*, *31*(1), 1–8. https://doi.org/10.1002/icd.2285

Christie, S., & Gentner, D. (2010). Where hypotheses come from: Learning new relations by structural alignment. *Journal of Cognition and Development*, *11*(3), 356–373. https://doi.org/10.1080/15248371003700015

Christie, S., & Gentner, D. (2014). Language helps children succeed on a classic analogy task. *Cognitive Science*, *38*(2), 383–397. https://doi.org/10.1111/cogs.12099

Clopper, C. G., & Pisoni, D. B. (2004). Effects of talker variability on perceptual learning of dialects. *Language and Speech*, *47*(3), 207–239. https://doi.org/10.1177/00238309040470030101

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

Corneille, O., Yzerbyt, V. Y., Pleyers, G., & Mussweiler, T. (2009). Beyond awareness and resources: Evaluative conditioning may be sensitive to processing goals. *Journal of Experimental Social Psychology*, *45*(1), 279–282. https://doi.org/10.1016/j.jesp.2008.08.020

Corneille, O., & Hütter, M. (2020). Implicit? What do you mean? A comprehensive review of the delusive implicitness construct in attitude research. *Personality and Social Psychology Review, 24*(3), 212–232. https://doi.org/10.1177/1088868320911325

Corneille, O., & Stahl, C. (2019). Associative attitude learning: A closer look at evidence and how it relates to attitude models. *Personality and Social Psychology Review, 23*(2), 161–189. https://doi.org/10.1177/1088868318763261

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science, 12*(2), 163–170.

https://doi.org/10.1111/1467-9280.00328

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711. https://doi.org/10.1038/nn1560

Dayan, P., & Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. *Cognitive, Affective and Behavioral Neuroscience*, *14*(2), 473–492. https://doi.org/10.3758/s13415-014-0277-8

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology*, *18*(2), 185–196. https://doi.org/10.1016/j.conb.2008.08.003

De Freitas Netto, S. V., Sobral, M. F. F., Ribeiro, A. R. B., & Soares, G. R. da L. (2020). Concepts and forms of greenwashing: a systematic review. *Environmental Sciences Europe*, *32*(1). https://doi.org/10.1186/s12302-020-0300-3

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. https://doi.org/10.1177/1745691611400238

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

De Houwer, J., Thomas, S., & Baeyens, F. (2001). Associative learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. *Psychological Bulletin, 127*(6). https://doi.org/10.1037//D033-29O9.127.6.853

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods, 47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

DeBruine, L. M., & Barr, D. J. (2021). Understanding mixed-effects models through data simulation. *Advances in Methods and Practices in Psychological Science, 4*(1). https://doi.org/10.1177/2515245920965119

Dedonder, J., Corneille, O., Yzerbyt, V., & Kuppens, T. (2010). Evaluative conditioning of high-novelty stimuli does not seem to be based on an automatic form of associative learning. *Journal of Experimental Social Psychology*, *46*(6), 1118–1121.

https://doi.org/10.1016/j.jesp.2010.06.004

Denrell, J. (2005). Why most people disapprove of me: Experience sampling in impression formation. *Psychological Review, 112*(4), 951–978. https://doi.org/10.1037/0033-295X.112.4.951

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*(781), 1-17. https://doi.org/10.3389/fpsyg.2014.00781

Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, *22*(6), 1075–1081. https://doi.org/10.1016/j.conb.2012.08.003

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*(1), 1–43. https://doi.org/10.1037/0033-295X.115.1.1

Dovidio, J. F., & Gaertner, S. L. (1999). Reducing prejudice: Combating intergroup biases. *Current Directions in Psychological Science*, *8*(4), 101–105. https://doi.org/10.1111/1467-8721.00024

Dwyer, D. M., Jarratt, F., & Dick, K. (2007). Evaluative conditioning with foods as CSs and body shape as USs: No evidence for sex differences, extinction, or overshadowing. *Cognition and Emotion*, *21*(2), 281–299. https://doi.org/10.1080/02699930600551592

Eagly, A. H., & Chaiken, S. (1995). Attitude strength, attitude structure, and resistance to change. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 413-432). Mahwah, NJ: Erlbaum.

Eagly, A. H., & Chaiken, S. (2007). The advantages of an inclusive definition of attitude. *Social Cognition*, *25*(5), 582–602. https://doi.org/10.1521/soco.2007.25.5.582

Estes, W. K., & Burke, C. J. (1953). A theory of stimulus variability in learning. *Psychological Review*, *60*(4), 276–286. https://doi.org/10.1037/h0055775

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical. *Behavioral Research Methods, 39*(2), 175–191. https://doi.org/10.3758/bf03193146

Fazio, R. H. (2001). On the automatic activation of associated evaluations: An overview. *Cognition and Emotion*, *15*(2), 115–141. https://doi.org/10.1080/02699930125908

Fazio, R. H., Eiser, J. R., & Shook, N. J. (2004). Attitude formation through exploration: Valence asymmetries. *Journal of Personality and Social Psychology, 87*(3), 293–311. https://doi.org/10.1037/0022-3514.87.3.293

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology, 69*(6), 1013–1027. https://doi.org/10.1037/0022-3514.69.6.1013

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology, 54*, 297–327. https://doi.org/10.1146/annurev.psych.54.101601.145225

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, *50*(2), 229–238. https://doi.org/10.1037/0022-3514.50.2.229

Fazio, R. H. (1986). How do attitudes guide behavior? In R. M. Sorrentino, & E.T. Higgins (Eds.), *The Handbook of Motivation and Cognition Foundations of Social Behavior* (pp. 204–243). Guilford Press. http://psycnet.apa.org/psycinfo/1986-98550-007

Fazio, R. H, Cunningham, W., Han, A., Jefferis, V., & Jones, C. (2007). Attitudes as object-evaluation associations of varying strength. *Social Cognition, 25*(5), 603-637. https://doi.org/10.1521/soco.2007.25.5.603

FeldmanHall, O., Dunsmoor, J. E., Tompary, A., Hunter, L. E., Todorov, A., & Phelps, E. A. (2018). Stimulus generalization as a mechanism for learning to trust. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(7), E1690–E1697. https://doi.org/10.1073/pnas.1715227115

Ferguson, M. J., & Fukukura, J. (2012). Likes and dislikes: A social cognitive perspective on attitudes. In S. T. Fiske & N. C. Macrae (Eds.), *The SAGE Handbook of Social Cognition* (pp. 165-186). SAGE Publications Ltd.

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review*, *107*(4), 659–676. https://doi.org/10.1037//0033-295x.107.4.659

Fiedler, K. (2014). From intrapsychic to ecological theories in social psychology: Outlines of a functional theory approach. *European Journal of Social Psychology*, *44*(7), 657–670.

https://doi.org/10.1002/ejsp.2069

Fiedler, K., McCaughey, L., & Prager, J. (2021). Quo vadis, methodology? The key role of manipulation checks for validity control and quality of science. *Perspectives on Psychological Science, 16*(4), 816–826. https://doi.org/10.1177/1745691620970602

Fiedler, K., & Unkelbach, C. (2011). Evaluative conditioning depends on higher order encoding processes. *Cognition and Emotion*, *25*(4), 639–656. https://doi.org/10.1080/02699931.2010.513497

Fitzgerald, C., Martin, A., Berner, D., & Hurst, S. (2019). Interventions designed to reduce implicit prejudices and implicit stereotypes in real world contexts: A systematic review. *BMC Psychology*, *7*(1). https://doi.org/10.1186/s40359-019-0299-7

Förderer, S., & Unkelbach, C. (2012). Hating the cute kitten or loving the aggressive pit-bull: EC effects depend on CS-US relations. *Cognition and Emotion*, *26*(3), 534–540. https://doi.org/10.1080/02699931.2011.588687

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in Simple Bargaining Experiments. *Games and Economic Behavior, 6*(3), 347–369. https://doi.org/10.1006/game.1994.1021

French, A. R., Franz, T. M., Phelan, L. L., & Blaine, B. E. (2013). Reducing Muslim/Arab stereotypes through evaluative conditioning. *Journal of Social Psychology*, *153*(1), 6–9. https://doi.org/10.1080/00224545.2012.706242

Fujita, K., Henderson, M. D., Eng, J., Trope, Y., & Liberman, N. (2006). Spatial distance and mental construal of social events. *Psychological Science*, *17*(4), 278–282. https://doi.org/10.1111/j.1467-9280.2006.01698.x

Galdi, S., Arcuri, L., & Gawronski, B. (2008). Automatic mental associations predict future choices of undecided decision-makers. *Science, 321*(5892), 1100–1100. https://doi.org/10.1126/science.1155998

Ganellen, R. J. (1988). Specificity of attributions and overgeneralization in depression and anxiety. *Journal of Abnormal Psychology, 97*(1), 83-86. https://doi.org/10.1037/0021-843X.97.1.83

Garagnani, M., Kirilina, E., & Pulvermüller, F. (2021). Semantic grounding of novel spoken words in the primary visual cortex. *Frontiers in Human Neuroscience, 15*(581847), 1-16.

https://doi.org/10.3389/fnhum.2021.581847

Gast, A. (2018). A declarative memory model of evaluative conditioning. *Social Psychological Bulletin, 13*(3), 1-23. https://doi.org/10.5964/spb.v13i3.28590

Gast, A., De Houwer, J., & De Schryver, M. (2012). Evaluative conditioning can be modulated by memory of the CS-US pairings at the time of testing. *Learning and Motivation*, *43*(3), 116–126. https://doi.org/10.1016/j.lmot.2012.06.001

Gast, A., Langer, S., & Sengewald, M. A. (2016). Evaluative conditioning increases with temporal contiguity. The influence of stimulus order and stimulus interval on evaluative conditioning. *Acta Psychologica*, *170*, 177–185. https://doi.org/10.1016/j.actpsy.2016.07.002

Gast, A., & Rothermund, K. (2011). I like it because I said that I like it: Evaluative conditioning effects can be based on stimulus-response learning. *Journal of Experimental Psychology: Animal Behavior Processes, 37*(4), 466–476. https://doi.org/10.1037/a0023077

Gawronski, B. (2019). Six lessons for a cogent science of implicit bias and its criticism. *Perspectives on Psychological Science, 14*(4), 574–595. https://doi.org/10.1177/1745691619826015

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*(5), 692–731. https://doi.org/10.1037/0033-2909.132.5.692

Gawronski, B., & Bodenhausen, G. V. (2009). Operating principles versus operating conditions in the distinction between associative and propositional processes. *Behavioral and Brain Sciences, 32*(2), 207–208. https://doi.org/10.1017/S0140525X09000958

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

Gawronski, B., & Brannon, S. M. (2018). Attitudes and the implicit-explicit dualism. In D. Albarracín, & B. T. Johnson (Eds.), *The Handbook of Attitudes* (2nd edition, pp. 158-196). New York: Taylor & Francis.

Gawronski, B., Brannon, S. M., Blask, K., & Walther, E. (2020). Exploring the contextual renewal of conditioned attitudes after counterconditioning. *Social Cognition*, *39*(4), 287-323. https://doi.org/10.1521/SOCO.2020.38.4.287

Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis, & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd edition, pp. 283-310). New York, NY: Cambridge University Press.

Gawronski, B., Hu, X., Rydell, R. J., Vervliet, B., & De Houwer, J. (2015). Generalization versus contextualization in automatic evaluation revisited: A meta-analysis of successful and failed replications. *Journal of Experimental Psychology: General*, *144*(4), e50–e64. https://doi.org/10.1037/xge0000079

Gawronski, B., Rydell, R. J., De Houwer, J., Brannon, S. M., Ye, Y., Vervliet, B., & Hu, X. (2018). Contextualized attitude change. *Advances in Experimental Social Psychology*, *57*, 1–52. https://doi.org/10.1016/bs.aesp.2017.06.001

Gawronski, B., Rydell, R. J., Vervliet, B., & De Houwer, J. (2010). Generalization versus contextualization in automatic evaluation. *Journal of Experimental Psychology: General*, *139*(4), 683–701. https://doi.org/10.1037/a0020315

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*, 155–170. https://doi.org/https://doi.org/10.1016/s0364-0213(83)80009-3

Gentner, D. (2005). The development of relational category knowledge. In L. Gershkoff-Stowe, & D. H. Rakison (Eds.), *Building object categories in developmental time* (pp. 245-275). Taylor & Francis. https://doi.org/10.4324/9781410612908

Gentner, D., & Hoyos, C. (2017). Analogy and abstraction. *Topics in Cognitive Science*, *9*(3), 672–693. https://doi.org/10.1111/tops.12278

Gentner, D., & Markman, A. B. (1997). Structure Mapping in analogy and similarity. *American Psychologist*, *52*(1), 45–56. https://doi.org/10.1037/0003-066X.52.1.45

Gentner, D., & Smith, L. A. (2013). Analogical learning and reasoning. In D. Reisberg (Eds.), *The oxford handbook of cognitive psychology* (pp. 668-681). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195376746.013.0042

Gibson, B. (2008). Can evaluative conditioning change attitudes toward mature brands? New evidence from the implicit association test. *Journal of Consumer Research*, *35*(1), 178–

188. https://doi.org/10.1086/527341

Gilead, M., Liberman, N., & Maril, A. (2014). From mind to matter: Neural correlates of abstract and concrete mindsets. *Social Cognitive and Affective Neuroscience*, *9*(5), 638–645. https://doi.org/10.1093/scan/nst031

Gilead, M., Trope, Y., & Liberman, N. (2020). Above and beyond the concrete: The diverse representational substrates of the predictive brain. *Behavioral and Brain Sciences, 43*. https://doi.org/10.1017/S0140525X19002000

Gilmour, J. (2015). Formation of stereotypes. *Behavioural Sciences Undergraduate Journal, 2*(1), 67–73. https://doi.org/10.29173/bsuj307

Glaser, T., & Kuchenbrandt, D. (2017). Generalization effects in evaluative conditioning: Evidence for attitude transfer effects from single exemplars to social categories. *Frontiers in Psychology*, *8*(103), 1–16. https://doi.org/10.3389/fpsyg.2017.00103

Glenberg, A., Willford, J., Gibson, B., Goldberg, A., & Zhu, X. (2012). Improving reading to improve math. *Scientific Studies of Reading*, *16*(4), 316–340. https://doi.org/10.1080/10888438.2011.564245

Green, P., & Macleod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Gries, P. H., & Su, J. (2013). Taiwanese views of China and the world: Party identification, ethnicity, and cross-strait relations. *Japanese Journal of Political Science*, *14*(1), 73–96. https://doi.org/10.1017/S1468109912000357

Grumm, M., Nestler, S., & Collani, G. von. (2009). Changing explicit and implicit attitudes: The case of self-esteem. *Journal of Experimental Social Psychology*, *45*(2), 327–335. https://doi.org/10.1016/j.jesp.2008.10.006

Guala, F., & Mittone, L. (2010). Paradigmatic experiments: The dictator game. *Journal of Socio-Economics*, *39*(5), 578–584. https://doi.org/10.1016/j.socec.2009.05.007

Hackel, L. M., Doll, B. B., & Amodio, D. M. (2015). Instrumental learning of traits versus rewards: Dissociable neural correlates and effects on choice. *Nature Neuroscience, 18*(9), 1233–1235. https://doi.org/10.1038/nn.4080

Hackel, L. M., Berg, J. J., Lindström, B. R., & Amodio, D. M. (2019). Model-based and model-free social cognition: Investigating the role of habit in social attitude formation and choice. *Frontiers in Psychology, 10*(2592), 1-11. https://doi.org/10.3389/fpsyg.2019.02592

Hahn, U., Bailey, T. M., & Elvin, L. B. C. (2005). Effects of category diversity on learning, memory, and generalization. *Memory and Cognition, 33*(2), 289–302. https://doi.org/10.3758/BF03195318

Hammerl, M., & Grabitz, H. J. (1993). Human evaluative conditioning: Order of stimulus presentation. *Integrative Physiological and Behavioral Science, 28*(2). https://doi.org/10.1007/BF02691227

Hamzani, O., Mazar, T., Itkes, O., Petranker, R., & Kron, A. (2019). Semantic and affective representations of valence: Prediction of autonomic and facial responses from feelings-focused and knowledge-focused self-reports. *Emotion, 20*(3), 486. https://doi.org/10.1037/emo0000567

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019). THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLoS ONE*, *14*(10). https://doi.org/10.1371/journal.pone.0223792

Hensels, I. S., & Baines, S. (2016). Changing “gut feelings” about food: An evaluative conditioning effect on implicit food evaluations and food choice. *Learning and Motivation*, *55*, 31–44. https://doi.org/10.1016/j.lmot.2016.05.005

Hess, Y. D., Carnevale, J. J., & Rosario, M. (2018). A construal level approach to understanding interpersonal processes. *Social and Personality Psychology Compass, 12*(8), 1–13. https://doi.org/10.1111/spc3.12409

Hewstone, M., Rubin, M., & Willis, H. (2002). Intergroup bias. *Annual Review of Psychology*, *53*(1), 575–604. https://doi.org/10.1146/annurev.psych.53.100901.135109

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. https://doi.org/10.1037/a0018916

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin*, *31*(10), 1369–1385.

https://doi.org/10.1177/0146167205275613

Högden, F., Stahl, C., & Unkelbach, C. (2020). Similarity-based and rule-based generalisation in the acquisition of attitudes via evaluative conditioning. *Cognition and Emotion*, *34*(1), 105–127. https://doi.org/10.1080/02699931.2019.1588709

Hohwy, J. (2020). New directions in predictive processing. *Mind and Language*, *35*(2), 209–223. https://doi.org/10.1111/mila.12281

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods, 24*(5), 539-556. https://doi.org/10.1037/met0000201

Hollands, G. J., & Marteau, T. M. (2016). Pairing images of unhealthy and healthy foods with images of negative and positive health consequences: Impact on attitudes and food choice. *Health Psychology, 35*(8), 847-851. https://doi.org/10.1037/hea0000293.supp

Homa, D., Sterling, S., & Trepel, L. (1981). Evaluation of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 638–648. https://doi.org/10.1037/0278-7393.10.4.638

Hoppe, D. B., Hendriks, P., Ramscar, M., & van Rij, J. (2022). An exploration of error-driven learning in simple two-layer networks from a discriminative learning perspective. *Behavior Research Methods*, *54*(5), 2221–2251. https://doi.org/10.3758/s13428-021-01711-5

Hoppe, D. B., van Rij, J., Hendriks, P., & Ramscar, M. (2020). Order matters! Influences of linear order on linguistic category learning. *Cognitive Science, 44*(11). https://doi.org/10.1111/cogs.12910

Hu, X., Gawronski, B., & Balas, R. (2017). Propositional versus dual-process accounts of evaluative conditioning: I. The effects of co-occurrence and relational information on implicit and explicit evaluations. *Personality and Social Psychology Bulletin, 43*(1), 17–32. https://doi.org/10.1177/0146167216673351

Huet, M., Jacobs, D. M., Camachon, C., Missenard, O., Gray, R., & Montagne, G. (2011). The education of attention as explanation of variability of practice effects: Learning the final approach phase in a flight simulator. *Journal of Experimental Psychology: Human Perception and Performance, 37*(6), 1841–1854. https://doi.org/10.1037/a0024386

Hummel, J. E. (2010). Symbolic versus associative learning. *Cognitive Science, 34*(6), 958–965. https://doi.org/10.1111/j.1551-6709.2010.01096.x

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review, 110*(2), 220–264. https://doi.org/10.1037/0033-295X.110.2.220

Hütter, M. (2022). An integrative review of dual- and single-process accounts of evaluative conditioning. *Nature Reviews Psychology, 1*(11), 640–653. https://doi.org/10.1038/s44159-022-00102-7

Hütter, M., & Fiedler, K. (2016). Editorial: Conceptual, theoretical, and methodological challenges in evaluative conditioning research. *Social Cognition, 34*(5), 343–356. https://doi.org/10.1521/soco.2016.34.5.343

Hütter, M., Kutzner, F., & Fiedler, K. (2014). What is learned from repeated pairings? On the scope and generalizability of evaluative conditioning. *Journal of Experimental Psychology: General*, *143*(2), 631–643. https://doi.org/10.1037/a0033409

Hütter, M., & Rothermund, K. (2020). Automatic processes in evaluative learning. *Cognition and Emotion, 34*(1), 1–20. https://doi.org/10.1080/02699931.2019.1709315

Hütter, M., & Tigges, D. (2019). On the external validity of evaluative conditioning: Evaluative responses generalize to modified instances of conditioned stimuli. *Journal of Experimental Social Psychology*, *84*(103824), 1–12. https://doi.org/10.1016/j.jesp.2019.103824

Hütter, M., Niese, Z. A., & Ihmels, M. (2022). Bridging the gap between autonomous and predetermined paradigms: The role of sampling in evaluative learning. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001172

Hütter, M., & Sweldens, S. (2018). Dissociating controllable and uncontrollable effects of affective stimuli on attitudes and consumption. *Journal of Consumer Research*, *45*(2), 320–349. https://doi.org/10.1093/jcr/ucx124

Hütter, M., Sweldens, S., Stahl, C., Unkelbach, C., & Klauer, K. C. (2012). Dissociating contingency awareness and conditioned attitudes: Evidence of contingency-unaware evaluative conditioning. *Journal of Experimental Psychology: General*, *141*(3), 539–557. https://doi.org/10.1037/a0026477

Isen, A. M., & Daubman, K. A. (1984). The influence of affect on categorization. *Journal of Personality and Social Psychology, 47*(6), 1206–1217. https://doi.org/10.1037//0022-3514.47.6.1206

Itkes, O., & Kron, A. (2019). Affective and semantic representations of valence: A conceptual framework. *Emotion Review, 11*(4), 283–293. https://doi.org/10.1177/1754073919868759

Jackson, J. C., Watts, J., List, J. M., Puryear, C., Drabble, R., & Lindquist, K. A. (2022). From text to thought: How analyzing language can advance psychological science. *Perspectives on Psychological Science, 17*(3), 805–826. https://doi.org/10.1177/17456916211004899

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2008). Generalization and similarity in exemplar models of categorization: Insights from machine learning. *Psychonomic Bulletin and Review, 15*(2), 256–271. https://doi.org/10.3758/PBR.15.2.256

Jensen-Fielding, H., Luck, C. C., & Lipp, O. V. (2018). Is the devil in the detail? Evidence for S-S learning after unconditional stimulus revaluation in human evaluative conditioning under a broader set of experimental conditions. *Cognition and Emotion, 32*(6), 1275–1290. https://doi.org/10.1080/02699931.2017.1408573

Jones, C. R., Fazio, R. H., & Olson, M. A. (2009). Implicit misattribution as a mechanism underlying evaluative conditioning. *Journal of Personality and Social Psychology, 96*(5), 933–948. https://doi.org/10.1037/a0014747

Jones, C. R., Olson, M. A., & Fazio, R. H. (2010). Evaluative conditioning. The “how” question. *Advances in Experimental Social Psychology, 43,* 205–255. https://doi.org/10.1016/S0065-2601(10)43005-1

Judd, C. M., & Park, B. (1988). Out-group homogeneity: Judgments of variability at the individual and group levels. *Journal of personality and social psychology, 54*(5), 778. https://doi.org/10.1037/0022-3514.54.5.778

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology, 68,* 601–625. https://doi.org/10.1146/annurev-psych-122414-033702

Jurchiș, R., Costea, A., Dienes, Z., Miclea, M., & Opre, A. (2020). Evaluative conditioning of artificial grammars: Evidence that subjectively-unconscious structures bias affective evaluations of novel stimuli. *Journal of Experimental Psychology: General*, *149*(9), 1800–1809. https://doi.org/10.1037/xge0000734

Kamin, L. (1969). Predictability, surprise, attention, and conditioning. In Campbell, R. A., & Church, R. M. (Eds.) *Punishment and aversive behavior*, (pp. 279–296). New York: Appleton-Century- Crofts.

Kattner, F. (2014). Reconsidering the (in)sensitivity of evaluative conditioning to reinforcement density and CS-US contingency. *Learning and Motivation*, *45*(1), 15–29. https://doi.org/10.1016/j.lmot.2013.09.002

Kattner, F., & Green, C. S. (2015). Cue competition in evaluative conditioning as a function of the learning process. *Acta Psychologica*, *162*, 40–50. https://doi.org/10.1016/j.actpsy.2015.09.013

Kaup, B., Ulrich, R., Bausenhart, K. M., Bryce, D., Butz, M. V., Dignath, D., Dudschig, C., Franz, V. H., Friedrich, C., Gawrilow, C., Heller, J., Huff, M., Hütter, M., Janczyk, M., Leuthold, H., Mallot, H., Nürk, H.-C., Ramscar, M., Said, N., Svaldi, J. & Wong, H. Y. (2023). Modal and amodal cognition: an overarching principle in various domains of psychology. *Psychological Research*, 1–31. https://doi.org/10.1007/s00426-023-01878-w

Kawakami, K., Amodio, D. M., & Hugenberg, K. (2017). Intergroup perception and cognition: An integrative framework for understanding the causes and consequences of social categorization. *Advances in Experimental Social Psychology, 55*, 1–80. https://doi.org/10.1016/bs.aesp.2016.10.001

Kerkhof, I., Vansteenwegen, D., Baeyens, F., & Hermans, D. (2011). Counterconditioning: An effective technique for changing conditioned preferences. *Experimental Psychology*, *58*(1), 31–38. https://doi.org/10.1027/1618-3169/a000063

Kiefer, A., & Hohwy, J. (2019). Representation in the prediction error minimization framework. *The Routledge Companion to Philosophy of Psychology*, 384–410. https://doi.org/10.4324/9780429244629-24

Kim, J. C., Sweldens, S., & Hütter, M. (2016). The symmetric nature of evaluative memory associations: Equal effectiveness of forward versus backward evaluative conditioning. *Social Psychological and Personality Science*, *7*(1), 61–68. https://doi.org/10.1177/1948550615599237

Kim, Y. J., Park, J., & Wyer, R. S. (2009). Effects of temporal distance and memory on consumer judgments. *Journal of Consumer Research, 36*(4), 634–645. https://doi.org/10.1086/599765

Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Mohr, A. H., Jzerman, H. I., Nilsonne, G., Vanpaemel, W., & Frank, M. C. (2018). A practical guide for transparency in psychological science. *Collabra: Psychology*, *4*(1), 1–15. https://doi.org/10.1525/collabra.158

Konovalova, E., & Le Mens, G. (2020). An information sampling explanation for the ingroup heterogeneity effect. *Psychological Review*, *127*(1), 47–73. https://doi.org/10.1037/rev0000160

Koppehele-Gossel, J., Hoffmann, L., Banse, R., & Gawronski, B. (2020). Evaluative priming as an implicit measure of evaluation: An examination of outlier-treatments for evaluative priming scores. *Journal of Experimental Social Psychology, 87*(103905). https://doi.org/10.1016/j.jesp.2019.103905

Kousta, S. T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General*, *140*(1), 14–34. https://doi.org/10.1037/a0021446

Krueger, J., & Rothbart, M. (1988). Use of categorical and individuating information in making inferences about personality. *Journal of Personality and Social Psychology, 55*(2), 187–195. https://doi.org/10.1037/0022-3514.55.2.187

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. https://doi.org/10.1177/2515245918771304

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review*, *103*(2), 284–308. https://doi.org/10.1037/0033-295X.103.2.284

Kurdi, B., & Banaji, M. R. (2019). Attitude change via repeated evaluative pairings versus evaluative statements: Shared and unique features. *Journal of Personality and Social Psychology*, *116*(5), 681–703. https://doi.org/10.1037/pspa0000151

Kurdi, B., & Charlesworth, T. E. S. (2023). A 3D framework of implicit attitude change. *Trends in Cognitive Sciences*, *27*(8), 745-758. https://doi.org/10.1016/j.tics.2023.05.009

Kurdi, B., Gershman, S. J., & Banaji, M. R. (2019). Model-free and model-based learning processes in the updating of explicit and implicit evaluations. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(13), 6035–6044. https://doi.org/10.1073/pnas.1820238116

Kurtz, K. J., Boukrina, O., & Gentner, D. (2013). Comparison promotes learning and transfer of relational categories. *Journal of Experimental Psychology: Learning Memory and Cognition*, *39*(4), 1303–1310. https://doi.org/10.1037/a0031847

Labiouse, C. L., & French, R. M. (2001). A connectionist model of person perception and stereotype formation. In J. A. Bullinaria, & D. W. Glasspool (Eds.), *Connectionist Models of Learning, Development and Evolution: Proceedings of the Sixth Neural Computation and Psychology Workshop, Liège, Belgium*, 16–18 September 2000 (pp. 209-218). London: Springer London. https://doi.org/10.1007/978-1-4471-0281-6_21

Lai, C. K., Marini, M., Lehr, M., Cerruti, S. A., Shin, C., Joy-Gaba, J. E. L., Ho, A. K., Teachman, B. A., Wojcik, S. P., Koleva, S. P., Frazier, R. S., Heiphetz, L., Chen, E. E., Turner, R. N., Haidt, J., Kesebir, S., Hawkins, C. B., Schaefer, H. S., Rubichi, S., Sartori, G., Dial, C. M., Sriram, N., Banaji, M. R., & Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General, 143*(4), 1765–1785. https://doi.org/10.1037/a0036260

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2020). Improving inferences about null effects with Bayes factors and equivalence tests. *Journals of Gerontology - Series B Psychological Sciences and Social Sciences, 75*(1), 45–57. https://doi.org/10.1093/geronb/gby065

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1997). International affective picture system (IAPS): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention, 1*, 39–58.

Le Pelley, M. E., Reimers, S. J., Calvini, G., Spears, R., Beesley, T., & Murphy, R. A. (2010). Stereotype formation: Biased by association. *Journal of Experimental Psychology: General*, *139*(1), 138–161. https://doi.org/10.1037/a0018210

Leach, C. W., van Zomeren, M., Zebel, S., Vliek, M. L. W., Pennekamp, S. F., Doosje, B., Ouwerkerk, J. W., & Spears, R. (2008). Group-level self-definition and self-investment: A hierarchical (multicomponent) model of in-group identification. *Journal of Personality* *and Social Psychology*, *95*(1), 144–165. https://doi.org/10.1037/0022-3514.95.1.144

Ledgerwood, A. (2014). Evaluations in their social context: Distance regulates consistency and context dependence. *Social and Personality Psychology Compass*, *8*(8), 436–447. https://doi.org/10.1111/spc3.12123

Ledgerwood, A., Eastwick, P. W., & Gawronski, B. (2020). Experiences of liking versus ideas about liking. Commentary on Gilead, Trope, & Liberman. *Behavioral and Brain Sciences*. https://doi.org/10.1017/S0140525X19003145

Ledgerwood, A., Eastwick, P. W., & Smith, L. K. (2018). Toward an integrative framework for studying human evaluation: Attitudes toward objects and attributes. *Personality and Social Psychology Review*, *22*(4), 378–398. https://doi.org/10.1177/1088868318790718

Ledgerwood, A., Trope, Y., & Chaiken, S. (2010). Flexibility now, consistency later: Psychological distance and construal shape evaluative responding. *Journal of Personality and Social Psychology*, *99*(1), 32–51. https://doi.org/10.1037/a0019843

Lee, I. C., Chen, E. E., Tsai, C. H., Yen, N. S., Chen, A. L. P., & Lin, W. C. (2016). Voting intention and choices: Are voters always rational and deliberative? *PLoS ONE*, *11*(2). https://doi.org/10.1371/journal.pone.0148643

Lee, I. C., & Pratto, F. (2011). Changing boundaries of ethnic identity and feelings toward ingroup/ outgroup: Examining Taiwan residents from a psycho-historical perspective. *Journal of Cross-Cultural Psychology*, *42*(1), 3–24. https://doi.org/10.1177/0022022110361776

Lee, I. C., Su, J. C., Gries, P. H., & Liu, F. C. S. (2018). When objective group membership and subjective ethnic identification don’t align: How identification shapes intergroup bias through self-enhancement and perceived threat. *Group Processes and Intergroup Relations*, *21*(4), 615–630. https://doi.org/10.1177/1368430216677301

Lee, J. C., Mills, L., Hayes, B. K., & Livesey, E. J. (2021). Modelling generalisation gradients as augmented Gaussian functions. *Quarterly Journal of Experimental Psychology*, *74*(1), 106–121. https://doi.org/10.1177/1747021820949470

Levey, A. B., & Martin, I. (1975). Classical conditioning of human ‘evaluative’ responses. *Behaviour Research and Therapy*, *13*(4), 221–226. https://doi.org/10.1016/0005-7967(75)90026-1

Li, J., Chen, B., & Zhang, Y. (2021). Adopting evaluative conditioning to improve coach–athlete relationships. *Frontiers in Psychology, 12*(November). https://doi.org/10.3389/fpsyg.2021.751990

Liberman, N., & Förster, J. (2009). The effect of psychological distance on perceptual level of construal. *Cognitive Science*, *33*(7), 1330–1341. https://doi.org/10.1111/j.1551-6709.2009.01061.x

Liberman, N., & Trope, Y. (2008). The psychology of transcending the here and now. *Science, 322*(5905), 1201–1205. https://doi.org/10.1126/science.1161958

Linville, P. W., Fischer, G. W., & Yoon, C. (1996). Perceived covariation among the features of ingroup and outgroup members: The outgroup covariation effect. *Journal of Personality and Social Psychology, 70*(3), 421–436. https://doi.org/10.1037/0022-3514.70.3.421

Linville, P. W., Fischer, G. W., & Salovey, P. (1989). Perceived distributions of the characteristics of in-group and out-group members: Empirical evidence and a computer simulation. *Journal of Personality and Social Psychology, 57*(2), 165–188. https://doi.org/10.1037/0022-3514.57.2.165

Lischetzke, T., Reis, D., & Arndt, C. (2015). Data-analytic strategies for examining the effectiveness of daily interventions. *Journal of Occupational and Organizational Psychology*, *88*(3), 587–622. https://doi.org/10.1111/joop.12104

Luck, C. C., Patterson, R. R., & Lipp, O. V. (2020). Be careful what you say!–Evaluative change based on instructional learning generalizes to other similar stimuli and to the wider category. *Cognition and Emotion*, *35*(1), 1–16. https://doi.org/10.1080/02699931.2020.1816912

Luguri, J. B., Napier, J. L., & Dovidio, J. F. (2012). Reconstruing intolerance: Abstract thinking reduces conservatives' prejudice against nonnormative groups. *Psychological Science, 23*(7), 756–763. https://doi.org/10.1177/0956797611433877

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods, 47*(4), 1122–1135. https://doi.org/10.3758/s13428-014-0532-5

Maass, A. (1999). Linguistic intergroup bias: Stereotype perpetuation through language. *Advances in Experimental Social Psychology, 31,* 79–121.

https://doi.org/10.1016/s0065-2601(08)60272-5

Macmillan, N. A., & Creelman, C. D. (2004). *Detection Theory: A User's Guide* (2nd edition). Lawrence Erlbaum Associates, Inc. https://doi.org/10.1016/j.appet.2020.105063

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology, 25*(4), 431–467. https://doi.org/10.1006/cogp.1993.1011

Marsman, M., Schönbrodt, F. D., Morey, R. D., Yao, Y., Gelman, A., & Wagenmakers, E. J. (2017). A Bayesian bird's eye view of 'Replications of important results in social psychology.' *Royal Scociety Open Science, 4*(160426), 137–141. https://doi.org/http://dx.doi.org/10.1098/rsos.160426

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* San Francisco: Freeman.

Masterton, S., Hardman, C., Halford, J. C. G., & Jones, A. (2021). Examining cognitive bias modification interventions for reducing food value and choice: Two pre-registered, online studies. *Appetite*, *159*. https://doi.org/10.1016/j.appet.2020.105063

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, *114*(2), 159–188. https://doi.org/10.1037/0096-3445.114.2.159

McConahay, J. B. (1986). Modern racism, ambivalence, and the Modern Racism Scale. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 91–125). Academic Press.

McCrea, S. M., Wieber, F., & Myers, A. L. (2012). Construal level mind-sets moderate self- and social stereotyping. *Journal of Personality and Social Psychology*, *102*(1), 51–68. https://doi.org/10.1037/a0026108

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan* (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315372495

McSweeney, F. K., & Bierley, C. (1984). Recent developments in classical conditioning. *Journal of Consumer Research, 11*(2), 619–631. https://doi.org/https://doi.org/10.1086/208999

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law, 7*(1), 3–

35). https://doi.org/10.1037/1076-8971.7.1.3

Meteyard, L., Cuadrado, S. R., Bahrami, B., & Vigliocco, G. (2012). Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex, 48*(7), 788–804. https://doi.org/10.1016/j.cortex.2010.11.002

Măgurean, S., Constantin, T., & Sava, F. A. (2016). The indirect effect of evaluative conditioning on smoking. *Journal of Substance Use*, *21*(2), 198–203. https://doi.org/10.3109/14659891.2015.1005183

Mierop, A., Hütter, M., Stahl, C., & Corneille, O. (2019). Does attitude acquisition in evaluative conditioning without explicit CS-US memory reflect implicit misattribution of affect? *Cognition and Emotion*, *33*(2), 173–184. https://doi.org/10.1080/02699931.2018.1435505

Milkman, K. L., Akinola, M., & Chugh, D. (2012). Temporal distance and discrimination: An audit study in academia. *Psychological Science*, *23*(7), 710–717. https://doi.org/10.1177/0956797611434539

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin*, *117*(3), 363–386. https://doi.org/10.1037/0033-2909.117.3.363

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

Montrey, M., & Shultz, T. R. (2019). Outgroup homogeneity bias causes ingroup favoritism. *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*. 2392–2398. https://doi.org/10.48550/arXiv.1908.08203

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2015). Processing goals moderate the effect of co-occurrence on automatic evaluation. *Journal of Experimental Social Psychology, 60*, 157–162. https://doi.org/10.1016/j.jesp.2015.05.009

Moran, T., Nudler, Y., & Anan, Y. B. (2023). Evaluative conditioning: Past, present, and future. *Annual Review of Psychology*, *74*, 14.1-14.25. https://doi.org/https://doi.org/10.1146/annurev-psych-032420-031815

Namy, L. L., & Gentner, D. (2002). Making a silk purse out of two sow’s ears: Young children's use of comparison in category learning. *Journal of Experimental Psychology: General, 131*(1), 5–15. https://doi.org/10.1037/0096-3445.131.1.5

Niese, Z. A., & Hütter, M. (2023). The malleability of sampling's impact on evaluation: Sampling goals moderate the evaluative impact of sampling a stimulus. *Journal of Experimental Social Psychology*, *109*(104516). https://doi.org/10.1016/j.jesp.2023.104516

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 700–708. https://doi.org/10.1037/0278-7393.14.4.700

Nosofsky, R. M. (2011). The generalized context model: an exemplar model of classification. In E. M. Pothos, & A. J. Wills (Eds.), *Formal Approaches in Categorization* (1st ed., pp. 18–39). Cambridge University Press. https://doi.org/10.1017/cbo9780511921322.002

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin and Review,26*(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science, 12*(5), 413–417. https://doi.org/10.1111/1467-9280.00376

Olson, M. A., & Fazio, R. H. (2006). Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin, 32*(4), 421–433. https://doi.org/10.1177/0146167205284004

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). https://doi.org/10.1126/science.aac4716

Osherson, D. N., Smith, E. E., & Lopez, A. (1990). Category-Based Induction. *Psychological Review, 97*(2), 185–200. https://doi.org/10.1037/0033-295X.97.2.185

Paas, F. G. W. C., & Van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology, 86*(1), 122–133. https://doi.org/10.1037//0022-0663.86.1.122

Paluck, E. L., Porat, R., Clark, C. S., & Green, D. P. (2021). Prejudice reduction: Progress and challenges. *Annual Review of Psychology, 72*, 533–560.

https://doi.org/10.1146/annurev-psych-071620

Park, B., & Hastie, R. (1987). Perception of variability in category development: Instance-versus abstraction-based stereotypes. *Journal of Personality and Social Psychology, 53*(4), 621–635. https://doi.org/10.1037/0022-3514.53.4.621

Park, B., & Judd, C. M. (1990). Measures and models of perceived group variability. *Journal of Personality and Social Psychology, 59*(2), 173–191. https://doi.org/10.1037/0022-3514.59.2.173

Park, B., Judd, C. M., & Ryan, C. S. (1991). Social categorization and the representation of variability information. *European Review of Social Psychology*, *2*(1), 211–245. https://doi.org/10.1080/14792779143000079

Park, B., & Rothbart, M. (1982). Perception of out-group homogeneity and levels of social categorization: Memory for the subordinate attributes of in-group and out-group members. *Journal of Personality and Social Psychology*, *42*(6), 1051–1068. https://doi.org/10.1037/0022-3514.42.6.1051

Park, B., Ryan, C. S., & Judd, C. M. (1992). Role of meaningful subgroups in explaining differences in perceived variability for in-groups and out-groups. *Journal of Personality and Social Psychology*, *63*(4), 553–567. https://doi.org/10.1037/0022-3514.63.4.553

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89*(3), 277–293. https://doi.org/10.1037/0022-3514.89.3.277

Payne, K., & Lundberg, K. (2014). The affect misattribution procedure: Ten years of evidence on reliability, validity, and mechanisms. *Social and Personality Psychology Compass, 8*(12), 672–686. https://doi.org/10.1111/spc3.12148

Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review*, *94*(1), 61–73. https://doi.org/10.1037//0033-295x.94.1.61

Perdue, C. W., Dovidio, J. F., Gurtman, M. B., & Tyler, R. B. (1990). Us and them: Social categorization and the process of intergroup bias. *Journal of Personality and Social Psychology, 59*(3), 475–486. https://doi.org/10.1037/0022-3514.59.3.475

Peters, J. F., Tozzi, A., Ramanna, S., & İnan, E. (2017). The human brain from above: an increase in complexity from environmental stimuli to abstractions. *Cognitive*

Neurodynamics, 11(4), 391–394. https://doi.org/10.1007/s11571-017-9428-2

Peters, K. R., & Gawronski, B. (2011). Are we puppets on a string? comparing the impact of contingency and validity on implicit and explicit evaluations. *Personality and Social Psychology Bulletin*, *37*(4), 557–569. https://doi.org/10.1177/0146167211400423

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology*, *19*, 123–205. https://doi.org/10.1016/S0065-2601(08)60214-2

Pleyers, G., Corneille, O., Yzerbyt, V., & Luminet, O. (2009). Evaluative conditioning may incur attentional costs. *Journal of Experimental Psychology: Animal Behavior Processes*, *35*(2), 279–285. https://doi.org/10.1037/a0013429

Posner, M., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3), 353–363. https://doi.org/10.1037/h0078016

R Core Team (2023). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Raes, F., Griffith, J. W., Craeynest, M., Williams, J. M. G., Hermans, D., Barry, T. J., Takano, K., & Hallford, D. J. (2023). Overgeneralization as a predictor of the course of depression over time: The role of negative overgeneralization to the self, negative overgeneralization across situations, and overgeneral autobiographical memory. *Cognitive Therapy and Research*, *47*(4), 598–613. https://doi.org/10.1007/s10608-023-10385-6

Ram, H., Grinfeld, G., & Liberman, N. (2023). Expected Variability Increases Generalization in Learning [Manuscript submitted for publication]. Department of Social & Health Sciences, Bar Ilan University.

Ramscar, M. (2021). A discriminative account of the learning, representation and processing of inflection systems. *Language, Cognition and Neuroscience* , *38*(4), 446–470. https://doi.org/10.1080/23273798.2021.2014062

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science, 34*(6), 909–957. https://doi.org/10.1111/j.1551-6709.2009.01092.x

Ranganath, K. A., & Nosek, B. A. (2008). Implicit attitude generalization occurs immediately; explicit attitude generalization takes time. *Psychological Science, 19*(3), 249–254. https://doi.org/10.1111/j.1467-9280.2008.02076.x

Raviv, L., Lupyan, G., & Green, S. C. (2022). How variability shapes learning and generalization. *Trends in Cognitive Sciences*, *26*(6), 462–483. https://doi.org/10.1016/j.tics.2022.03.007

Reed, S. K. (2016). A taxonomic analysis of abstraction. *Perspectives on Psychological Science, 11*(6), 817–837. https://doi.org/10.1177/1745691616646304

Reichmann, K., Hütter, M., Kaup, B., & Ramscar, M. (2023). Variability and abstraction in evaluative conditioning : Consequences for the generalization of likes and dislikes. *Journal of Experimental Social Psychology, 108*, 104478. https://doi.org/10.1016/j.jesp.2023.104478

Rescorla, R. A. (1968). Probability of shock in the presence and absence of Cs in fear conditioning. *Journal of Comparative and Physiological Psychology*, *66*(1), 1–5. https://doi.org/10.1037/h0025984

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

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803–814. https://doi.org/10.1037/0278-7393.21.4.803

Rosch, E. (1988). Principles of categorization. *Readings in Cognitive Science*, 312–322. https://doi.org/https://doi.org/10.1016/b978-1-4832-1446-7.50028-5

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, *12*(2), 339–349. https://doi.org/10.1111/j.1467-7687.2008.00786.x

Rydell, R. J., & Gawronski, B. (2009). I like you, I like you not: Understanding the formation of context-dependent automatic attitudes. *Cognition and Emotion*, *23*(6), 1118–1152. https://doi.org/10.1080/02699930802355255

Rydell, R. J., McConnell, A. R., Strain, L. M., Claypool, H. M., & Hugenberg, K. (2007).

Implicit and explicit attitudes respond differently to increasing amounts of counterattitudinal information. *European Journal of Social Psychology, 37*(5), 867–878. https://doi.org/10.1002/ejsp.393

Schechtman, E., Laufer, O., & Paz, R. (2010). Negative valence widens generalization of learning. *Journal of Neuroscience*, *30*(31), 10460–10464. https://doi.org/10.1523/JNEUROSCI.2377-10.2010

Semin, G. R., & Fiedler, K. (1988). The cognitive functions of linguistic categories in describing persons: Social cognition and language. *Journal of Personality and Social Psychology, 54*(4), 558–568. https://doi.org/10.1037/0022-3514.54.4.558

Shanks, D. R. (2010). Learning: From association to cognition. *Annual Review of Psychology, 61*, 273–301. https://doi.org/10.1146/annurev.psych.093008.100519

Shapira, O., Liberman, N., Trope, Y., & Rim, S. (2012). Levels of Mental Construal. In S. T. Fiske, & C. N. Macrae (Eds.), *The SAGE Handbook of Social Cognition* (pp. 2022–2030). SAGE Publications. https://doi.org/10.4135/9781446247631.n12

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*(4820), 1317–1323. https://doi.org/10.1126/science.3629243

Sherman, S. J., Rose, J. S., Koch, K., Presson, C. C., & Chassin, L. (2003). Implicit and explicit attitudes toward cigarette smoking: The effects of context and motivation. *Journal of Social and Clinical Psychology*, *22*(1), 13–39. https://doi.org/10.1521/jscp.22.1.13.22766

Siegel, S., & Allan, L. G. (1996). The widespread influence of the Rescorla-Wagner model. *Psychonomic Bulletin and Review*, *3*(3), 314–321. https://doi.org/10.3758/BF03210755

Simon, B. (1992). The perception of ingroup and outgroup homogeneity: Reintroducing the intergroup context. *European Review of Social Psychology, 3*(1), 1–30. https://doi.org/10.1080/14792779243000005

Simon, B., & Brown, R. (1987). Perceived intragroup homogeneity in minority-majority contexts. *Journal of Personality and Social Psychology*, *53*(4), 703–711. https://doi.org/10.1037/0022-3514.53.4.703

Simon, B., & Mummendey, A. (1990). Perceptions of relative group size and group homogeneity: We are the majority and they are all the same. *European Journal of Social*

Psychology, 20(4), 351–356. https://doi.org/10.1002/ejsp.2420200406

Smith, J. D. (2014). Prototypes, exemplars, and the natural history of categorization. *Psychonomic Bulletin and Review*, *21*(2), 312–331. https://doi.org/10.3758/s13423-013-0506-0

Smith, P. K., & Trope, Y. (2006). You focus on the forest when you're in charge of the trees: Power priming and abstract information processing. *Journal of Personality and Social Psychology*, *90*(4), 578–596. https://doi.org/10.1037/0022-3514.90.4.578

Soderberg, C. K., Callahan, S. P., Kochersberger, A. O., Amit, E., & Ledgerwood, A. (2015). The effects of psychological distance on abstraction: Two meta-analyses. *Psychological Bulletin*, *141*(3), 525–548. https://doi.org/10.1037/bul0000005.supp

Son, J. Y., Smith, L. B., & Goldstone, R. L. (2008). Simplicity and generalization: Short-cutting abstraction in children's object categorizations. *Cognition*, *108*(3), 626–638. https://doi.org/10.1016/j.cognition.2008.05.002

Sperlich, L. M., & Unkelbach, C. (2022). When do people learn likes and dislikes from co-occurrences? A dual-force perspective on evaluative conditioning. *Journal of Experimental Social Psychology*, *103*, 104377. https://doi.org/10.1016/j.jesp.2022.104377

Spruyt, A., Klauer, K. C., Gast, A., Schryver, M. De, & Houwer, J. De. (2014). Feature-specific attention allocation modulates the generalization of recently acquired likes and dislikes. *Experimental Psychology*, *61*(2), 85–98. https://doi.org/10.1027/1618-3169/a000228

Stahl, C., & Unkelbach, C. (2009). Evaluative learning with single versus multiple unconditioned stimuli: The role of contingency awareness. *Journal of Experimental Psychology: Animal Behavior Processes*, *35*(2), 286–291. https://doi.org/10.1037/a0013255

Stan Development Team (2023). *RStan: the R interface to Stan* (Version 2.32.3), https://mc-stan.org/

Stephan, E., Liberman, N., & Trope, Y. (2011). The effects of time perspective and level of construal on social distance. *Journal of Experimental Social Psychology*, *47*(2), 397–402. https://doi.org/10.1016/j.jesp.2010.11.001

Stuart, E. W., Shimp, T. A., & Engle, R. W. (1987). Classical conditioning of consumer attitudes: Four experiments in an advertising context. *Journal of Consumer Research, 14*(3), 334–349. https://doi.org/https://doi.org/10.1086/209117

Sweldens, S., Van Osselaer, S. M. J., & Janiszewski, C. (2010). Evaluative conditioning procedures and the resilience of conditioned brand attitudes. *Journal of Consumer Research, 37*(3), 473–489. https://doi.org/10.1086/653656

Szollosi, A., & Donkin, C. (2021). Arrested theory development: The misguided distinction between exploratory and confirmatory research. *Perspectives on Psychological Science, 16*(4), 717–724. https://doi.org/10.1177/1745691620966796

Taylor, P., Hobbs, J. N., Burroni, J., & Siegelmann, H. T. (2015). The global landscape of cognition: Hierarchical aggregation as an organizational principle of human cortical networks and functions. *Scientific Reports, 5*, 1–18. https://doi.org/10.1038/srep18112

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285. https://doi.org/10.1126/science.1192788

Thew, G. R., Gregory, J. D., Roberts, K., & Rimes, K. A. (2017). Self-critical thinking and overgeneralization in depression and eating disorders: An experimental study. *Behavioural and Cognitive Psychotherapy*, *45*(5), 510–523. https://doi.org/10.1017/S1352465817000327

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review, 117*(2), 440–463. https://doi.org/10.1037/a0018963

Tummeltshammer, K., Amso, D., French, R. M., & Kirkham, N. Z. (2017). Across space and time: infants learn from backward and forward visual statistics. *Developmental Science, 20*(5). https://doi.org/10.1111/desc.12474

Tussing, A. A., & Greene, R. L. (1999). Differential effects of repetition on true and false recognition. *Journal of Memory and Language*, *40*(4), 520–533. https://doi.org/10.1006/jmla.1999.2636

Unkelbach, C., & Fiedler, K. (2016). Contrastive CS-US relations reverse evaluative conditioning effects. *Social Cognition*, *34*(5), 413–434. https://doi.org/10.1521/soco.2016.34.5.413

Unkelbach, C., Koch, A., & Alves, H. (2019). The evaluative information ecology: On the frequency and diversity of "good" and "bad." *European Review of Social Psychology, 30*(1), 216–270. https://doi.org/10.1080/10463283.2019.1688474

Vallacher, R. R., & Wegner, D. M. (1989). Levels of personal agency: individual variation in action identification. *Journal of Personality and Social Psychology, 57*(4), 660–671. https://doi.org/10.1037/0022-3514.57.4.660

Van Den Heuvel, T. J., Derksen, J. J. L., Eling, P. A. T. M., & Van Der Staak, C. P. F. (2012). An investigation of different aspects of overgeneralization in patients with major depressive disorder and borderline personality disorder. *British Journal of Clinical Psychology*, *51*(4), 376–395. https://doi.org/10.1111/j.2044-8260.2012.02034.x

Vanbrabant, K., Boddez, Y., Verduyn, P., Mestdagh, M., Hermans, D., & Raes, F. (2015). A new approach for modeling generalization gradients: A case for hierarchical models. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.00652

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Vigliocco, G., Kousta, S. T., Della Rosa, P. A., Vinson, D. P., Tettamanti, M., Devlin, J. T., & Cappa, S. F. (2014). The neural representation of abstract words: The role of emotion. *Cerebral Cortex*, *24*(7), 1767–1777. https://doi.org/10.1093/cercor/bht025

Von Helversen, B., Herzog, S. M., & Rieskamp, J. (2014). Haunted by a doppelgänger: Irrelevant facial similarity affects rule-based judgments. *Experimental Psychology, 61*(1), 12–22. https://doi.org/10.1027/1618-3169/a000221

Vukatana, E., Graham, S. A., Curtin, S., & Zepeda, M. S. (2015). One is not enough: Multiple exemplars facilitate infants' generalizations of novel properties. *Infancy*, *20*(5), 548–575. https://doi.org/10.1111/INFA.12092

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E. J., van Doorn, J., Šmíra, M., Epskamp, S., Etz, A., Matzke, D., … Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin and Review*, *25*(1), 58–76. https://doi.org/10.3758/s13423-017-1323-7

Walther, E. (2002). Guilty by mere association: Evaluative conditioning and the spreading attitude effect. *Journal of Personality and Social Psychology, 82*(6), 919–934. https://doi.org/10.1037/0022-3514.82.6.919

Walther, E., Gawronski, B., Blank, H., & Langer, T. (2009). Changing likes and dislikes through the back door: The US-revaluation effect. *Cognition and Emotion*, *23*(5), 889–917. https://doi.org/10.1080/02699930802212423

Walther, E., Halbeisen, G., & Blask, K. (2018). What You feel is what you see: A binding perspective on evaluative conditioning. *Social Psychological Bulletin, 13*(3). https://doi.org/10.5964/spb.v13i3.27551

Walther, E., & Nagengast, B. (2006). Evaluative conditioning and the awareness issue: Assessing contingency awareness with the four-picture recognition test. *Journal of Experimental Psychology: Animal Behavior Processes, 32*(4), 454–459. https://doi.org/10.1037/0097-7403.32.4.454

Walther, E., Nagengast, B., & Trasselli, C. (2005). Evaluative conditioning in social psychology: Facts and speculations. *Cognition and Emotion, 19*(2), 175–196. https://doi.org/10.1080/02699930441000274

Watkins, E. R., Baeyens, C. B., & Read, R. (2009). Concreteness training reduces dysphoria: Proof-of-principle for repeated cognitive bias modification in depression. *Journal of Abnormal Psychology*, *118*(1), 55–64. https://doi.org/10.1037/a0013642

Wicker, A. W. (1969). Attitudes versus actions: The relationship of verbal and overt behavioral responses to attitude objects. *Journal of Social Issues*, *25*(4), 41–78. https://doi.org/10.1111/j.1540-4560.1969.tb00619.x

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., & Chang, W. (2014). Package ggplot2: An implementation of the grammar of graphics. *Create Elegant Data Visualisations Using the Grammar of Graphics*, *2*(1), 1–189. https://mran.microsoft.com/snapshot/2015-01-08/web/packages/ggplot2/ggplot2.pdf

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour, 2,* 915-924.

https://doi.org/10.1038/s41562-018-0467-4

Zerhouni, O., Houben, K., El Methni, J., Rutte, N., Werkman, E., & Wiers, R. W. (2019). I didn’t feel like drinking, but I guess why: Evaluative conditioning changes on explicit attitudes toward alcohol and healthy foods depends on contingency awareness. *Learning and Motivation*, *66*, 1–12. https://doi.org/10.1016/j.lmot.2019.02.001

SUPPLEMENTARY MATERIALS

Chapter 1:

Variability and abstraction in evaluative conditioning: Consequences for the generalization of likes and dislikes

Kathrin Reichmann, Mandy Hütter, Barbara Kaup, & Michael Ramscar

Eberhard Karls Universität Tübingen

Supplement A

Predictions for variable versus invariable CSs in evaluative conditioning can be expressed in quantitative terms by calculating associative strengths between CS cues and US valence with the Rescorla-Wagner (RW) model (Rescorla & Wagner, 1972). See *Figure 1* in the manuscript for an illustration of the variable versus invariable learning conditions. Formally stated, the RW assumes a discrepancy function (1) $\Delta V_{ij}^{n+1} = \alpha_i \beta_j \ (\lambda_j - V_{TOTAL})$ and an updating rule (2) $V_{ij}^{n+1} = V_{ij}^{n} + \Delta V_{ij}^{n+1}$.

- ΔV_{ij}^{n+1} is the change in associative strength (V) of cue *i* and an outcome *j* on a given trial n.
- V_{TOTAL} is the sum of associative strengths for all cues present on trial *n*.
- V_{ij}^{n+1} is the associative strength after trial *n*.
- β_j denotes the rate of learning for outcome j, and was set to .3 for all calculations.
- λ_j is the maximum associative strength that an outcome *j* can support, and was set to 1 for positive pairings and -1 for negative pairings.
- α_i denotes the salience of a cue *i*, and was set to .3 if a CS cue was present and to 0 if a CS cue was absent.

To calculate the associative strength for each CS cue, a vector of i = (1, 2, 3, 4) for V_{ij} represents Cues 1 to 4, with 1 representing the fixed Cue 1, and 2-4 the unique cues of CS1 to CS3. In the *invariable* condition, α_i = (.3, .3, 0, 0) because only Cue 1 and Cue 2 are presented. In the *variable* condition, α_i was set to .3 for *i* = 1 (given that Cue 1 is present in all the CSs) and was randomly set to .3 for one of Cues 2 to 4, thus for *i* = 2, 3 or 4, e.g., α_i = (.3, 0, 0, .3). Associative strengths were calculated for *n* = 20 learning trials, corresponding to the number of learning trials in the reported studies. Outcomes for both positive and negative pairings were calculated but obviously result in the same values of associative strength. For each combination of valence and conditioning procedure, 100 simulations were run with random orders of stimulus presentations. In line with the qualitative predictions, averaged associative values for cues *i* = (1, 2) amounted to |0.49| in the *invariable* condition, and |0.49| (Cue 1) and |0.16| (Cue 2) in the *variable* condition.

Supplement B

Supplement B presents Experiment 2S that was conducted in between Experiments 1 and 2. The experiment employed a similar experimental procedure as Experiment 2, but presented the dependent measures in a different temporal sequence (recognition memory measure at the end of the experiment). Moreover, in this experiment evaluative ratings of CS components were interspersed with ratings of CSs and GSs. The pre-registration of the experiment can be accessed via https://osf.io/zse65/?view_only=057c4d459a70420a8107d9faec789f18, the data of the experiment are available under https://osf.io/xg5p4/?view_only=50497c1690c44af2be24015ca213230b.[42]

Experiment 2S

Method

Participants

The required sample size for a replication of the three-way interaction of US valence, CS variability, and stimulus type obtained in Experiment 1 (direct evaluative measure, *B* = 32.55, *SE* = 5.45) was based on simulated data from the model specified in Experiment 1. To achieve a power greater than .95 and to account for 20% possible data exclusions, the anticipated sample size was set to *N* = 132. The study was conducted online with the study link distributed via the university's mailing list. After excluding 18 participants (of total *N* = 136) who reported that they had not paid attention during learning or who spoke Chinese, a total of *N* = 118 data sets were included in the analysis. This sample size provided 80% power to detect a standardized beta coefficient of β = 0.46 or greater (for the three-way interaction of US valence, stimulus type and CS variability on direct evaluations, with an alpha-level of .05). The sample consisted of students with different majors aged between 18 and 32 years (*M* = 22.57, *SD* = 3.32), with 89 identifying as female, 27 as male, and 2 as diverse. In total, the study took about 15 minutes. Participants received course credit or could sign up for a raffle for 20 x 25€ vouchers for a local bookstore.

Procedure

The materials were identical to Experiment 2. With regards to the study procedure, participants completed the recognition memory task at the end of the experiment, after the AMP and the direct evaluative ratings. Additionally, participants evaluated the components of

[42] Two additional experimental conditions were included for exploratory purposes. One condition included only one CS per category and presented each CS-US pair only once, *N* = 63. The second condition included one CS per category and 16 filler stimuli, *N* = 61. Data of the additional conditions are available on OSF: https://osf.io/xg5p4/?view_only=50497c1690c44af2be24015ca213230b.

the CSs on a direct rating scale (-100 to 100) together with CSs and GSs and not in a separate task. We changed the latter aspect in Experiment 2 to avoid that the ratings of CSs and GSs were affected by the decomposition of CSs into their components.

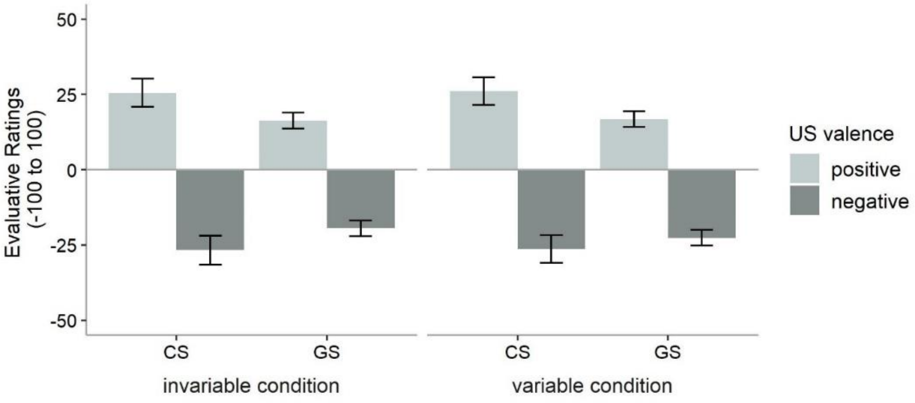
Results

Direct Evaluative Ratings

Mean evaluative responses are displayed in *Figure 1S*. Evaluative responses were submitted to the same mixed-effect model as in Experiment 1, including US valence (positive vs. negative), stimulus type (CS vs. GS), and CS variability (invariable vs. variable) as fixed effects (effect coded), and random by-subject intercepts and slopes for US valence and stimulus type. There was a main effect of US valence, *B* = 44.97, *SE* = 4.37, *t*(134.85) = 10.28, *p* < .001, β = 0.84, indicating an overall EC effect. On average, stimuli linked to positive valence were evaluated 44.97 points (on a scale from -100 to 100) higher than stimuli linked to negative valence. The EC effect was reduced for GSs, as implied by a significant, negative parameter of the two-way interaction between valence and stimulus type, *B* = -14.77, *SE* = 4.45, *t*(1652) = -3.32, *p* < .001, β = -0.28. To assess the degree of generalization for positive versus negative US valence, we calculated simple slopes for stimulus type per level of US valence. For negative pairings, ratings were not significantly lower for GSs than CSs, *B* = 5.52, *SE* = 3.15, *t*(1606.12) = 1.75, *p* = .080, β = 0.10. For positive pairings, ratings were on average 9.96 points lower for GSs than CSs, *B* = -9.26, *SE* = 3.15, *t*(1606.12) = -2.94, *p* = .003, β = -0.17. This additional analysis points towards stronger generalization for negatively than positively paired categories. The three-way interaction of valence and stimulus type was not qualified by CS variability, *B* = 3.46, *SE* = 8.90, *t*(1652) = 0.39, *p* = .698, β = 0.06.

Figure 1S

Mean Evaluative Ratings in Experiment 2S



Note. CS = Conditioned Stimuli, GS = Generalization Stimuli. Error bars display standard errors.

AMP

“Pleasant” (1) versus “unpleasant” (0) responses obtained in the AMP were analyzed with a generalized linear mixed-effect model (0.6% of timed out trials were excluded). There was a main effect of US valence, *B* = 1.64, *SE* = 0.13, *z* = 3.77, *p* < .001, indicating a significant overall EC effect. On average, the odds to respond “pleasant” were 1.64 times higher for stimuli linked to positive than negative valence. The three-way interaction of US valence, stimulus type, and CS variability did not reach significance, *B* = 0.77, *SE* = 0.44, *z* = -0.60, *p* = .551, and the effect points in the opposite direction than the one observed in Experiment 1. All other parameter estimates did not differ significantly from zero (smallest *p* = .104).

Evaluative Ratings of CS Components

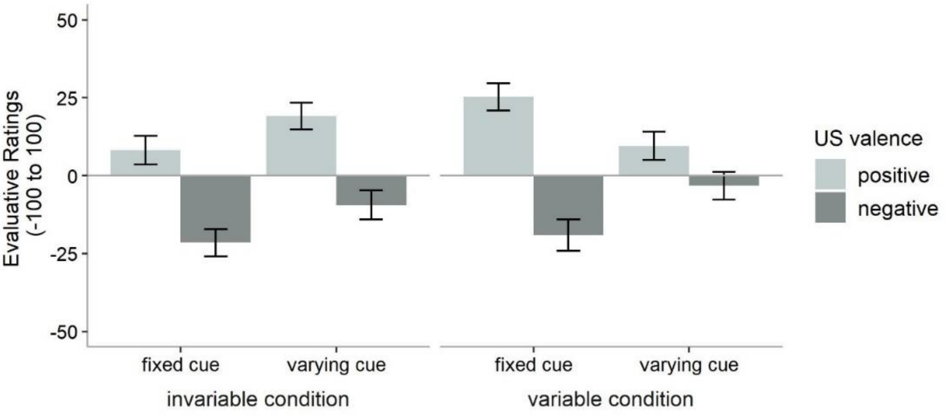
Continuous evaluative ratings of CS components were submitted to a linear mixed-model with US valence, CS variability, and stimulus component (fixed vs. varying) as fixed effects, and random by-subject intercepts and slopes for US valence and stimulus components. The aggregated ratings and their standard errors are depicted in *Figure 2S*. Overall, there was a main effect of US valence indicating a significant EC effect also for CS components, *B* = 28.58, *SE* = 4.05, *t*(118) = 7.12, *p* < .001, β = 00.56. Importantly, the three-way interaction of US valence, CS variability, and component type was significant, *B* = -30.46, *SE* = 11.58, *t*(590.01) = -2.63, *p* = .009, β = -0.59, indicating that the stimulus components showed different EC effects depending on the CS variability condition.

We further analyzed the three-way interaction by calculating separate interaction effects (US valence × component type) for each CS variability condition. This allowed us to test whether evaluations differed between the two types of components depending on CS variability. In the invariable condition, the two-way interaction between US valence and component type was not significant, *B* = -1.18, *SE* = 8.33, *t*(590) = -0.14, *p* = .887, β = -0.02. However, while the ratings in this condition did not differ between components for positively paired components, *B* = 10.92, *SE* = 6.09, *t*(360.52) = 1.79, *p* = .074, β = 0.21, they were more negative for fixed than varying components for negatively paired components, *B* = 12.11, *SE* = 6.09, *t*(360.52) = 1.99, *p* = .048, β = 0.23.

In the variable condition, the two-way interaction was significant, *B* = -31.64, *SE* = 8.05, *t*(590) = -3.93, *p* < .001, β = -0.61. As expected, the EC effect was larger for the fixed component than the component varying across characters in this condition. Ratings of positively paired components were higher for fixed than variable components, *B* = -15.78, *SE* = 5.89, *t*(360.52) = -2.68, *p* = .008, β = -0.31, and ratings of negatively paired components were lower for fixed than variable components, *B* = 15.86, *SE* = 5.89, *t*(360.49) = 2.69, *p* = .007, β = 0.31. This effect was as expected based on the abstraction account of variability and did not depend on valence.

Figure 2S

Mean Evaluative Ratings of CS Components in Experiment 2S



Note. Fixed cue = CS component fixed across characters of a category, varying cue = CS component varying between characters. Error bars display standard errors.

Recognition Memory Task

The signal detection measure *d’* was calculated for the “old” and “new” responses taken from every participant in the recognition memory task. These were then analyzed using the same linear mixed model as specified in Experiment 2. Overall, *d’* values were higher in the invariable condition (*M* = 0.58, *SD* = 1.39) than the variable condition (*M* = 0.20, *SD* = 1.34), although this difference was not statistically significant, *B* = -0.38, *SE* = 0.20, *t*(118) =

-1.94, *p* = .055. Moreover, *d'* values did not depend on US valence, *B* = -0.22, *SE* = 0.16, *t*(118) = -1.40, *p* = .166, and the two-way interaction between CS variability and US valence also did not reach significance, *B* = -0.05, *SE* = 0.31, *t*(118) = -0.15, *p* = .882. See *Figure 3S* for the aggregated d' scores.

An additional analysis of raw "old" versus "new" responses per stimulus type displayed a similar result pattern (see *Table 1S*, and *Supplement E* for a detailed description of the analysis).

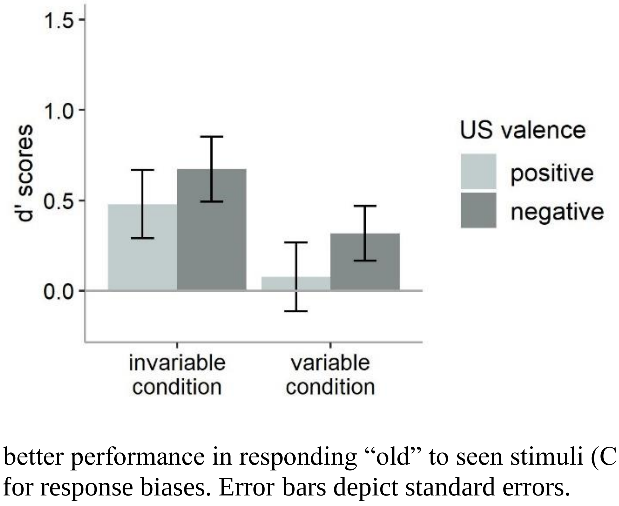
Table 1S

Proportion (standard deviations) of 'old' responses in the invariable versus variable condition for CSs, GSs, and distractors across experiments, with simple slopes testing for the difference between CS variability conditions

	'old' responses *Invariable*	'old' responses *Variable*	*OR*	*SE*	*z*	*p*
Exp. 2S						
CS	0.89 (0.16)	0.89 (0.15)	0.87	0.33	-0.41	.682
GS	0.53 (0.27)	0.66 (0.16)	1.79	0.16	3.54	< .001
Distractor	0.05 (0.11)	0.10 (0.17)	2.49	0.40	2.27	.023

Figure 3S

Mean d' scores in Experiment 2S



Note. Higher *d'* scores reflect a better performance in responding "old" to seen stimuli (CSs) and "new" to novel stimuli (GSs) while controlling for response biases. Error bars depict standard errors.

Discussion

The additional experiment reported here sought to replicate the effect of CS variability on evaluative generalization from Experiment 1, while keeping the number of CSs evaluated at test fixed between CS variability conditions. Further, a recognition memory task was added to the end of the experiment, and CS components were included in the evaluative rating phase.

The results of the added measures indicated that participants formed different types of representations depending on the conditioning procedure. Evaluative ratings of CS components were in line with the predictions of the abstraction account. Namely, the size of the EC effect did not differ between individual components in the invariable condition but was more pronounced for the fixed than varying component in the variable condition. This finding supports the notion that participants learn to distinguish the fixed from the varying component in the variable condition. Moreover, on a descriptive level the results of the recognition memory task indicated that participants had greater difficulties distinguishing CSs from GSs in the variable condition than in the invariable condition. Nevertheless, the results should be interpreted with caution as they were not significant. In addition, some participants mentioned in the open comment section that they did not understand which stimuli to classify as “old” – stimuli they saw during learning (CSs only, as anticipated), or stimuli they saw during learning or testing (CSs and GSs, because they saw GSs when evaluating the stimuli).

We therefore conducted Experiment 2 reported in the main article to test whether the effect on generalization reported in Experiment 1 can be replicated, while improving the employed experimental procedure. Given that Experiment 2 demonstrated the predicted effect of CS variability on generalization, we deem carry-over effects between tasks next to random fluctuation the most plausible explanation for the lack of significant results in this experiment.

Supplement C

Supplement C presents the instructions provided to participants before the conditioning phase, as well as the AMP and Recognition Memory Task. Instructions are translated from German.

Conditioning Phase

In this study we want to investigate how people react to different pictures. First, you will participate in a simple perception task. You will see Chinese characters together with positive or negative pictures on the screen for a short period of time. Please look at the pictures and characters carefully.

AMP

You will now see the Chinese characters for only a very short period of time. Then, different syllables will appear that are Chinese pronunciations (e.g. “lin“). The syllables and characters are presented together in a random manner; thus, they are independent of each other. Your task will be to indicate for each syllable if it could mean something pleasant or unpleasant. Please rely mainly on your intuition to do this task. The characters are presented only for orientation. Therefore, you don't have to pay attention to them. Focus primarily on the syllables.

Press “d“ if the syllable could mean something unpleasant. Press “k“ if it could mean something pleasant. Please put the index finger of your left hand on key “d“ and the index finger of your right hand on key “k“. React on each syllable as fast as you can. We are interested in your spontaneous responses.

Recognition Memory Task

Please indicate for each Chinese character if you just saw it together with a positive or negative picture or not. Press key “x“ if you see the character for the first time, i.e. if it is new to you. Press key “m“ if you have seen the character together with a picture in the perception task, i.e. if it is familiar to you. Please put your index finger of your left hand on key “x“ and the index finger of your right hand on key “m“. Answer as fast as and as accurate as you can. We are interested in your spontaneous responses.

Supplement D

In Supplement D, we present all parameter coefficients of fixed and random effects of the mixed models conducted for each experiment and dependent measure (Experiments 1-3 reported in the main article, and Experiment 2S reported in Supplement B).

Experiment 1

Table S2

Results of the Mixed Model conducted on Evaluative Ratings in Experiment 1

Effect	*B*	*SE*	*df*	*t*	*p*	β
Fixed effects						
(Intercept)	3.83	(1.06)	247.41	3.62	< .001	.02
US valence	47.70	(3.16)	217.12	15.10	< .001	.91
CS variability	-6.92	(2.12)	247.41	-3.27	.001	-.13
Stimulus Type	-3.50	(1.36)	4047.14	-2.57	.010	-.07
US valence × CS variability	5.89	(6.32)	217.12	0.93	.352	.11
US valence × Stimulus Type	-24.03	(2.72)	4388.62	-8.82	< .001	-.46
CS variability × Stimulus Type	1.23	(2.73)	4047.14	0.45	.652	.02
US valence × Stimulus Type × CS Variability	32.55	(5.45)	4388.62	5.97	< .001	.62

		Correlation with random effect for	
	Variance	Intercept	US valence
Random effects			
(Intercept)	131.42		
US valence	1624.13	-.24	
Stimulus Type	0.82	-.59	.93

Note. Fixed effects estimates (top) and variance-covariance estimates (bottom) for the mixed model conducted on direct evaluative ratings (-100 to 100). The model was specified as lmer(ratings ~ USvalence * stimulus type * CS variability + (USvalence + stimulus type | subject) in R, using the lme4 package. Effect coding: US valence (-0.5 negative, 0.5 positive), stimulus type (-0.5 CS, 0.5 GS), CS variability (-0.5 invariable, 0.5 variable). All *p*-values in this table are two-tailed. Number of level-1 observations = 4784; Number of level-2 clusters = 200.

Table S3

Results of the Generalized Mixed Model conducted on Responses of the AMP in Experiment 1

Effect	*log-odds*	*OR*	*SE*	*Wald Z*	*p*
Fixed effects					
(Intercept)	0.25	1.28	(0.04)	6.09	< .001
US valence	0.78	2.17	(0.11)	7.36	< .001
CS variability	-0.21	0.81	(0.08)	-2.66	.008
Stimulus type	-0.05	0.95	(0.07)	-0.71	.475
US valence × CS variability	0.34	1.40	(0.21)	1.59	.111
US valence × Stimulus Type	-0.27	0.76	(0.15)	-1.87	.061
CS variability × Stimulus Type	-0.06	0.94	(0.15)	-0.41	.681
US valence × Stimulus Type × CS Variability	0.28	1.33	(0.29)	0.97	.330

		Correlation with random effect for	
	Variance	Intercept	US valence
Random effects			
(Intercept)	0.06		
US valence	1.16	.01	
Stimulus Type	0.00	.99	-.17

Note. Fixed effects estimates (top) and variance-covariance estimates (bottom) for the generalized mixed model (binominal link function) conducted on responses of the AMP (0 = unpleasant, 1 = pleasant). The model was specified as glmer(responses ~ USvalence * stimulus type * CS variability + (USvalence + stimulus type | subject, family = binomial) in R, using the lme4 package. Effect coding: US valence (-0.5 neg, 0.5 pos), stimulus type (-0.5 CS, 0.5 GS), CS variability (-0.5 invariable, 0.5 variable). All *p*-values in this table are two-tailed. Number of level-1 observations = 4784; Number of level-2 clusters = 200.

Experiment 2S

Table S4

Results of the Mixed Model conducted on Evaluative Ratings in Experiment 2S

Effect	*B*	*SE*	*df*	*t*	*p*	β
Fixed effects						
(Intercept)	-1.30	(1.87)	120.39	-0.69	.488	.01
US valence	44.97	(4.37)	134.85	10.28	< .001	.84
CS variability	-0.43	(3.73)	120.39	-0.11	.909	-.01
Stimulus Type	-1.87	(2.23)	1468.92	-0.84	.402	-.03
US valence × CS variability	1.95	(8.75)	134.85	0.22	.824	.04
US valence × Stimulus Type	-14.77	(4.45)	1652.00	-3.32	.001	-.28
CS variability × Stimulus Type	-1.79	(4.46)	1468.92	-0.40	.688	-.03
US valence × Stimulus Type × CS Variability	3.46	(8.90)	1652.00	0.39	.698	.06

		Correlation with random effect for	
	Variance	Intercept	US valence
Random effects			
(Intercept)	264.68		
US valence	1670.77	-.12	
Stimulus Type	2.20	-.82	-.47

Note. See Experiment 1 for a description of the model. Number of level-1 observations = 1888; Number of level-2 clusters = 118.

Table S5

Results of the Mixed Model conducted on Evaluative Ratings of CS Components in Experiment 2S

Effect	*B*	*SE*	*df*	*t*	*p*	β
Fixed effects						
(Intercept)	1.08	(1.86)	118.00	0.58	.562	.00
US valence	28.85	(4.05)	118.00	7.12	< .001	.56
CS variability	4.01	(3.73)	118.00	1.08	.284	.08
Component Type	5.78	(3.09)	118.00	1.87	.064	.11
US valence × CS variability	-0.54	(8.11)	118.00	-0.07	.947	-.01
US valence × Component Type	-16.41	(5.79)	590.01	-2.83	.005	-.32
CS variability × Component Type	-11.47	(6.18)	118.00	-1.86	.066	-.22
US valence × Component Type × CS Variability	-30.46	(11.58)	590.01	-2.63	.009	-.59

		Correlation with random effect for	
	Variance	Intercept	US valence
Random effects			
(Intercept)	162.80		
US valence	947.40	-.38	
Component Type	137.60	.04	.45

Note. Fixed effects estimates (top) and variance-covariance estimates (bottom) for the mixed model conducted on direct evaluative ratings of CS components (-100 to 100). The model was specified as lmer(ratings ~ USvalence * component type * CS variability + (USvalence + component type | subject) in R, using the lme4 package. Effect coding: US valence (-0.5 negative, 0.5 positive), component type (-0.5 fixed, 0.5 varying), CS variability (-0.5 invariable, 0.5 variable). All *p*-values in this table are two-tailed. Number of level-1 observations = 944; Number of level-2 clusters = 118.

Table S6

Results of the Generalized Mixed Model conducted on Responses of the AMP in Experiment 2S

Effect	*log-odds*	*OR*	*SE*	*Wald Z*	*p*
Fixed effects					
(Intercept)	0.13	1.14	(0.06)	2.13	.034
US valence	0.49	1.64	(0.13)	3.77	< .001
CS variability	0.02	1.02	(0.12)	0.16	.871
Stimulus type	0.10	1.10	(0.11)	0.89	.376
US valence × CS variability	0.11	1.12	(0.26)	0.43	.671
US valence × Stimulus Type	0.36	1.43	(0.22)	1.63	.104
CS variability × Stimulus Type	0.24	1.28	(0.22)	1.09	.274
US valence × Stimulus Type × CS Variability	-0.26	0.77	(0.44)	-0.60	.551

		Correlation with random effect for	
	Variance	Intercept	US valence
Random effects			
(Intercept)	0.08		
US valence	0.57	.34	
Stimulus Type	0.03	.47	-.67

Note. See Experiment 1 for a description of the model. Number of level-1 observations = 1874; Number of level-2 clusters = 118.

Table S7

Results of the Generalized Mixed Model conducted on responses of the Recognition Memory Task in Experiment 2S

Effect	*log-odds*	*OR*	*SE*	*Wald Z*	*p*
Fixed effects					
(Intercept)	-0.02	0.98	(0.10)	-0.21	.833
Stimulus Type I: Distractor vs. CS	4.58	97.20	(0.25)	18.39	< .001
Stimulus Type II: Distractor vs. GS	0.87	2.39	(0.18)	4.85	< .001
CS variability: CS	-0.13	0.87	(0.33)	-0.41	.682
CS variability: GS	0.58	1.79	(0.16)	3.54	< .001
CS variability: Distractor	0.91	2.49	(0.40)	2.27	.023
	Variance				
Random effects					
(Intercept)	0.41				

Note. Fixed effects estimates (top) and variance-covariance estimates (bottom) for the generalized mixed model (binominal link function) conducted on responses of the Recognition Memory Task (1 = old, 0 = new). The model was specified as glmer (memory response ~ stimulus type I + stimulus type II + CS variability : stimulus type_dummy + (1|subject), binomial) in R, using the lme4 package. Other than pre-registered, by-subject random slopes for stimulus type were removed due to convergence issues. Effect coding: CS variability (-0.5 invariable, 0.5 variable). Dummy Coding: Stimulus type I (0 distractor, 1 CS), Stimulus Type II (0 distractor, 1 GS), stimulus type_dummy: level of interest is set to 0. All *p*-values in this table are two-tailed. Number of level-1 observations = 2352; Number of level-2 clusters = 118.

Table S8

Results of the Mixed Model conducted on d' scores obtained in Experiment 2S

Effect	*B*	*SE*	*df*	*t*	*p*
Fixed effects					
(Intercept)	0.39	(0.10)	118	3.95	< .001
CS variability	-0.38	(0.20)	118	-1.94	.055
US valence	-0.22	(0.16)	118	-1.40	.166
US valence × CS variability	-0.05	(0.31)	118	-0.15	.882
	Variance				
Random effects					
(Intercept)	0.41				

Note. Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for the mixed model conducted on *d'* scores. The model was specified as lmer(d' ~ CS variability * US valence+ (1|subject)) in R, using the lme4 package. Effect coding: CS variability (-0.5 invariable, 0.5 variable), US valence (-0.5 negative, 0.5 positive). All *p*-values in this table are two-tailed. Number of level-1 observations = 236; Number of level-2 clusters = 118.

Experiment 2

Table S9

Results of the Mixed Model conducted on Evaluative Ratings in Experiment 2

Effect	*B*	*SE*	*df*	*t*	*p*	β
Fixed effects						
(Intercept)	10.49	(1.61)	134.56	6.53	< .001	.03
US valence	29.65	(4.09)	148.61	7.24	< .001	.60
CS variability	-1.73	(3.21)	134.56	-0.54	.590	-.04
Stimulus Type	-7.15	(2.19)	739.73	-3.27	.001	-.14
US valence × CS variability	0.54	(8.19)	148.61	0.07	.947	.01
US valence × Stimulus Type	-14.53	(4.28)	1806.00	-3.39	.001	-.29
CS variability × Stimulus Type	2.80	(4.38)	739.73	0.64	.523	.06
US valence × Stimulus Type × CS Variability	26.23	(8.57)	1806.00	3.06	.002	.53

		Correlation with random effect for	
	Variance	Intercept	US valence
Random effects			
(Intercept)	175.59		
US valence	1491.91	-.33	
Stimulus Type	25.46	.21	.85

Note. See Experiment 1 for a description of the model. Number of level-1 observations = 2064; Number of level-2 clusters = 129.

Table S10

Results of the Mixed Model conducted on Evaluative Ratings of CS Components in Experiment 2

Effect	*B*	*SE*	*df*	*t*	*p*	β
Fixed effects						
(Intercept)	11.82	(1.68)	129.24	7.01	< .001	.00
US valence	19.69	(3.91)	129.01	5.03	< .001	.41
CS variability	-0.86	(3.37)	129.24	-0.26	.789	-.02
Component Type	1.02	(2.71)	600.96	0.38	.708	.02
US valence × CS variability	14.53	(7.83)	129.01	1.86	.066	.31
US valence × Component Type	-6.82	(5.39)	774.00	-1.26	.206	-.14
CS variability × Component Type	-9.15	(5.42)	600.96	-1.69	.092	-.19
US valence × Component Type × CS Variability	-17.87	(10.79)	774.00	-1.66	.098	-.38

		Correlation with random effect for	
	Variance	Intercept	US valence
Random effects			
(Intercept)	124.76		
US valence	985.86	-.39	
Component Type	8.97	.92	.01

Note. See Experiment 2S for a description of the model. Number of level-1 observations = 1032; Number of level-2 clusters = 129.

Table S11

Results of the Generalized Mixed Model conducted on Responses of the AMP in Experiment 2

Effect	*log-odds*	*OR*	*SE*	*Wald Z*	*p*
Fixed effects					
(Intercept)	0.22	1.24	(0.05)	3.97	< .001
US valence	0.73	2.07	(0.12)	5.95	< .001
CS variability	-0.19	0.83	(0.11)	-1.75	.080
Stimulus type	0.07	1.08	(0.11)	0.67	.503
US valence × CS variability	0.37	1.44	(0.24)	1.49	.136
US valence × Stimulus Type	-0.53	0.59	(0.22)	-2.41	.016
CS variability × Stimulus Type	0.25	1.29	(0.22)	1.16	.247
US valence × Stimulus Type × CS Variability	0.52	1.68	(0.44)	1.19	.235

		Correlation with random effect for	
	Variance	Intercept	US valence
Random effects			
(Intercept)	0.00		
US valence	0.38	-1.00	
Stimulus Type	0.01	1.00	-1.00

Note. See Experiment 1 for a description of the model. Number of level-1 observations = 2061; Number of level-2 clusters = 129.

Table S12

Results of the Generalized Mixed Model conducted on Responses of the Recognition Memory Task in Experiment 2

Effect	*log-odds*	*OR*	*SE*	*Wald Z*	*p*
Fixed effects					
(Intercept)	-0.64	0.53	(0.09)	-7.27	< .001
Stimulus Type I: Distractor vs. CS	4.06	57.80	(0.20)	20.66	< .001
Stimulus Type II: Distractor vs. GS	-0.44	0.65	(0.15)	-2.87	.004
CS variability: CS	-0.48	0.62	(0.25)	-1.93	.054
CS variability: GS	1.20	3.31	(0.16)	7.35	< .001
CS variability: Distractor	0.95	2.58	(0.34)	2.80	.005
	Variance				
Random effects					
(Intercept)	0.37				

Note. See Experiment 2S for a description of the model. Number of level-1 observations = 2580; Number of level-2 clusters = 129.

Table S13

Results of the Mixed Model conducted on d' scores obtained in Experiment 2

Effect	*B*	*SE*	*df*	*t*	*p*
Fixed effects					
(Intercept)	1.04	(0.06)	129	16.50	< .001
CS variability	-0.60	(0.13)	129	-4.74	< .001
US valence	-0.36	(0.12)	129	-2.94	.004
US valence × CS variability	-0.81	(0.24)	129	-3.30	< .001
	Variance				
Random effects					
(Intercept)	0.03				

Note. See Experiment 2S for a description of the model. Number of level-1 observations = 258; Number of level-2 clusters = 129.

Experiment 3

Table S14

Results of the Mixed Model conducted on Evaluative Ratings in Experiment 3

Effect	*B*	*SE*	*df*	*t*	*p*	β
Fixed effects						
(Intercept)	7.21	(1.35)	173.78	5.33	< .001	.02
US valence	38.22	(3.65)	191.11	10.48	< .001	.74
CS variability	-13.06	(2.70)	173.78	-4.83	< .001	-.25
Stimulus Type	-2.70	(1.87)	1895.26	-1.44	.149	-.05
US valence × CS variability	16.65	(7.29)	191.11	2.28	.024	.32
US valence × Stimulus Type	-23.64	(3.73)	2337.98	-6.34	< .001	-.46
CS variability × Stimulus Type	8.89	(3.74)	1895.26	2.38	.018	.17
US valence × Stimulus Type × CS Variability	13.73	(7.45)	2337.98	1.84	.066	.27

		Correlation with random effect for	
	Variance	Intercept	US valence
Random effects			
(Intercept)	158.82		
US valence	1622.77	-.50	
Stimulus Type	4.43	-.74	0.95

Note. See Experiment 1 for a description of the model. Number of level-1 observations = 2672; Number of level-2 clusters = 167.

Table S15

Results of the Mixed Model conducted on Evaluative Ratings of CS Components in Experiment 3

Effect	*B*	*SE*	*df*	*t*	*p*	β
Fixed effects						
(Intercept)	12.81	(1.69)	167.33	7.57	< .001	.01
US valence	29.33	(3.23)	171.93	9.08	< .001	.55
CS variability	-11.60	(3.38)	167.33	-3.43	.001	-.22
Component Type	4.32	(2.62)	544.02	1.65	.099	.08
US valence × CS variability	14.89	(6.46)	171.93	2.30	.022	.28
US valence × Component Type	-8.76	(5.00)	1001.99	-1.73	.084	-.16
CS variability × Component Type	-8.78	(5.23)	544.02	-1.68	.094	-.16
US valence × Component Type × CS Variability	-42.06	(10.13)	1001.99	-4.15	< .001	-.79

		Correlation with random effect for	
	Variance	Intercept	US valence
Random effects			
(Intercept)	208.31		
US valence	666.02	-.74	
Component Type	71.95	-.85	.98

Note. See Experiment 2S for a description of the model. Number of level-1 observations = 1336; Number of level-2 clusters = 167.

Table S16

Results of the Generalized Mixed Model conducted on Responses of the AMP in Experiment 3

Effect	*log-odds*	*OR*	*SE*	*Wald Z*	*p*
Fixed effects					
(Intercept)	0.31	1.36	(0.05)	5.90	< .001
US valence	0.64	1.90	(0.12)	5.36	< .001
CS variability	-0.17	0.84	(0.10)	-1.70	.090
Stimulus type	0.00	1.00	(0.10)	0.00	.997
US valence × CS variability	0.10	1.10	(0.24)	0.40	.686
US valence × Stimulus Type	-0.21	0.81	(0.19)	-1.08	.282
CS variability × Stimulus Type	-0.14	0.87	(0.19)	-0.73	.466
US valence × Stimulus Type × CS Variability	0.54	1.72	(0.38)	1.43	.154

		Correlation with random effect for	
	Variance	Intercept	US valence
Random effects			
(Intercept)	0.06		
US valence	0.83	-.51	
Stimulus Type	0.03	-.98	.31

Note. See Experiment 1 for a description of the model. Number of level-1 observations = 2667; Number of level-2 clusters = 167.

Table S17

Results of the Generalized Mixed Model conducted on Responses of the Recognition Memory Task in Experiment 3

Effect	*log-odds*	*OR*	*SE*	*Wald Z*	*p*
Fixed effects					
(Intercept)	-0.40	0.67	(0.07)	-5.45	< .001
Stimulus Type I: Distractor vs. CS	3.27	26.40	(0.16)	20.64	< .001
Stimulus Type II: Distractor vs. GS	-0.32	0.73	(0.12)	-2.68	.007
CS variability: CS	-0.84	0.43	0.22	-3.80	< .001
CS variability: GS	0.57	1.77	0.14	3.99	< .001
CS variability: Distractor	-0.49	0.61	0.25	-1.97	.049
	Variance				
Random effects					
(Intercept)	0.44				

Note. See Experiment 2S for a description of the model. Number of level-1 observations = 3335; Number of level-2 clusters = 167.

Table S18

Results of the Mixed Model conducted on d' scores obtained in Experiment 3

Effect	*B*	*SE*	*df*	*t*	*p*
Fixed effects					
(Intercept)	0.89	(0.07)	167	13.34	< .001
CS variability	-0.57	(0.13)	167	-4.29	< .001
US valence	-0.08	(0.13)	167	-0.65	.517
US valence × CS variability	-0.41	(0.26)	167	-1.62	.108
	Variance				
Random effects					
(Intercept)	0.06				

Note. See Experiment 2S for a description of the model. Number of level-1 observations = 334; Number of level-2 clusters = 167.

Supplement E

Supplement E includes a detailed description of the analyses of “old” versus “new” responses obtained in the recognition memory task in Experiments 2S, 2 and 3. An overview of the results is displayed in *Table 1* in the manuscript.

Recognition Memory Task, Experiment 2S

“Old” (0) and “new” (1) responses collected in the recognition memory task were analyzed for each type of stimulus (CSs, distractors, and GSs). Overall, participants recognition memory performance was high for CSs (*M* = .89, *SD* = .16; “old” responses are correct responses), and distractors (*M* = .08, *SD* = .15, “old” responses are false alarms), but around chance level for GSs (*M* = .60, *SD* = .23, “old” responses are false alarms). We tested the difference between learning conditions for each stimulus type. “Old” (1) versus “new” (0) responses were submitted to a generalized linear mixed-effect model with random by-subject intercepts. Simple slopes were calculated for CS variability at each level of the factor stimulus type. The four practice trails, as well as the trials that timed-out (0.33% of trials) were excluded from data analysis. Exponentiated parameter estimates are reported.

For CSs, the odds for responding “old” (correct responses) did not differ between CS variability conditions, *B* = 0.88, *SE* = 0.33, *z* = -0.41, *p* = .682. For both the distractors, *B* = 2.49, *SE* = 0.40, *z* = 2.27, *p* = .023, and GSs, *B* = 1.79, *SE* = 0.16, *z* = 3.54, *p* < .001, odds for responding “old” were higher in the *variable* than the *invariable* condition. Thus, false alarm rates were higher in the learning condition that included many CSs of a category, rather than a single one, while hit rates did not differ between the two conditions. This observation further substantiates the effects reported for the signal detection measure *d’* in the main manuscript.

Recognition Memory Task, Experiment 2

“Old” and “new” responses collected in the recognition memory task in Experiment 2 were analyzed using the same model as in Experiment 2S. There were no timed-out trials. For CSs, the odds for choosing the correct response (“old”) were lower in the *variable* than in the *invariable* condition, *B* = 0.62, *SE* = 0.25, *z* = -1.93, *p* = .054, even though this effect did not reach significance. Thus, participants made fewer errors in identifying CSs correctly as “old” when they were in the *invariable* condition, compared to the *variable* condition. For distractors, *B* = 2.58, *SE* = 0.34, *z* = 2.80, *p* = .005, and GSs, *B*= 3.31, *SE* = 0.16, *z* = 7.35, *p* < .001, odds for making “old” responses were higher in the *invariable* than the *variable* condition. This result represented higher false memory rates when experiencing many CSs per category, relative to the condition that presented only a single CS per category.

Recognition Memory Task, Experiment 3

Results of the recognition memory task of Experiment 3 were again analyzed with the same model as specified for Experiment 2S. Exponentiated parameter estimates are reported. Trials that timed out (0.15% of all trials) were excluded before calculating the model. For each simple slope, we set the respective stimulus type to 0 and report the fixed effect of CS variability. For CSs, odds for responding "old" were higher in the *invariable* than in the *variable* condition, *B* = 0.43, *SE* = 0.22, *z* = -3.80, *p* < .001, indicating higher "hit" rates when only one CS per category was included in learning. For distractors, "old" responses were also higher in the *invariable* than the *variable* condition, *B* = 0.61, *SE* = 0.25, *z* = -1.97, *p* = .049. Thus, participants in the *invariable* condition performed worse in correctly classifying distractors than participants in the *variable* condition. Lastly, for GSs the odds for responding "old" were higher in the *variable* than the *invariable* condition, *B* = 1.77, *SE* = 0.14, *z* = 3.99, *p* < .001, showing higher false memory rates for GSs when learning conditions allowed for the extraction of the predictive cues.

SUPPLEMENTARY MATERIALS

Chapter 2:

Are attitudes towards outgroup members more resistant to change? On the role of social categories in attitude change via evaluative conditioning

Kathrin Reichmann[1], I-Ching Lee[2], Mandy Hütter[1]
[1]*Eberhard Karls Universität Tübingen, Tübingen, Germany*
[2]*National Taiwan University, Taipei, Taiwan*

Supplement A

Supplement A includes the questions used to measure self-identification. They were translated to Chinese for Experiment 1, and to German for Experiment 2. Responses were obtained on a 7-point scale, with the endpoints "does not apply to me at all" and "applies to me very much".

Based on Leach et al. (2008):

I am glad to be Taiwanese [German].

I think that Taiwan [Germany] has a lot to be proud of.

It is pleasant to be Taiwanese [German].

Being Taiwanese [German] gives me a good feeling.

I have a lot in common with the average Taiwanese [German] person.

I am similar to the average Taiwanese [German] person.

Based on Lee et al. (2018):

I identify myself as Taiwanese [German].

Generally speaking, I enjoy being Taiwanese [German].

I feel personally criticized when others criticize Taiwan [Germany].

I like using "I am Taiwanese [German]" to express my identity.

Supplement B

Supplement B presents the instructions presented to participants in both experiments. Instructions were presented in Chinese in Experiment 1, and German in Experiment 2.

General Study Information

Welcome to the study! We are happy that you decided to participate. In this study, we are interested in the ways people process information and images, especially in social contexts.

Evaluative Priming Task

In the next task, you will see positive and negative adjectives. Please classify each adjective as either positive or negative. Please respond as quick and as accurate as you can. It is important that you try to be very fast, but also make very few mistakes.

Press the key 'a' if the adjective has a negative meaning.

Press the key 'l' if the adjective has a positive meaning.

You will notice that there are also some images appearing in between adjectives. You don’t have to pay attention to the images. Please put the index finger of your left hand on the 'a' key and the index finger of your right hand on the 'l' key.

Evaluative Conditioning Phase

You will now see every individual with either a positive, or a negative image. The images convey information about the individuals. We choose the images to help you form a correct impression of the person on the screen. Please watch closely. There will be surprise tasks coming up, so you should pay attention. You will have the chance to take two breaks and look away from the screen.

Continuous Evaluative Ratings

You will now see all people again. We would like to ask you to evaluate each person. Please use the scale to indicate how negative or positive your impression of the person is. You can rely on your gut feeling to do this. We are interested in your spontaneous reactions.

Recognition Memory Task [Experiment 2 only]

A short task will follow about the presented people. Please indicate for every individual whether you just saw them or not. Use the keys x and m for this purpose.

Press the key 'x' if you have never seen the person before.

Press the key 'm' if you have seen the person earlier.

Please place the index finger of your left hand on the 'x' key and the index finger of your right hand on the 'm' key. React to each person as quickly as possible. You have two seconds to respond.

Supplement C

In Supplement C, we present all parameter coefficients of fixed and random effects of the mixed models conducted for dependent measure (evaluative priming task, continuous evaluative ratings, recognition memory) for Experiments 1 and 2.

Experiment 1

Table S1

Results of the mixed model conducted on EPT scores in Experiment 1

Effect	*B*	*SE*	*df*	*t*	*p*
Fixed effects					
(Intercept)	13.60	3.32	76.41	4.23	.000
US valence (Pre)	- 0.12	5.12	705.86	0.07	.981
Time of Measurement	2.48	4.50	76.80	0.73	.564
Social Category (Pre)	4.35	5.25	317.04	0.77	.381
US valence × Time of Measurement	1.95	6.90	1090.89	0.28	.778
US valence × Social Category (Pre)	-9.78	9.76	1091.05	-1.00	.317
Social Category × Time of Measurement	-2.12	6.90	1090.89	-0.31	.758
US valence × Social Category × Time of Measurement	18.61	13.80	1090.89	1.35	.178

		Correlation with random effect for		
	Variance	Intercept	Time of Measurement	US Valence
Random effects				
(Intercept)	386.35			
Time of Measurement	543.81	-0.67		
US Valence	22.43	-0.34	-0.21	
Social Category	62.64	0.44	-0.30	-0.81

Note. Fixed effects estimates (top) and variance-covariance estimates (bottom) for the mixed model conducted on evaluative priming scores (higher scores indicate more positive indirect evaluations). The model was specified as lmer(EPTscore ~ time * US valence * social category + (time + US valence + social category | subject) in R, using the lme4 package. Effect coding: US valence (-0.5 negative, 0.5 positive), social category (-

0.5 outgroup, 0.5 ingroup). Dummy coding: time of measurement (0 pre-conditioning, 1 post conditioning). All *p*-values in this table are two-tailed. Number of level-1 observations = 1343; Number of level-2 clusters = 84.

Table S2

Results of the mixed model conducted on continuous evaluative ratings in Experiment 1

Effect	*B*	*SE*	*df*	*t*	*p*	β
Fixed effects						
(Intercept)	5.35	2.85	84	1.88	.064	0.00
US valence	13.14	3.43	84	3.90	.000	0.30
Social Category	9.33	3.12	84	2.99	.004	0.21
US valence × Social Category	-3.60	5.57	420	- 0.64	.519	-0.08

		Correlation with random effect for	
	Variance	Intercept	US Valence
Random effects			
(Intercept)	519.1		
US Valence	333.1	-0.01	
Social Category	163.2	-0.31	-0.11

Note. Fixed effects estimates (top) and variance-covariance estimates (bottom) for the mixed model conducted on direct evaluative ratings (-100 to 100), collected post conditioning. The model was specified as lmer(ratings ~ USvalence * social group + (USvalence + social category | subject) in R, using the lme4 package. Effect coding: US valence (-0.5 negative, 0.5 positive), social category (-0.5 outgroup, 0.5 ingroup). All *p*-values in this table are two-tailed. Number of level-1 observations = 672; Number of level-2 clusters = 84.

Experiment 2

Table S3

Results of the mixed model conducted on continuous evaluative ratings in Experiment 2

Effect	*B*	*SE*	*df*	*t*	*p*	β
Fixed effects						
(Intercept)	11.70	1.91	130.78	6.13	.000	0.03
US valence (Pre)	-1.36	1.68	311.88	-0.81	.417	-0.04
Time of Measurement	-2.32	1.07	3219.00	-2.17	.030	-0.06
Social Category (Pre)	-11.26	1.96	224.55	-5.75	.000	-0.30
US valence × Time of Measurement	10.20	2.14	3219.00	4.76	.000	0.27
US valence × Social Category (Pre)	-1.65	3.03	3219.00	-0.54	.586	-0.04
Social Category × Time of Measurement	2.96	2.14	3219.00	1.38	.167	0.08
US valence × Social Category × Time of Measurement	6.38	4.28	3219.00	1.49	.137	0.17

		Correlation with random effect for	
	Variance	Intercept	US Valence
Random effects			
(Intercept)	339.89		
US Valence	57.18	-0.25	
Social Category	170.89	-0.06	0.34

Note. The model was specified as lmer(ratings ~ time * US valence * social category + (US valence + social category| subject). Effect coding: US valence (-0.5 negative, 0.5 positive), social category (-0.5 outgroup, 0.5 ingroup). Dummy coding: time of measurement (0 pre-conditioning, 1 post conditioning). All *p*-values in this table are two-tailed. Number of level-1 observations = 3552; Number of level-2 clusters = 111.

Table S4

Results of the mixed model conducted on EPT scores in Experiment 2

Effect	*B*	*SE*	*df*	*t*	*p*
Fixed effects					
(Intercept)	19.96	3.25	104.06	6.14	.000
US valence	-1.84	4.18	669.76	-0.44	.660
Social Category	-2.65	4.19	619.38	-0.63	.527
US valence × Social Category	20.45	8.34	727.16	2.45	.014

		Correlation with random effect for	
	Variance	Intercept	US Valence
Random effects			
(Intercept)	646.79		
US Valence	8.30	-1.00	
Social Category	18.20	1.00	-1.00

Note. Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for the mixed model on evaluative priming scores (higher scores indicate more positive indirect evaluations. The model was specified as lmer(ratings ~ USvalence * social group + (USvalence + social category | subject) in R, using the lme4 package. Effect coding: US valence (-0.5 negative, 0.5 positive), social category (-0.5 outgroup, 0.5 ingroup). All *p*-values in this table are two-tailed. Number of level-1 observations = 831; Number of level-2 clusters = 104.

Table S5

Results of the mixed model conducted on the responses of the recognition memory task in Experiment 3, collected post conditioning

Effect	*log-odds*	*OR*	*SE*	*Wald Z*	*p*
Fixed effects					
(Intercept)	2.22	9.20	0.10	23.02	.000
US valence	0.06	1.06	0.16	0.36	.717
Social Category	0.47	1.59	0.18	2.55	.011
Stimulus Type	-4.72	0.01	0.14	-33.01	.000
US valence × Social Category	0.45	1.56	0.32	1.40	.161
US valence × Stimulus Type	0.10	1.10	0.24	0.40	.689
Social Category × Stimulus Type	-0.94	0.39	0.29	-3.28	.001
US valence × Stimulus Type × Social Category	0.37	1.44	0.48	0.76	.447

		Correlation with random effect for
	Variance	Intercept
Random effects		
(Intercept)	0.16	
Social Category	0.21	0.48

Note. Fixed effects estimates (top) and variance-covariance estimates (bottom) for the generalized mixed model conducted on new (0) and old (1) responses collected in the recognition memory task. The model was specified as glmer(response ~ stimulus type * social category * US valence + (social category | subject). Effect coding: US valence (-0.5 negative, 0.5 positive), social category (-0.5 outgroup, 0.5 ingroup). Dummy coding: stimulus type (0 CS, 1 lure). Number of level-1 observations = 3464; Number of level-2 clusters = 111.

SUPPLEMENTARY MATERIALS

Chapter 3:

Abstract representations of attitudes: Do they make evaluative conditioning resistant to US revaluation? A study of ecological conditions

Kathrin Reichmann, & Mandy Hütter
Eberhard Karls Universität Tübingen

Supplement A

Supplement A presents the instructions provided to participants before the conditioning phase, the US revaluation phase, and the direct evaluative ratings. Instructions are translated from German.

Conditioning Phase

Experiment 1, one-to-many [one-to-one]

Dear participant, welcome to our study! We are happy that you decided to participate. The present study is about the way people form impressions and how finite impressions are. We will now present you various logos on the screen. The logos were created by several people in a joint effort as part of a graphic workshop. [*one-to-one*: The logos were created by different people as part of a graphic workshop]. In the following, you will now see the newly designed logo and right after that, one of the people who was involved in the design of the logo [*one-to-one*: the person who designed the logo]. From time to time, people who did not participate in the workshop will appear. These people appear separately, without following a logo. CAUTION! It is important that you remember well which logo was created by whom. We will later ask you about your impression of the different logos. In between, a short task will test whether you are still paying attention. Let’s get started!

Experiment 2

Dear participant, welcome to our study! We are happy that you decided to participate. You will now take part in a simple perception task. For this purpose, various logos will be presented on the screen. The logos were created by several people in a joint effort as part of a graphic workshop. Accompanying each logo, you will always see the image of one of the people who participated in the design of the respective logo. Please attend to the people and logos carefully. [*Experiment 3*: First, familiarize yourself with the logos and the people. We will then ask you about your impression of the various logos. Please look at the presentation attentively. If you miss the presentation of some logos, you will be less familiar with them, which could affect the results.] In between learning blocks, there are small attention checks. Here, you should click a button on the screen in time, so stay alert. You will have three opportunities to take a short break. [*Experiment 3*: CAUTION! You will see each person just once. It is important that you memorize the people well! Let’s get started!] Let’s get started!

US Revaluation

In the following, you will receive additional information about the people you saw earlier. We want to study whether the first impression about a person can be revised by explicitly learned, new information. You will shortly see each person together with information that describes the person in more detail. Your task is to carefully read the information about the people and try to remember it. For this purpose, you will see each person and the additional information two times. Once you have read the information about a person, go to the next one by clicking 'Next'. Take as much time as you need while reading the statements.

Evaluative Ratings

Now we would like to ask you to indicate on a continuous scale how pleasant or unpleasant your impression of the depicted logo is. We are interested in your spontaneous reaction.

Supplement B

In Supplement B, we present all parameter coefficients of fixed and random effects of the mixed models conducted for each experiment on direct evaluative ratings (collected on a scale from -100 to 100).

Experiment 1

Table S1

Results of the Mixed Model conducted on Evaluative Ratings in Experiment 1

Effect	*B*	*SE*	*df*	*t*	*p*	β
Fixed effects						
(Intercept)	18.39	2.59	54.14	7.11	< .001	0.00
US valence	23.73	2.79	697.04	8.50	< .001	0.49
Pairing schedule	-6.03	3.24	242.06	-1.86	.064	-0.12
US revaluation	-2.19	2.80	699.06	-0.78	.434	-0.04
US valence × Pairing schedule	18.11	5.66	714.25	3.20	.001	0.37
US valence × US revaluation	-23.25	5.61	703.61	-4.14	< .001	-0.48
Pairing schedule × US revaluation	-1.40	5.61	703.33	-0.25	.803	-0.03
US valence × Pairing schedule × US revaluation	23.83	11.24	706.43	2.12	.034	0.49

	Variance
Random effects	
Participants (Intercept)	151.40
CS (Intercept)	192.20

Note. lmer(ratings ~ USvalence * revaluation * pairing schedule + (1| subject) + (1|CS)) in R, using the lme4 package (Bates et al., 2015). Effect coding: US valence (-0.5 neg, 0.5 pos), US revaluation (-0.5 congruent, 0.5 incongruent), pairing schedule (-0.5 one-to-one, 0.5 one-to-many). All *p*-values in this table are two-tailed. Number of level-1 observations = 952; Number of level-2 clusters (subjects) = 238; Number of level-2 clusters (CSs) = 47.

Experiment 2

Table S2

Results of the Mixed Model conducted on Evaluative Ratings in Experiment 2

Effect	*B*	*SE*	*df*	*t*	*p*	β
Fixed effects						
(Intercept)	7.85	2.80	53.18	2.81	.007	-0.01
US valence	31.50	2.48	1211.54	12.68	< .001	0.60
Sequencing	2.33	2.55	180.35	0.91	.362	0.04
US revaluation	0.00	2.48	1207.53	0.00	.999	0.00
US valence × Sequencing	6.79	4.97	1212.50	1.36	.173	0.13
US valence × US revaluation	-12.35	5.00	1221.56	-2.47	.014	-0.24
Sequencing × US revaluation	3.57	5.00	1220.85	0.71	.475	0.07
US valence × Sequencing × US revaluation	5.39	9.91	1207.82	0.54	.587	0.10

	Variance
Random effects	
Participants (Intercept)	13.15
CS (Intercept)	335.83

Note. Effect coding: US valence (-0.5 neg, 0.5 pos), US revaluation (-0.5 congruent, 0.5 incongruent), sequencing (-0.5 backward, 0.5 forward). Number of level-1 observations = 1432; Number of level-2 clusters (subjects) = 179; Number of level-2 clusters (CSs) = 54.

Experiment 3

Table S3

Results of the Mixed Model conducted on Evaluative Ratings in Experiment 3

Effect	*B*	*SE*	*df*	*t*	*p*	β
Fixed effects						
(Intercept)	11.01	2.85	53.85	3.86	< .001	0.01
US valence	35.54	2.69	1199.73	13.19	< .001	0.66
Sequencing	3.34	2.70	1203.36	1.24	.216	0.06
US revaluation	-5.31	2.71	1204.90	-1.96	.050	-0.10
US valence × Sequencing	-12.02	5.41	1203.26	-2.22	.026	-0.22
US valence × US revaluation	-6.99	5.42	1206.34	-1.29	.198	-0.13
Sequencing × US revaluation	0.68	5.42	1207.77	0.13	.900	0.01
US valence × Sequencing × US revaluation	-15.22	10.77	1199.32	-1.41	.158	-0.28

	Variance
Random effects	
Participants (Intercept)	0.00
CS (Intercept)	341.20

Note. Number of level-1 observations = 1236; Number of level-2 clusters (subjects) = 309; Number of level-2 clusters (CSs) = 54.

Supplement C

Supplement C includes tables with the fixed and random parameter coefficients of the models fitted to the memory data of Experiments 1 and 3. In the task, participants had to select the CS out of four CSs that a US occurred with. Responses were coded as either correct (1) or incorrect (0).

Experiment 1

Table S4

Results of the Generalized Mixed Model conducted on Memory Data in Experiment 1

Effect	*log-odds*	*OR*	*SE*	*Wald Z*	*p*
Fixed effects					
(Intercept)	0.43	1.54	0.08	5.24	< .001
US valence	0.14	1.15	0.12	1.15	.249
Pairing schedule	-2.05	7.76	0.16	-12.86	< .001
US revaluation	0.05	1.05	0.12	0.40	.692
US valence × Pairing schedule	-0.15	1.16	0.25	-0.60	.546
US valence × US revaluation	-0.24	0.78	0.25	-0.98	.330
Pairing schedule × US revaluation	-0.01	1.01	0.25	-0.03	.980
US valence × Pairing schedule × US revaluation	0.09	0.92	0.50	0.18	.860

	Variance
Random effects	
Participants (Intercept)	0.45
CS (Intercept)	0.03

Note. glmer(responseCorrect ~ US valence * pairing schedule * US revaluation + (1| subject) + (1|CS)) in R, using the lme4 package (Bates et al., 2015). Effect coding: US valence (-0.5 neg, 0.5 pos), US revaluation (-0.5 congruent, 0.5 incongruent), pairing schedule (-0.5 one-to-one, 0.5 one-to-many). All *p*-values in this table are two-tailed. Number of level-1 observations = 2824; Number of level-2 clusters (subjects) = 238; Number of level-2 clusters (CSs) = 47.

Experiment 3

Table S5

Results of the Generalized Mixed Model conducted on Memory Data in Experiment 3

Effect	*log-odds*	*OR*	*SE*	*Wald Z*	*p*
Fixed effects					
(Intercept)	-0.39	0.68	0.04	-9.74	< .001
US valence	0.08	1.08	0.05	1.58	.114
Sequencing	-0.23	0.80	0.08	-2.89	.004
US revaluation	-0.12	0.88	0.05	-2.52	.012
US valence × Sequencing	0.10	1.10	0.10	1.01	.314
US valence × US revaluation	-0.31	0.73	0.10	-3.16	.002
Sequencing × US revaluation	-0.01	0.99	0.10	-0.06	.949
US valence × Sequencing × US revaluation	0.00	1.00	0.20	-0.01	.988

	Variance
Random effects	
Participants (Intercept)	0.30
CS (Intercept)	0.00

Note. Effect coding: US valence (-0.5 neg, 0.5 pos), US revaluation (-0.5 congruent, 0.5 incongruent), sequencing (-0.5 backward, 0.5 forward). Number of level-1 observations = 7416; Number of level-2 clusters (subjects) = 309; Number of level-2 clusters (CSs) = 54.