School of Psychology

An Investigation of Assimilation and Contrast Effects in Backward Evaluative Conditioning

Luke John Stanley Green

0000-0002-1756-2048

This thesis is presented for the degree of

Doctor of Philosophy

of

Curtin University

August 2020

Declaration

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made. This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

The research presented and reported in this thesis was conducted in accordance with the National Health and Medical Research Council National Statement on Ethical Conduct in Human Research (2007) – updated March 2014. The proposed research study received human research ethics approval from the Curtin University Human Research Ethics Committee (EC00262), Approval Number # HRE2017-0776

Signature: …….

Date: ……..……

Abstract

Valence, which refers to how much we like or dislike stimuli, people, and events, has been shown to influence many aspects of our lives – from making important decisions, to forming attitudes about social, cultural, and political issues, and even as a potential barrier to successful treatment for anxiety disorders. Thus, understanding how and under which circumstances valence can be acquired and changed is an important psychological question. One way that valence can change is by repeatedly pairing a neutral stimulus, known as a conditional stimulus (CS), with a valenced stimulus, known as an unconditional stimulus (US). This process is called evaluative conditioning (EC) and can be easily modelled in lab studies by presenting neutral and valenced stimuli, such as pictures and sounds. Evaluative conditioning is measured using both self-report measures, such as explicit valence ratings, and implicit measures, such as affective priming and the startle blink reflex. In most circumstances, evaluative conditioning results in assimilation effects, in that the CS acquires the valence of the US. However, under certain circumstances contrast effects can occur and the CS acquires valence that is opposite to that of the US. In the current thesis, I investigated under which conditions assimilation and contrast effects occur in evaluative backward conditioning, when the US is presented before the CS (i.e. US-CS). Both backward conditioning only (US-CS) and concurrent forward and backward conditioning (CS-US-CS) procedures with picture and sound USs were used to assess the effects that instructional manipulations, affective relief, US intensity, US predictability, and US offset predictability have on the acquisition of backward CS valence.

Previous picture-picture paradigms using instructional manipulations that highlight the relationship between CSs and USs revealed backward CS contrast effects in concurrent forward and backward evaluative conditioning designs (CS-US-CS), whereas assimilation effects emerged in simple backward designs (US-CS) without such instructions. Chapter 2 examined whether affective relief at US offset was responsible for these backward CS contrast effects without the need for an instructional manipulation and whether past discrepant results were potentially due to different procedural features between conditioning paradigms. Contrary to our predictions, only the groups receiving instructions about CS-US/US-CS relations exhibited backward CS contrast effects, regardless of the conditioning procedure

employed. Thus, affective relief at US offset was not responsible for backward CS contrast effects in picture-picture paradigms. As previous fear conditioning research has shown that affective relief experienced at the offset of aversive USs elicits backward CS contrast effects without instructional manipulations, it is possible that the picture USs employed here were not intense enough to trigger affective relief.

Chapter 3 examined whether previously shown opposing effects on explicit and implicit measures of backward CS valence occur as a function of US intensity in concurrent forward and backward conditioning designs (CS-US-CS; Experiment 1). Instructional manipulations that highlighted the relationship between CSs and USs were presented and pleasant and unpleasant sounds of differing intensities were used as USs. The startle blink reflex was measured during acquisition and explicit valence ratings were recorded after acquisition. Startle responding in Experiment 1 revealed that both US intensities elicited affective relief (i.e. contrast effects), though greater affective relief was observed at the offset of high intensity USs compared to low intensity USs. Additionally, and contrary to our predictions, backward CS contrast effects were observed on explicit valence ratings in only the high intensity group. A second experiment assessed whether the offset of a pleasant US could elicit an opponent emotional response mirroring affective relief (i.e. disappointment) by adding a neutral US in a simple backward conditioning design (US-CS) using similar instructions as Experiment 1. Startle responding during backward CSs presented at the offset of the pleasant US was larger than startle during backward CSs after neutral and unpleasant USs, confirming our prediction that disappointment occurs after a pleasant event. Unexpectedly, the explicit valence ratings revealed assimilation effects, even though relational instructions had been employed. Taken together, these experiments suggest that instructional manipulations may have no bearing on backward CS valence ratings when US onset is unpredictable in a simple backward conditioning design (US-CS). They also showed that opponent emotional responses occurred at the offset of pleasant and unpleasant USs, resulting in backward CS contrast effects on the startle blink reflex regardless of US intensity or US predictability (CS-US-CS vs US-CS). Finally, US intensity within the same modality could not explain why previous research found both assimilation and contrast effects on implicit measures of CS valence.

Chapter 4 examined whether instructional manipulations highlighting the role of the CSs in starting or stopping the USs were responsible for previous startle

modulation results when using sound USs and whether the instructions were required to observe backward CS contrast effects on explicit valence ratings. As predicted, backward CS contrast effects were observed for both instruction groups for startle modulation and only in the group presented with relational instructions for explicit valence ratings. These findings suggest that startle modulation is driven by the relief and disappointment experienced at US offset, whilst explicit valence ratings vary as a function of the instructions presented before the conditioning task, as demonstrated in previous picture-picture paradigms.

Chapter 5 examined whether increasing US offset predictability would elicit backward CS contrast effects when no instructions highlighting CS-US/US-CS relations were presented in a picture-picture paradigm. US offset predictability was manipulated by varying US durations or overlapping backward CS onset with US presentations. Experiment 1 used a backward conditioning only design (US-CS) and Experiment 2 used a concurrent forward and backward conditioning design (CS-US-CS) in order to compare backward CS valence acquisition as a function of US onset predictability. Assimilation effects were found for backward conditioning in both experiments with a larger effect in Experiment 1 (US-CS design). These findings suggest that manipulating US offset predictability does not elicit backward CS contrast effects. Moreover, they show that presenting a concurrent forward CS (CS-US-CS) inhibits backward CS acquisition in comparison to a backward conditioning only design (US-CS).

In the current thesis, backward CS contrast effects were observed on explicit valence ratings only when instructional manipulations highlighting CS-US/US-CS relations were presented. Affective relief and US offset predictability did not influence backward CS acquisition on valence ratings. However, US onset predictability appeared to moderate the size of backward CS assimilation effects and reduce the effectiveness of relational instructions in producing contrast effects. Startle modulation was not influenced by instructions, US intensity, or US onset predictability, suggesting that affective relief and disappointment at US offset were responsible for the startle blink results. These findings suggest that explicit valence ratings and the startle blink reflex may be measuring different aspects of the acquisition experience. Variations in results across experiments suggest that characteristics of the US such as US onset predictability, US intensity, and US

modality, may have a larger effect on backward CS valence acquisition than previously thought.

Acknowledgements

Firstly, I would like to thank my supervisor, Professor Ottmar Lipp, for his support and guidance throughout my PhD. Thank you for all the opportunities you've provided me and for showing me what the life of a researcher entails – from the pragmatic mindset required to complete experimental research, to the frustrations and satisfactions involved with publishing empirical work, and the enjoyment (and stress!) that comes with attending and presenting research at national and international conferences. I very much appreciate you sharing your knowledge and experience with me and helping me to learn and grow as an early career researcher.

I would also like to thank my associate supervisor, Dr Camilla Luck, for the day to day encouragement and for always being there to answer my questions, discuss learning theories, talk through ideas, troubleshoot problems, and for teaching me about the world of research. Thank you for your friendship and support, and for everything else you've done for me during my PhD journey.

Thank you to my fellow PhD students – Sophie, Sofie, Enrique, and Rachel, for sharing your knowledge and wisdom with me and for all the laughs we've shared during both the pleasant and unpleasant times. I would also like to thank Welber Marinovic for his positivity and enthusiasm for all things research – thank you for enforcing the use of R and instilling in me the benefits of learning how to program.

Thank you to my parents for their love and support during my PhD (and before!). Thank you for always believing in me and for encouraging me to chase my dreams. Thank you to my Dad for the Saturdays and Satur-nights spent in the studio mixing, talking, and listening to music. Those times helped me keep things in perspective and turn off my PhD brain. And, thank you to my Mum, for cooking for us and putting up with the noise and the late nights that ensued from these times!

Finally, I would like to thank my wife, Tamara Green. I doubt that I would have survived this experience without you. Thank you for listening to me ramble about my research even when you were tired or had no idea what I was talking about. Thank you for running our lives so I could focus on my PhD, and thank you for not making me feel bad about gallivanting around the world presenting my research. I love you and am forever grateful you are my wife.

List of Publications Included as Part of the Thesis

Green, L. J. S., Luck, C. C., Gawronski, B., & Lipp, O. V. (2019). Contrast effects in backward evaluative conditioning: Exploring effects of affective relief/disappointment versus instructional information. *Emotion*. Advance online publication. https://dx.doi.org/10.1037/emo0000701

Green, L. J. S., Luck, C., & Lipp, O. V. (2020). How disappointing: Startle modulation reveals conditional stimuli presented after pleasant unconditional stimuli acquire negative valence. *Psychophysiology, 57*(8), 1-16. https://dx.doi.org/10.1111/psyp.13563

Green, L. J. S., Luck, C., & Lipp, O. V. (in press). Startle modulation during backward conditioning is not modulated by instructions. *Psychophysiology.* https://dx.doi.org/10.1111/psyp.13679

I warrant that I have obtained, where necessary, permission from the copyright owners to use any third-party copyright material reproduced in the thesis, or to use any of my own published work in which the copyright is held by another party.

Attribution Statements of Co-Authored Works

The following acknowledgements show the input of the candidate and the co-authors for each empirical chapter.

…….	…….
Luke Green	Ottmar Lipp
(Candidate)	(Supervisor)

Chapter 2: Green, L. J. S., Luck, C. C., Gawronski, B., & Lipp, O. V. (2019). Contrast effects in backward evaluative conditioning: Exploring effects of affective relief/disappointment versus instructional information. *Emotion*. Advance online publication. https://dx.doi.org/10.1037/emo0000701

	Conception and design	**Data acquisition**	**Data conditioning and manipulation**	**Data analysis and statistical method**	**Interpretation and discussion**
Luke Green	✓	✓	✓	✓	✓
Co-Author 1: Camilla Luck	✓				✓
Co-Author 1 Acknowledgment: I acknowledge that these represent my contribution to the above research output Signed:					
Co-Author 2: Bertram Gawronski				✓	✓
Co-Author 2 Acknowledgment: I acknowledge that these represent my contribution to the above research output Signed:					
Co-Author 3: Ottmar Lipp	✓			✓	✓
Co-Author 3 Acknowledgment: I acknowledge that these represent my contribution to the above research output Signed:					

Chapter 3: Green, L. J. S., Luck, C. C., & Lipp, O. V. (2020). How disappointing: Startle modulation reveals conditional stimuli presented after pleasant unconditional stimuli acquire negative valence. *Psychophysiology*, *57*(8), 1-16. https://dx.doi.org/10.1111/psyp.13563

	Conception and design	**Data acquisition**	**Data conditioning and manipulation**	**Data analysis and statistical method**	**Interpretation and discussion**
Luke Green	✓	✓	✓	✓	✓
Co-Author 1: Camilla Luck	✓				✓
Co-Author 1 Acknowledgment: I acknowledge that these represent my contribution to the above research output Signed:					
Co-Author 2: Ottmar Lipp	✓				✓
Co-Author 2 Acknowledgment: I acknowledge that these represent my contribution to the above research output Signed:					

Chapter 4: Green, L. J. S., Luck, C. C., & Lipp, O. V. (2020). Startle during backward evaluative conditioning is not modulated by instructions.

	Conception and design	**Data acquisition**	**Data conditioning and manipulation**	**Data analysis and statistical method**	**Interpretation and discussion**
Luke Green	✓	✓	✓	✓	✓
Co-Author 1: Camilla Luck	✓				✓
Co-Author 1 Acknowledgment: I acknowledge that these represent my contribution to the above research output Signed:					
Co-Author 2: Ottmar Lipp	✓				✓
Co-Author 2 Acknowledgment: I acknowledge that these represent my contribution to the above research output Signed:					

Chapter 5: Green, L. J. S., Luck, C. C., & Lipp, O. V. (2020). Assimilation or contrast effects in backward evaluative conditioning: The role of US offset predictability.

	Conception and design	**Data acquisition**	**Data conditioning and manipulation**	**Data analysis and statistical method**	**Interpretation and discussion**
Luke Green	✓	✓	✓	✓	✓
Co-Author 1: Camilla Luck	✓				✓
Co-Author 1 Acknowledgment: I acknowledge that these represent my contribution to the above research output Signed:					
Co-Author 2: Ottmar Lipp	✓				✓
Co-Author 2 Acknowledgment: I acknowledge that these represent my contribution to the above research output Signed:					

Funding Source Acknowledgement

This research was supported by an Australian Government Research Training Program (RTP) Scholarship.

Table of Contents

Chapter 1: Introduction

Evaluation of stimuli, people, and events is part of the human experience. One dimension on which we evaluate things is known as valence and refers to how much we like or dislike something. Every decision we make includes some kind of evaluation about the valence of certain aspects involved in the decision – whether it be the valence of the object we are making a decision about, i.e. do I like the colour of that shirt I want to buy (the object), or the valence of the decision itself, i.e. am I going to be pleased with my decision to buy that shirt. While buying a new shirt is a relatively trivial decision, valence has been shown to influence profound decisions, from romantic relationships to political affiliations and career aspirations (e.g., Galdi, Arcuri, & Gawronski, 2008; Gibson, 2008; LeBel & Campbell, 2009). Moreover, our likes and dislikes are believed to influence all facets of behaviour and have been implicated in the formation of attitudes that contribute to the perpetuation of racism and poor health related outcomes and have been suggested as a potential factor in determining the likelihood of relapse after successful anxiety treatment (Allport, 1935; Dirikx, Hermans, Vansteenwegen & Baeyens, 2004; Matsuda, Garcia, Catagnus, & Brandt. 2020; Sheeran et al., 2016; Walther, 2002; Zbozinek, Hermans, Prenoveau, Liao & Craske, 2015). Thus, understanding the processes and moderating factors responsible for the acquisition of valence and how it can be manipulated to benefit humankind is of the utmost importance in the current landscape of psychological science.

1.1 Valence Change Procedures

There are several procedures through which valence can be learned and manipulated (De Houwer, 2007). The most basic of these is known as mere exposure (Zajonc, 1968; see Moreland & Topolinski, 2010, for a review). The mere exposure effect refers to the increase in positive valence towards a stimulus after being repeatedly exposed to that stimulus. Another procedure that can change stimulus valence is known as operant evaluative conditioning and occurs when valence is transferred from an operant contingency to a neutral stimulus. For example, pressing a key that results in viewing a pleasant picture can change the valence of a neutral stimulus when that key press is later paired with the neutral stimulus (Eder, Krishna, & Van Dessel, 2019; see also Kawakami, Phills, Steele, & Dovidio, 2007). The most complex valence change procedure revolves around the idea of intersecting

regularities (Hughes, De Houwer, & Perugini, 2016). The notion of intersecting regularities suggests that evaluative learning can occur when any aspect of an operant contingency (i.e. stimulus, response, or outcome) or Pavlovian contingency intersects with any aspect of another operant contingency or Pavlovian contingency. For example, a stimulus (S1; "Press the 'J' key") followed by a response (R1; Participant presses the 'J' key) with two equal probability outcomes (O1 – neutral picture; O2 – valenced picture) can result in valence transferring from O2 to O1 because both outcomes intersect with the same stimulus (S1) and response (R1; see Hughes et al., 2016 for a thorough explanation of intersecting regularities). The valence change procedure of concern in this thesis is known as evaluative conditioning (EC; see Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010 for a review and meta-analysis). Evaluative conditioning procedures involve pairing a neutral stimulus known as a conditional stimulus (CS) with a valenced stimulus known as an unconditional stimulus (US) in an attempt to change CS valence (De Houwer, Thomas, & Baeyens, 2001; Hofmann et al., 2010). Importantly, while evaluative conditioning can be considered as a conditioning procedure, i.e. pairing valenced and neutral stimuli in a conditioning experiment, it can also be considered as an effect, i.e. the change in stimulus valence resulting from a conditioning procedure (De Houwer, 2007). This distinction is important because evaluative conditioning can occur from both evaluative conditioning procedures and other conditioning procedures such as fear conditioning, where the US is both fear inducing and evaluated as containing negative properties (Hofmann et al., 2010; Lipp, Neumann, & Mason, 2001; Zanna, Kiesler, & Pilkonis, 1970). Thus, the term evaluative conditioning can be used interchangeably to refer to both the procedure and the effect. However, of interest in this thesis is the effect (i.e. valence change) resulting from stimulus pairing regardless of whether it occurs in an evaluative or fear conditioning procedure.

1.2 Measuring Evaluative Conditioning

Changes in valence and other responses following conditioning can be assessed using explicit and implicit measures. Explicit measures involve asking participants to self-report on a specific criterion, such as stimulus valence (explicit valence ratings) or how afraid of a stimulus they are (self-reported fear ratings; De Houwer, 2007). These measures are quick and easy to implement with little to no burden to participants. However, they have the disadvantage of being subject to

demand characteristics. Implicit measures include reaction time based behavioural measures such as affective priming and the implicit association task (IAT), or physiological measures such as the startle blink reflex (De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009; Fazio, Jackson, Dunton, & Williams, 1995; Greenwald, McGhee, & Schwartz, 1998; Lang, Bradley, & Cuthbert, 1990). Reaction time based implicit measures capitalise on the fact that we are faster to categorise stimuli that are congruent in valence with each other than stimuli that are incongruent with each other. For example, in affective priming, participants are shown a prime stimulus (CS1 or CS2) followed by a target word that is either pleasant or unpleasant and instructed to categorise the valence of the word as quickly as possible. If CS1 is more pleasant than CS2, pleasant words should be categorised faster after CS1 than CS2. Implicit measures are less susceptible to demand characteristics than explicit measures, however, they are more time consuming to implement and more effortful for participants to complete. Implicit measures can also suffer from problems with low reliability, defining exactly what is meant by the term 'implicit', and the question of what processes are actually being assessed by these measures (i.e. Brownstein, Madva, & Gawronski, 2019; Corneille & Hutter, 2020; Corneille & Stahl, 2019; Gawronski & De Houwer, 2014; Gawronski, Morrison, Phills, & Galdi, 2017). The implicit physiological measure used to assess changes in stimulus valence is the startle blink reflex. The startle blink reflex is part of the defensive response exhibited by humans and other animals upon being presented with a startling stimulus such as a short and loud burst of white noise, a brief and bright flash of light, or electrical stimulation (Blumenthal et al., 2005; Lang et al., 1990). The startle reflex is accompanied by defensive postural changes including a protective eye-blink which can be measured via electromyography (EMG) of the orbicularis oculi muscle. The magnitude of the blink elicited by the startle eliciting probe has been shown to be larger when viewing negative pictures and smaller when viewing positive pictures, though generally only when using high arousal stimuli (Cuthbert, Bradley, & Lang, 1996). Thus, stimulus valence can be assessed by comparing blink magnitude between stimuli, such that larger startles are elicited during stimuli with negative valence than stimuli of neutral or positive valence. Overall, the startle blink reflex is a useful measure in that it provides a reliable physiological assessment of an organism's state, however, it is time consuming to implement and its interaction with arousal can lead to difficulties with interpretation.

1.3 Evaluative Conditioning Theories

Prominent theories of EC revolve around two different mechanisms thought to lead to valence change – associative mechanisms and propositional mechanisms (Gawronski & Bodenhausen, 2006, 2011, 2018; Mitchell, De Houwer & Lovibond, 2009). EC theories can be single process associative based theories, single process propositional based theories, or dual-process theories such as the associative-propositional evaluative (APE) model. Associative mechanisms of valence change work through the mere pairing of two stimuli in space or time resulting in the neutral stimulus acquiring the valence of the valenced stimulus. Associations are believed to be formed purely due to the concurrent activation of stimulus representations as links that form quickly and without assessment of the truth value of the relationship between them. Propositional mechanisms, on the other hand, are believed to result in valence change through deliberate and effortful evaluation of the truth value regarding the relationship between stimuli. This means that valence change occurs because we reason about how one stimulus is related to another stimulus, i.e. whether A causes B, A is an effect of B, or there is some other relationship beyond cause and effect between stimuli. Older theories of EC such as the referential account (Baeyens, Eelen, Crombez, & Van den Bergh, 1992), the holistic account (Levey & Martin, 1975; Martin & Levey, 1978, 1994), or the implicit misattribution account (Jones, Fazio, & Olson, 2009) are largely based on single process associative mechanisms, while the conceptual categorisation account (Davey, 1994; Field & Davey, 1999) is based on mental processes similar to single process propositional mechanisms. However, these older theories have less empirical support than current single process propositional (Mitchell et al., 2009) and dual process models (Gawronski & Bodenhausen, 2006, 2011, 2018). Moreover, they are not relevant to the current thesis and will not be discussed further (see Hoffman et al., 2010, for a summary of these theories).

Dual process theories from which the APE model is derived posit that both associative and propositional mechanisms play a role in evaluative conditioning (Gawronski & Bodenhausen, 2006, 2011, 2018). Recent research on EC has provided convincing support for the role of propositional mechanisms in changing stimulus valence (Moran & Bar-Anan, 2013; Hu, Gawronski, & Balas, 2017a; Hu, Gawronski, & Balas, 2017b). However, evidence for the role of associative mechanisms in changing stimulus valence is less convincing but has not been conclusively ruled out

(see Corneille & Stahl, 2019). It is worth noting that while theories surrounding associative and propositional mechanisms specifically are commonly tested in current EC research, the goal of this thesis was to investigate the mechanisms underlying valence change in general, not to test specific aspects of current EC theories. Thus, the experiments were not designed in such a manner as to specifically test propositional and associative mechanisms. However, as some of the findings from the thesis are relevant to the debate they will be considered in terms of associative and propositional mechanisms in the general discussion.

1.4 Evaluative Conditioning Effects

Staats and Staats (1957) provided one of the earliest demonstrations of evaluative conditioning by showing that syllables paired with positive and negative words acquired valence in the direction of the word they were paired with. Martin and Levey (1975) built upon this work with their original picture-picture paradigm in which participants sorted pictures into liked, neutral, and disliked categories before the neutral pictures were paired with the liked, neutral, and disliked pictures (CS-US). The valenced pictures were also presented before the neutral pictures in two conditions – disliked picture-neutral picture and liked picture-neutral picture. Presenting the valenced stimulus first in this manner is known as backward conditioning (i.e. US-CS), whereas presenting the neutral stimulus before the valenced stimulus as in the other conditions is known as forward conditioning (i.e. CS-US). Participants showed a preference for the neutral pictures forwardly paired with the liked pictures and disliked the neutral pictures forwardly paired with the disliked pictures. Numerically, the mean difference between backwardly paired CSs suggest evaluative conditioning may have occurred, though no significant difference was observed (potentially due to low power). Martin and Levey's (1975) picture-picture paradigm has been adapted over the years and has become a popular procedure in the EC literature that is still used today (Benedict & Gast, 2020; Hofmann et al., 2010).

The standard evaluative conditioning effect demonstrated for both forward and backward conditioning by Martin and Levey (1975) is known as an assimilation effect and has been shown in many publications since (Hofmann et al., 2010; Kim, Sweldens, & Hütter, 2016). The opposing effect, where the CS valence acquired is the opposite of the US valence has also been shown under certain circumstances for both forward and backward conditioning and is known as a contrast effect

(Andreatta, Mühlberger, Yarali, Gerber, & Pauli, 2010; Fiedler & Unkelbach, 2011; Moran & Bar-Anan, 2013). The main aim of the current thesis is to determine the mechanism/s responsible for contrast effects in backward conditioning, for which there appear to be two main candidates. The first of these is affective relief, which has been demonstrated in differential backward fear conditioning studies by Andreatta et al. (2010; Andreatta, Mühlberger, Glotzbach-Schoon, & Pauli, 2013). The second is driven by a propositional process resulting from instructional manipulations as shown by Moran and Bar-Anan (2013). These more recent studies of backward conditioning showing that both assimilation and contrast effects can emerge highlight the fact that backward conditioning procedures provide an avenue to investigate the processes responsible for evaluative conditioning and their boundary conditions. The following sections on affective relief and the use of instructional manipulations will address how these avenues of inquiry have been utilised so far in the literature, highlight potential alternative explanations for the results and processes at hand, and provide testable hypotheses relating to the processes and variables that may influence evaluative conditioning that will be assessed in this thesis.

1.5 Affective Relief

The offset of a painful or aversive stimulus elicits relief in both human and non-human animals (Becerra, Navratilova, Porreca, & Borsook, 2013; Porreca & Navratilova, 2017). Relief experienced at pain offset has been shown to increase with pain intensity and to depend on context (Leknes, Brooks, Wiech, & Tracey, 2008; Leknes, et al., 2013). Leknes et al. (2013) showed that a moderately painful stimulus elicited relief in a context where the other stimulus presented was intensely painful, but not in a context where the other stimulus presented was painless. Relief from pain has also been shown to activate similar brain areas to reward, though differences between the neural circuitry underlying relief and reward are apparent (Leknes, Lee, Berna, Andersson, & Tracey, 2011; Seymour et al., 2005).

Several accounts have been provided to explain what relief from pain reflects (Porreca & Navratilova, 2017). Franklin et al. (2010) suggests that relief from pain may be the representation of negative affective valence reducing towards neutral valence after cessation of the painful stimulus (see also Selby & Joiner, 2009), while Leknes et al. (2011) suggests that relief from pain results from an increase in positive valence due to the violation of negative expectations occurring from pain offset.

According to Franklin, Lee, Hanna, and Prinstein (2013), both of these perspectives are required to adequately conceptualise relief. In any case, relief from pain is a complex emotion that requires further research (Porreca & Navratilova, 2017). However, as this thesis is interested in the mechanisms underlying 'relief learning' rather than relief from pain, the rest of this section will focus on how affective relief can lead to stimuli acquiring positive properties, i.e. 'relief learning'.

Affective relief experienced at the offset of an aversive event can give rise to 'relief learning', which occurs when a neutral stimulus is presented during the period of relief (Deutsch, Smith, Kordts-Freudinger, & Reichardt, 2015; Gerber et al., 2014). The result of this stimulus pairing is a conditioned response to the neutral stimulus (backward CS) that opposes the valence of the aversive US – this means the backward CS acquires positive properties after being paired with affective relief. The term 'contrast effect' can also be used to describe the result of 'relief learning', as the backward CS acquires valence that is opposite to the US it was paired with. However, while 'relief learning' is used to describe a conditioned response following an aversive US only, contrast effects refer to changes in CS responding after pairings with either aversive or pleasant USs.

The phenomenon of 'relief learning' appears to follow the processes outlined by Solomon's (1980) opponent-process theory. According to Solomon (1980), the *a*-process, which is an emotional reaction to the US, elicits an opposing reaction, the *b*-process, which suppresses the initial reaction but is slow to initiate and slow to decay. In the case of 'relief learning', the negative emotional reaction to the aversive US (*a*-process) elicits the opposing reaction of 'relief' (*b*-process), resulting in 'relief learning' occurring to any stimulus presented after the US, i.e. the backward CS. Previous research with both humans and non-humans suggest that opponent-process theory provides a solid theoretical framework in which to understand the basis of 'relief learning' (Andreatta et al., 2010; Andreatta et al., 2013; Deutsch et al., 2015; Gerber et al., 2014).

Andreatta et al. (2010) demonstrated relief learning in humans in a differential fear conditioning experiment by presenting three groups with either forward, backward, or a control conditioning procedure. Geometric shapes were employed as CSs and an aversive shock was used as the US. In each group CSs were presented for 8 seconds and one CS was paired with the US (CS+) and one CS was presented alone (CS-). In the forward conditioning group, CS+ trials were CS

presentations followed immediately by the US. In the backward conditioning group the CS+ trials consisted of a US followed by a 6s gap followed by presentation of the backward CS. For the control group, CS+ trials mirrored the backward conditioning group but in a forward conditioning procedure, i.e. CSs were followed by a 6s gap and the US. Explicit valence ratings were taken before and after acquisition and the startle blink reflex was measured during the extinction phase. In the forward and backward conditioning groups, the CS+ was rated more negatively after training compared with pre-test and compared to the control group. A dissociation between explicit valence ratings and the startle blink reflex was found as blink magnitude was inhibited during CS presentations in the backward conditioning group compared to the CS- and compared to the average startle response of all groups. Thus, the physiological measure suggests that the backward CS acquired positive valence whereas the explicit measure suggests that the backward CS acquired negative valence.

Andreatta et al. (2010) interpreted their findings as potential support for a dual process theory of learning (Strack & Deutsch, 2004). It was suggested that relief learning was occurring as indexed by the startle reflex and that the negative valence acquired on explicit ratings was due to the pain of the shock making the event (US-CS) unpleasant overall. The 6s gap between US offset and backward CS onset raised the possibility that the backward CS was observed as a CS- instead of a backward CS+, meaning that the backward CS may have appeared as a separate stimulus that was not paired with the shock. It is possible that the dissociation between measures could have resulted from startle being measured during extinction and valence being measured before and after acquisition. Luck and Lipp (2017) investigated these alternatives by replicating the paradigm used by Andreatta et al. (2010) with three main differences. First, a 100ms gap between US offset and backward CS onset was included instead of a gap of 6s. Second, the startle reflex was measured during acquisition instead of extinction, and third, explicit valence was measured on a trial by trial basis. This means that startle and explicit valence were measured at close to the same time for each stimulus. Under these conditions a dissociation between startle and valence was still observed, suggesting that the dissociation found by Andreatta et al. (2010) was not the result of measuring valence at different time points.

Andreatta et al. (2013) replicated the paradigm used by Andreatta et al. (2010) with the addition of a forward CS before the shock US (CS-US-6s gap-CS). The addition of the forward CS served to make US onset predictable, which was believed may influence the expression of explicit valence. This was confirmed as backward CSs were rated as pleasant and startle during backward CSs was inhibited when US onset was made predictable, meaning both measures showed relief learning. Thus, the dissociation between explicit ratings and blink magnitude does not seem to occur when US onset is predictable in this paradigm. As discussed below, US onset predictability is one of the variables that may influence backward CS contrast effects independent of affective relief and will be addressed in this thesis.

1.6 Instructions and Evaluative Conditioning

Contrast effects have been shown to result from instructions in simple impression formation experiments and studies that manipulate the validity of stimulus pairings (Gregg, Seibt, & Banaji, 2006; Förderer & Unkelbach, 2012; Moran & Bar-Anan, 2013; Unkelbach & Fiedler, 2016). For example, Gregg et al. (2006) instructed participants to suppose that one fictitious social group embodied pleasant characteristics and behaved accordingly, while a second social group did not. As expected, the first social group was evaluated pleasantly while the second was evaluated as unpleasant. In a subsequent experiment after learning this information participants were told there had been a mistake and that the pleasant group was actually unpleasant and the unpleasant group was actually pleasant. A valence reversal was observed in that the initially pleasant group was rated as unpleasant and the initially unpleasant group was rated as pleasant. Förderer and Unkelbach (2012) demonstrated a similar pattern of results in a forward evaluative conditioning experiment where participants were shown pictures of men (CSs) followed by pictures of landscapes and animals (USs). With each CS-US pairing a relational qualifier was presented suggesting that the CS either loved or loathed the US that followed. When the CS loved a positive or negative US an assimilation effect was found, i.e. CSs loving positive USs were rated as pleasant and CSs loving negative USs were rated as unpleasant. When the CS loathed a positive or negative US a contrast effect was found in that CSs loathing positive USs were rated as unpleasant and CSs loathing negative USs were rated as pleasant.

The contrast effects described above have also been demonstrated on backwardly conditioned CSs as a result of instructional manipulations. Moran and Bar-Anan (2013) presented participants with a concurrent forward and backward conditioning procedure (CS-US-CS) where they showed pictures of cartoon aliens as CSs and played either a pleasant melody (USpos) or an unpleasant human scream (USneg) as the US. The CSs were four different sets of four different cartoon aliens (16 CSs in total), with each set differing in colour and head shape to signify family membership. Before the conditioning task, participants were instructed that each family would have a different role to play in the experiment and that their task was to learn what each family's role was for a memory test at the end. They were told that one family would start the melody, one would stop the melody, one would start the human scream, and that one would stop the human scream. CSs were presented for 2.5s and USs were played for between 10 and 30s. The forward CS was presented for 500ms before the US began playing and remained on the screen for a further 2s. The backward CS was presented 2s before the US stopped and remained on the screen for a further 500ms after the US ended. Explicit valence ratings were measured after the conditioning task. Assimilation effects were found for forward CSs, in that CSs paired with the melody were evaluated as pleasant and CSs paired with the scream were evaluated as unpleasant. Backward CSs showed contrast effects, as CSs following the melody were evaluated as unpleasant and CSs following the scream were evaluated as pleasant. These findings are in line with the instructions presented and demonstrate that propositional processes can change EC when the association between two stimuli should elicit an assimilation effect, i.e. CS paired with unpleasant US should elicit an unpleasant response to the CS. Moran, Bar-Anan, and Nosek (2016) replicated the results from Moran and Bar-Anan (2013) in a series of online experiments utilising a picture-picture paradigm with large sample sizes. In three experiments they used various terms to explain the relations between the CS and US, such as starts and ends, gives and takes away, and allows and prevents the unpleasant USs. These high powered replication studies suggest that backward CS contrast effects driven by instructional manipulations are robust and reliable and can be obtained with different stimuli, in the laboratory and online, and with different relational qualifiers.

It is worth noting that some of the studies reviewed here also employed implicit measures of CS valence that generally dissociated from explicit measures as

they revealed assimilation instead of contrast for backward CSs (Gregg et al., 2006; Moran & Bar-Anan, 2013). These studies were initially taken as support for dual process theories such as the APE model. However, recent studies have shown that implicit measures can also be shown to yield contrast effects under certain conditions such as presenting instructions before every trial (Hu et al., 2017a) and that the IAT used by Moran and Bar-Anan (2013) can be structured in such a way that either assimilation or contrast effects emerge (Bading, Stahl, & Rothermund, 2019). The debate surrounding the utility of implicit reaction time based measures in determining the underlying processes involved in EC is a lively one, however, and further discussion is beyond the scope of this thesis (see Brownstein, Madva, & Gawronski, 2019; Corneille & Stahl, 2019).

1.7 Thesis Aims

The aim of the current thesis was to determine under which conditions contrast effects could be elicited with backward conditioning. The results observed by Moran and Bar-Anan (2013) and Andreatta et al. (2010; 2013) provide several avenues of investigation upon which to answer this question. The main processes suggested to explain these results include affective relief and propositional reasoning determined by instructional manipulations. It is possible that affective relief elicits backward CS contrast effects on the startle blink reflex and potentially explicit valence ratings when US onset is predictable (Andreatta et al., 2013). However, relief learning and how it is expressed in different conditioning procedures and on explicit and implicit measures is not well understood. The instructional manipulations used by Moran and Bar-Anan (2013) also appear to be capable of producing backward CS contrast effects, although there are several plausible alternative explanations for these results, including affective relief, that suggest instructions may not be required for backward CS contrast effects to occur. To investigate these explanations eight experiments were performed across four studies that pit affective relief against instructional manipulations and introduce the possibility that procedural features such as US intensity, US/CS overlap, US onset predictability, and US offset predictability (among others) may also influence whether contrast effects occur during backward evaluative conditioning.

The literature reviewed above which addresses the outcome of backward conditioning across different fear and evaluative conditioning paradigms provides several avenues in which to investigate the processes responsible for backward CS

contrast effects and evaluative conditioning in general. Andreatta et al. (2010; 2013) show that backward CSs are evaluated as pleasant only when a concurrent forward CS is presented, an observation which is consistent with the backward CS contrast effects observed by Moran and Bar-Anan (2013) in their concurrent forward and backward conditioning design (CS-US-CS). However, there are several differences between these studies that may explain why backward CS contrast effects occur with and without instructions in a concurrent forward and backward conditioning design (CS-US-CS). First, Moran and Bar-Anan (2013) present an instructional manipulation that is most likely responsible for the observed backward CS contrast effect. However, as Andreatta et al. (2013) also found a contrast effect without an instructional manipulation it is possible that the concurrent forward and backward conditioning procedure (CS-US-CS) itself is conducive to backward CS contrast effects without an instructional manipulation. This would mean that findings similar to those of Moran and Bar-Anan (2013) may be obtained without an instructional manipulation. Second, the shock USs used by Andreatta et al. (2010; 2013) were arguably more intense than the sound and picture USs employed by Moran and Bar-Anan (2013) and Moran et al. (2016). Third, Moran and Bar-Anan (2013) overlapped CS and US presentations by 2s and varied the duration of the US on each trial, while Andreatta et al. (2010) had no US/CS overlap or variation in US duration. The empirical chapters that address the aforementioned procedural differences and their potential effects are outlined below.

1.8 Thesis Outline

The first paper empirical chapter of this thesis is titled ‘Contrast effects in backward evaluative conditioning: Exploring effects of affective relief/disappointment versus instructional information’. In this chapter two pilot studies and one main experiment were performed. The first pilot study compared the instructions from Moran and Bar-Anan (2013) with the instructions from Mallan, Lipp, and Libera (2008) to assess whether the instructions were responsible for the findings from Moran and Bar-Anan (2013). The second pilot study replicated Moran et al. (2016) as closely as possible because pilot study one did not show backward CS contrast effects in the Moran and Bar-Anan (2013) instructions condition. The main experiment compared the picture-picture paradigms and instructions from Moran et al. (2016) and Mallan et al. (2008) to determine whether the instructions, paradigm, or affective relief alone could account for backward CS contrast effects.

The second empirical chapter of this thesis is called ‘How disappointing: Startle modulation reveals conditional stimuli presented after pleasant unconditional stimuli acquire negative valence’. In the first experiment the effects of US intensity on backward CS contrast effects were investigated in a lab study measuring the startle blink reflex and explicit valence ratings. The aim of this study was to determine whether higher intensity USs were more conducive to backward CS contrast effects than lower intensity USs. In the second experiment a backward conditioning only procedure was used. The aim was to see whether the same pattern of results would occur without a concurrent forward CS when US onset was no longer predictable (i.e. US-CS instead of CS-US-CS).

The third empirical chapter of this thesis is called ‘Startle during backward evaluative conditioning is not modulated by instructions’. In this chapter groups are compared with and without an instructional manipulation focussing on the role of the CSs in controlling the USs. The sound USs used by Moran and Bar-Anan (2013) were employed with the addition of a neutral US. Explicit valence ratings and the startle blink reflex were measured. The aim was to determine whether an instructional manipulation focussing on the role of the CSs in controlling the USs would modulate the startle blink reflex in the same manner as explicit valence ratings.

The fourth and final empirical chapter is called ‘Assimilation or contrast effects in backward evaluative conditioning: The role of US offset predictability’. Two experiments look at whether US/CS overlap and US duration are capable of driving backward CS contrast effects in a picture-picture paradigm. Experiment 1 employed a backward conditioning only procedure (US-CS), while Experiment 2 used a concurrent forward and backward conditioning paradigm (CS-US-CS). Backward conditioning is also compared across experiments to observe the influence of US onset predictability (i.e. having a concurrent forward CS; CS-US-CS vs US-CS) on backward CS valence.

The aim is to determine the relative influence of instructions, affective relief, US intensity, US onset predictability, and US offset predictability on the emergence of backward CS contrast effects in evaluative conditioning. The general discussion will synthesise the findings of the current thesis with the existing literature to address the influence of each of these factors on backward evaluative conditioning in general.

1.9 References

Allport, G. W. (1935). Attitudes. In C. Murchison (Ed.), Handbook of social psychology (Vol. 2, pp. 798–844). Worcester, MA: Clark University Press.

Andreatta, M., Mühlberger, A., Glotzbach-Schoon, E., & Pauli, P. (2013). Pain predictability reverses valence ratings of a relief-associated stimulus. Frontiers in Systems Neuroscience, 7(53), 1-12, http://dx.doi.org/10.3389/fnsys.2013.00053

Andreatta, M., Mühlberger, A., Yarali, A., Gerber, B., & Pauli, P. (2010). A rift between implicit and explicit conditioned valence in human pain relief learning. *Proceedings of the Royal Society of London B: Biological Sciences.* http://dx.doi.org /10.1098/rspb.2010.0103

Bading, K., Stahl, C., & Rothermund, K. (2019). Why a standard IAT effect cannot provide evidence for association formation: The role of similarity construction. *Cognition and Emotion.* http://dx.doi.org/10.1080/02699931.2019.1604322.

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,* 133-142. https://dx.doi.org/10.1016/0005-7967(92)90136-5

Becerra, L., Navratilova, E., Porreca, F., & Borsook, D. (2013). Analogous responses in the nucleus accumbens and cingulate cortex to pain onset (aversion) and offset (relief) in rats and humans. *Journal of Neurophysiology, 110,* 1221-1226. https://dx.doi.org/10.1152/jn.00284.2013

Benedict, T., & Gast, A. (2020). Evaluative conditioning with fear- and disgust-evoking stimuli: No evidence that they increase learning without explicit memory. *Cognition and Emotion, 34,* 42-56. https://dx.doi.org/10.1080/02699931.2019.1646213

Blumenthal, T. D., Cuthbert, B. N., Filion, D. L., Hackley, S., Lipp, O. V., & Van Boxtel, A. (2005). Committee report: Guidelines for human startle eyeblink

electromyographic studies. *Psychophysiology, 42,* 1-14. http://dx.doi.org/10.1111/j.1469-8986.2005.00271.x

Brownstein, M., & Madva, A., & Gawronski, B. (2019). What do implicit measures measure? *Wiley Interdisciplinary Reviews: Cognitive Science*. http://dx.doi.org /10.10.1002/wcs.1501

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,* 1-21. http://dx.doi.org/10.1177/1088868320911325

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

Cuthbert, B. N., Bradley, M. M., & Lang, P. J. (1996). Probing picture perception: Activation and emotion. *Psychophysiology, 33,* 103-112. http://dx.doi.org/10.1111/j.1469-8986.1996.tb02114.x

Davey, G. C. I. (1994). Is evaluative conditioning a qualitatively distinct form a classical conditioning? *Behaviour Research and Therapy, 32.* 291-299. https://dx.doi.org/10.1016/0005-7967(94)90124-4

De Houwer, J. (2007). A conceptual and theoretical analysis of evaluative conditioning. *The Spanish Journal of Psychology, 10,* 230–241. http://dx.doi.org/10.1017/S1138741600006491

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin, 13*5, 347–368. http://dx.doi.org/10.1037/a0014211

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,* 853–869. http://dx.doi.org/10.1037/0033-2909.127.6.853

Deutsch, R., Smith, K. J. M., Kordts-Freudinger, R., & Reichardt, R. (2015). How absent negativity relates to affect and motivation: An integrative relief model. *Frontiers in Psychology, 6*(152), 1-23. https://dx.doi.org/10.3389/fpsyg.2015.00152

Dirikx, T., Hermans, D., Vansteenwegen, D., Baeyens, F., & Eelen, P. (2004). Reinstatement of extinguished conditioned responses and negative stimulus valence as a pathway to return of fear in humans. *Learning & Memory*, *11*(5), 549-554. http://dx.doi.org/lm.78004

Eder, A. B., Krishna, A., & Van Dessel, P. (2019). Operant evaluative conditioning. *Journal of Experimental Psychology: Animal, Learning, and Cognition, 45,* 102-110. https://dx.doi.org/10.1037/xan0000189

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,* 1013–1027. http://dx.doi.org/10.1037/0022-3514.69.6.1013

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

Field, A. P., & Davey, G. C. L. (1999). Reevaluating evaluative conditioning: A nonassociative explanation of conditioning effects in the visual evaluative conditioning paradigm. *Journal of Experimental Psychology: Animal Behavior Processes, 25,* 211–224. https://dx.doi.org/10.1037/0097-7403.25.2.211

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

Franklin, J. C., Hessel, E. T., Aaron, R. V., Arthur, M. S., Heilbron, N., Prinstein, M. J. (2010). The functions of nonsuicidal self-injury: Support for cognitive-affective regulation and opponent processes from a novel

psychophysiological paradigm. *Journal of Abnormal Psychology, 119,* 850-862. https://dx.doi.org/10.1037/a0020896

Franklin, J. C., Lee, K. M., Hanna, E. K., and Prinstein, M. J. (2013). Feeling worse to feel better: Pain-offset relief simultaneously stimulates positive affect and reduces negative affect. *Psychological Science, 24,* 521-529. https://dx.doi.org/10.1177/0956797612458805

Galdi, S., Arcuri, L., & Gawronski, B. (2008). Automatic mental associations predict future choices of undecided decision makers. *Science, 321*, 1100-1102. http://dx.doi.org/ 10.1126/science.ll60769

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

Gawronski, B., & Bodenhausen, G. V. (2011). The associative propositional evaluation model: Theory, evidence, and open questions. *Advances in Experimental Social Psychology, 44,* 59–127. http://dx.doi.org/10.1016/B978-0-12-385522-0.00002-0

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

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., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal stability of implicit and explicit measures: A longitudinal analysis. *Personality and Social Psychology Bulletin, 43,* 300-312. https://dx.doi.org/10.1177/0146167216684131

Gerber, B., Yarali, A., Diegelmann, S., Wotjak, C. T., Pauli, P., & Fendt, M. (2014). Pain- relief learning in flies, rats, and man: Basic research and applied perspectives. *Learning and Memory, 21*, 232-252. http://dx.doi.org/10.1101/lm.032995.113

Gibson, B. (2008). Can evaluative conditioning change attitudes toward mature brands? New evidence from the Implicit Association Test. *Journal of Consumer Research, 35*, 178-188. http://dx.doi.org/10.1086/527341

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*, 1464–1480. http://dx.doi.org/10.1037/0022-3514.74.6.1464

Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality & Social Psychology, 90*(1), 1-20. https://dx.doi.org/10.1037/0022-3514.90.1.1

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. Psychological Bulletin, 136, 390-421. http://dx.doi.org/10.1037/a0018916

Hughes, S., De Houwer, J., & Perugini, M. (2016). Expanding the boundaries of evaluative learning research: How intersecting regularities shape our likes and dislikes. *Journal of Experimental Psychology: General, 145,* 731-754. https://dx.doi.org/10.1037/xge0000100

Hu, X., Gawronski, B., & Balas, R. (2017a). 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,* 17–32. http://dx.doi.org/10.1177/0146167216673351

Hu, X., Gawronski, B., & Balas, R. (2017b). Propositional versus dual process accounts of evaluative conditioning: II. The effectiveness of counter-conditioning and counter-instructions in changing implicit and explicit

evaluations. *Social Psychological & Personality Science, 8,* 858–866. http://dx.doi.org/10.1177/1948550617691094

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,* 933-948. https://dx.doi.org/10.1037/a0014747

Kawakami, K., Phills, C. E., Steele, J. R., & Dovidio, J. F. (2007). (Close) distance makes the heart grow fonder: Improving implicit racial attitudes and interracial interactions through approach behaviors. *Journal of Personality & Social Psychology, 92,* 957-971. https://dx.doi.org/10.1037/0022-3514.92.6.957

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,* 61-68. http://dx.doi.org/10.1177/1948550615599237

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1990). Emotion, attention, and the startle reflex. *Psychological Review, 97,* 377-395. https://dx.doi.org/10.1037/0033-295X.97.3.377

LeBel, E. P., & Campbell, L. (2009). Implicit partner affect, relationship satisfaction, and the prediction of romantic breakup. *Journal of Experimental Social Psychology, 45,* 1291-1294. http://dx.doi.org/10.1016/j.jesp.2009.07.003

Leknes, S., Berna, C., Lee, M. C., Snyder, G. D., Biele, G., & Tracey, I. (2013). The importance of context: When relative relief renders pain pleasant. *Pain, 154,* 402-410. http://dx.doi.org/10.1016/j.pain.2012.11.018

Leknes, S., Brooks, J. C. W., Wiech, K., & Tracey, I. (2008). Pain relief as an opponent process: A psychophysical investigation. *European Journal of Neuroscience, 28,* 794-810. https://doi.org/10.1111/j.1460-9568.2008.06380.x

Leknes, S., Lee, M., Berna, C., Andersson, J., & Tracey, I. (2011). Relief as a reward: Hedonic and neural responses to safety from pain. *PLoS one, 6*(4), e17870. https://dx.doi.org/10.1371/journal.pone.0017870

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

Lipp, O. V., Neumann, D. L., & Mason, V. (2001). Stimulus competition in affective and relational learning, *Learning and Motivation, 32*, 306-331. https://dx.doi.org/10.1006/lmot.2001.1087

Luck. C. C., & Lipp, O. V. (2017). Startle modulation and explicit valence evaluations dissociate during backward fear conditioning. *Psychophysiology, 54*, 673-683. http://dx.doi.org/10.1111/psyp.12834

Mallan, K. M., Lipp, O. V., & Libera, M. (2008). Affect, attention, or anticipatory arousal? Human blink startle modulation in forward and backward affective conditioning. *International Journal of Psychophysiology, 69*. 9-17. http://dx.doi.org/10.1016/j.ijpsycho.2008.02.005

Martin, I., & Levey, A. B. (1978). Evaluative conditioning. *Advances in Behaviour Research and Therapy, 1*, 57-101. https://dx.doi.org/10.1016/0146-6402(78)90013-9

Martin, I., & Levey, A. B. (1994). The evaluative response: Primitive but necessary. *Behaviour Research and Therapy, 32*, 301–305. https://dx.doi.org/10.1016/0005-7967(94)90125-2

Matsuda, K., Garcia, Y., Catagnus, R., Brandt, J. A. (2020). Can behavior analysis help us understand and reduce racism? A review of the current literature. *Behavior Analysis in Practice, 13*, 336-347. http://dx.doi.org/10.1007/s40617-020-00411-4

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

Moran, T., and Bar-Anan, Y. (2013). The effect of object-valence relations on automatic evaluation. *Cognition and Emotion, 27,* 743-752. http://dx.doi.org/10.1080/02699931.2012.732040

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2016). The assimilative effect of co-occurrence on evaluation above and beyond the effect of relational qualifiers. *Social Cognition, 34,* 435-461. http://dx.doi.org/101521soco2016345435

Moreland, R. L., & Topolinski, S. (2010). The mere exposure phenomenon: A lingering melody by Robert Zajonc. *Emotion Review, 2,* 329-339. https://dx.doi.org/10.1177/1754073910375479

Porreca, F. & Navratilova, E. (2017). Reward, motivation, and emotion of pain and its relief. *Pain, 158,* S43-S49. https://dx.doi.org/10.1097/j.pain.0000000000000798

Selby, E. A., & Joiner, T. E. (2009). Cascades of emotion: The emergence of borderline personality disorder from emotional and behavioral dysregulation. *Review of General Psychology, 13,* 219-229. https://dx.doi.org/10.1037/a0015687

Seymour, B., O'Doherty, J. P., Koltzenburg, M., Wiech, K., Frackowiak, R., Friston, K., & Dolan, R. (2005). Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. *Nature Neuroscience, 8,* 1234-1240. https://dx.doi.org/10.1038/nn1527

Sheeran, P., Maki, A., Montanaro, E., Avishai-Yitshak, A., Bryan, A., Klein, W. M. P., … Rothman, A. J. (2016). The impact of changing attitudes, norms, and self-efficacy on health-related intentions and behaviour: A meta-analysis. *Health Psychology, 35,* 1178-1188. http://dx.doi.org/10.1037/hea0000387

Solomon, R. L. (1980). The opponent-process theory of acquired motivation: The costs of pleasure and the benefits of pain. *American Psychologist, 35,* 691-712. https://doi.org/10.1037/0003-066X.35.8.691

Staats, C. K., & Staats, A. W. (1957). Meaning established by classical conditioning. *Journal of Experimental Psychology, 54,* 74–80. http://dx.doi.org/10.1037/h0047716

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behaviour. *Personality and Social Psychology Review, 8,* 220-247. https://dx.doi.org/10.1207/s15327957pspr0803_1

Unkelbach, C., Fiedler, K. (2016). Contrast CS-US relations reverse evaluative conditioning effects. *Social Cognition, 34*. 413-434. http://dx.doi.org/10.1521/soco.2016.34.5.413

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

Zanna, M. P., Kiesler, C. A., & Pilkonis, P. A. (1970). Positive and negative attitudinal affect established by classical conditioning. *Journal of Personality and Social Psychology, 14,* 321-328. https://dx.doi.org/10.1037/h0028991

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology Monographs, 9*, 1–27.

Zbozinek, T. D., Hermans, D., Prenoveau, J. M., Liao, B., & Craske, M. G. (2015). Post-extinction conditional stimulus valence predicts reinstatement fear: Relevance for long-term outcomes of exposure therapy. *Cognition and Emotion, 29*(4), 654-667. http://dx.doi.org/10.1080/02699931.2014.930421

Chapter 2: Contrast Effects in Backward Evaluative Conditioning: Exploring Effects of Affective Relief/Disappointment versus Instructional Information[1]

Luke J. S. Green
Curtin University

Camilla C. Luck
Curtin University

Bertram Gawronski
University of Texas at Austin

Ottmar V. Lipp
Curtin University

Author Notes

This work was supported by an Australian Government Research Training Program Scholarship to Luke Green and grants DP180111869 and SR120300015 from the Australian Research Council to Ottmar Lipp.

Correspondence concerning this article should be sent to: Luke J S Green, School of Psychology, Curtin University, GPO Box U1987 Perth WA 6845, Australia. Email: luke.green2@postgrad.curtin.edu.au.

2.1 Abstract

Past studies of backward evaluative conditioning (EC) have found an assimilation effect, in that neutral conditional stimuli (CS) were found to acquire the valence of co-occurring unconditional stimuli (US). Recent studies employing a concurrent forward and backward conditioning paradigm with instructions suggesting a contrastive relation between the US and the backward CS have resulted in contrast effects, in that backward CSs acquired valence opposite to the US. The current research investigated whether these effects were in fact due to the instructions highlighting the contrastive relation between the US and CS, or whether affective relief/disappointment experienced at US offset could account for this result. Consistent with the hypothesized role of instructions, backward CS contrast effects occurred only when instructions highlighted the valence of the US and attributed control of that US to the CSs. In contrast to the affective relief/disappointment hypothesis, no backward CS contrast effects were found without such instructions.

Keywords: associative learning, backward conditioning, evaluative conditioning, attitudes, propositional learning

2.2 Introduction

How we evaluate people, events, and stimuli has been shown to influence interpersonal relationships, voting behaviour, consumer behaviour, and career aspirations (e.g., Galdi, Arcuri, & Gawronski, 2008; Gibson, 2008; LeBel & Campbell, 2009). One method through which these evaluations are acquired and changed is known as *evaluative conditioning* (EC), which occurs when the evaluation of a neutral conditional stimulus (CS) is changed by its co-occurrence with a valenced unconditional stimulus (US; De Houwer, 2007). Prominent examples of EC include advertising campaigns that present a consumer product (CS) with a well-liked celebrity (US), leading to the product becoming positive. EC is of great interest to psychologists, because encountering co-occurring stimuli of differing valence is un-avoidable and ever-present in our daily lives.

Although past EC research has predominantly found assimilation effects (i.e., CS-US pairings produce CS evaluations that are in line with the valence of the co-occurring US), some studies have found contrast effects under certain conditions (i.e., CS-US pairings produce CS evaluations that are opposite to the valence of the co-occurring US). The main goal of the current research was to investigate the contribution of relief/disappointment learning and instructions about CS-US relations to the emergence of contrast effects in EC. We were particularly interested in whether relief/disappointment experienced at the offset of valenced USs could account for previously obtained contrast effects that have been interpreted to be the result of instructions about CS-US relations (Moran & Bar-Anan, 2013). If so, this would provide further support for the notion that EC and fear conditioning may share a common underlying mechanism.

2.2.1 Evaluative Conditioning

In a typical EC procedure, participants are presented with neutral stimuli (e.g., images of geometric shapes; CSs). Some of the neutral stimuli are presented together with pleasant stimuli (e.g., images of puppies; USpos), while others are presented together with unpleasant stimuli (e.g., images of snakes; USneg). After repeated pairings, evaluations of the neutral stimuli paired with pleasant stimuli tend to become more positive, whereas evaluations of the neutral stimuli paired with unpleasant stimuli tend to become more negative (De Houwer, Thomas, & Baeyens, 2001). This effect is quite robust, in that it has been shown with stimuli from various

modalities (i.e., visual, auditory, olfactory), when both the CSs and USs are from the same or different modalities, and when the CSs are presented with the same or different USs across trials (for a meta-analysis, see Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010).

EC-related changes in CS valence can be measured using either explicit or implicit measures. Explicit measures of CS valence involve asking participants to rate how much they like the CSs, or how pleasant they find the CSs (self-reported valence ratings). Implicit measures of CS valence, such as the implicit association test (IAT; Greenwald, McGhee, & Schwartz, 1998) or affective priming (Fazio, Jackson, Dunton, & Williams, 1995), are performance-based measures that infer CS evaluations from the speed of categorisation responses on different kinds of trials (De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009; Gawronski & De Houwer, 2014). Implicit measures exploit the fact that categorisation tends to be faster on valence-congruent trials than valence-incongruent trials. In affective priming, for example, responses to pleasant target words tend to be faster when they are preceded by a positive prime stimulus than when they are preceded by a negative prime stimulus (Fazio et al., 1995). In research on EC, implicit and explicit measures typically reveal similar patterns of results (e.g., Mallan, Lipp, & Libera, 2008; Olson & Fazio, 2001). However, as we will discuss below, dissociations between explicit and implicit measures have been found to emerge under certain circumstances (e.g., Hu, Gawronski, & Balas, 2017a, 2017b; Moran & Bar-Anan, 2013).

2.2.2 Forward vs. Backward Conditioning

CS-US pairings can differ in terms of the sequence in which a CS co-occurs with a US. *Forward conditioning* involves cases in which a CS precedes a US (CS-US); *backward conditioning* involves cases in which a CS follows a US (US-CS); and *simultaneous conditioning* involves cases in which a US appears simultaneously with a CS (CS+US). Mallan et al. (2008) used a between-subjects design to compare EC effects in forward, backward, and simultaneous conditioning paradigms, using geometric shapes as CSs, and valenced pictures as USs. Participants were instructed to pay attention to the pictures, as they would be asked questions about them at the end of the experiment. After conditioning, explicit valence ratings and affective priming revealed similar EC effects in all groups, such that CSs paired with positive USs became more positive than CSs paired with negative USs. These results suggest that both forward and backward conditioning lead to assimilation effects, such that

the CS acquires the valence of the US it was paired with (see also Kim, Sweldens, & Hütter, 2016). This assimilation effect was evident on both explicit and implicit measures (see also Hofmann et al., 2010).

However, different from Mallan et al.'s findings (2008), Moran and Bar-Anan (2013) found a dissociation between explicit and implicit measures for backward conditioning in a study that compared forward and backward conditioning using a within-subjects design (CS-US-CS). On positive trials of the EC task, participants were presented with an image of one member of a family of alien creatures (forward CSpos), followed by a pleasant melody (USpos), and an image of a member of a second family of alien creatures (backward CSpos). On negative trials of the EC task, participants were presented with an image of a member of a third family of alien creatures (forward CSneg), followed by an aversive human scream (USneg), and an image of a member of a fourth family of alien creatures (backward CSneg). Before the EC task, participants were told that each alien family (CSs) had a different role to play: start the melody, stop the melody, start the human scream, or stop the human scream. Participants were instructed to learn the role of each family of aliens for a memory test at the end of the study. For forward CSs, Moran and Bar-Anan (2013) found assimilation effects on both explicit and implicit measures of CS valence. That is, CSs that preceded the pleasant melody were found to be more positive than CSs that preceded the unpleasant scream. For backward CSs, however, explicit and implicit measures revealed different results. Whereas implicit measures of backward CS valence showed assimilation effects (i.e., CSs presented after the pleasant melody elicited more favourable responses than CSs presented after the unpleasant scream), explicit measures of backward CS valence revealed contrast effects (i.e., CSs presented after the unpleasant scream were rated more positively than the CSs presented after the pleasant melody).

Moran and Bar-Anan (2013) interpreted the obtained dissociation in terms of two functionally distinct learning processes underlying evaluations on implicit and explicit measures. Drawing on the associative-propositional evaluation (APE) model (Gawronski & Bodenhausen, 2006, 2011, 2018), they suggested that implicit measures are more sensitive to effects of associative learning, reflecting the mere co-occurrence of a CS and a US regardless of their relation. In contrast, explicit measures were assumed to be more sensitive to effects of propositional learning, reflecting the particular relation between a CS and a co-occurring US. However,

different from this interpretation in terms of two functionally distinct learning mechanisms, recent research suggests that the observed backward CS assimilation effect on implicit measures might be due to factors during the measurement of CS evaluations, in that backward CS contrast effects occur on both implicit and explicit measures when these factors are controlled (see Bading, Stahl, & Rothermund, 2019; Hu et al., 2017a; Moran & Bar-Anan, 2019; for a review, see Corneille & Stahl, 2019). These findings shift the explanatory burden from the reported dissociation between implicit and explicit measures to the question of what causes the backward CS contrast effects observed by Moran and Bar-Anan (2013).

According to Moran and Bar-Anan (2013), the backward CS contrast effect observed in their study is the direct result of the instructional manipulation, which stated that the backward CSs would stop the preceding USs. An alternative mechanism that may account for the backward CS contrast effects reported by Moran and Bar-Anan (2013) without the need for an instructional manipulation is affective relief/disappointment. Research on relief learning has shown that presenting a CS at the offset of an aversive stimulus (US-CS) can result in this backward CS gaining positive valence (Andreatta, Mühlberger, Yarali, Gerber, & Pauli, 2010; Gerber et al., 2014; Luck & Lipp, 2017). This occurs because feelings of relief from the aversive stimulus ending become conditioned to the backward CS, which is presented simultaneously with feelings of relief. Preliminary research suggests that the same contrastive process also occurs at the offset of a positive stimulus, resulting in disappointment (Green, Luck, & Lipp, submitted). Although Andreatta et al. (2010) investigated contrast effects in fear conditioning rather than evaluative conditioning, these findings raise the possibility that affective relief/disappointment may also affect the direction of backward EC effects. If the negative/positive US used by Moran and Bar-Anan (2013) was sufficiently aversive/pleasant to drive relief/disappointment learning, affective relief/disappointment may be sufficient to explain the backward CS contrast effect in their study. Although the boundary conditions of relief and disappointment learning are still not well understood, it is conceivable that affective relief/disappointment at the offset of the aversive/pleasant US in Moran and Bar-Anan (2013) may have contributed to the observed backward CS contrast effects.

A second alternative explanation is that Moran and Bar-Anan (2013) assessed forward and backward conditioning concurrently within subjects (CS-US-CS),

whereas Mallan et al. (2008) assessed forward and backward conditioning between groups (CS-US vs. US-CS). Research by Andreatta, Mühlberger, Glotzbach-Schoon, and Pauli (2013) suggests that the presence of a forward CS may moderate backward conditioning effects. These researchers found that making an aversive electro-tactile US predictable by presenting a forward CS resulted in positive explicit valence ratings of a backward CS paired with this US, but a backward CS was rated negatively when no forward CS was presented (see also Andreatta & Pauli, 2017). This occurs because the forward CS becomes more aversive when the onset of the US is predictable. The result is a larger discrepancy between the conditioned valence of the forward and backward CSs, possibly making the backward CS appear to be the opposite valence of the forward CS. This finding was replicated by Green et al. (submitted). Using both positive and negative USs and the same stimuli and instructions as Moran and Bar-Anan (2013), Green et al. found assimilation effects for backward CSs when no forward CS was present. Thus, counter to Moran and Bar-Anan's (2013) argument that the backward CS contrast effect in their study is the result of instructions about contrastive CS-US relations, the concurrent forward and backward conditioning procedure may be conducive to backward CS contrast effects without any instructions about contrastive CS-US relations. In either of the aforementioned cases, contrast effects should be observed using a similar paradigm without the instructions employed by Moran and Bar-Anan (2013).

Based on these considerations, the main goal of the current research was to investigate whether presenting within-subjects forward and backward conditioning without instructions could result in backward CS contrast effects as a result of the combination between US predictability and affective relief/disappointment.[2]

2.3 Pilot Studies

In addition to the main experiment reported below, we conducted two pilot studies to determine the instructions and experimental parameters required to address whether within-subjects forward and backward conditioning and affective relief/disappointment alone would elicit backward CS contrast effects (for more details, see Supplementary Materials). Pilot Study 1 compared the instructions used by Moran and Bar-Anan (2013) and Mallan et al. (2008) in a within-subjects forward

[2] All materials, data, and analysis files are available at https://osf.io/ur5kd/.

and backward conditioning picture-picture paradigm using stimulus presentation and timing parameters adapted from Mallan et al. (2008). No evidence of backward CS contrast effects on explicit measures were found in either group. Thus, we performed a direct replication of Moran, Bar-Anan, and Nosek (2016) to ensure that backward CS contrast effects could be obtained in a picture-picture paradigm using instructions. The instructions used in this second pilot study differed from the ones in the first pilot study, in that they highlighted the agency of the CSs in controlling the US (i.e., the CSs control which event happens to you, either gold bars or garbage) and the valence of the outcome (i.e., whether this event is happy [gold bars] or sad [garbage]). We also investigated whether the lack of backward CS contrast effects observed in Pilot Study 1 was the result of presenting explicit valence ratings and affective priming before the learning phase. It is possible that evaluating stimulus valence before the learning phase puts participants in an 'evaluative mindset', thus resulting in amplified or erroneous effects that could strengthen assimilative EC, and thereby conceal potential contrast effects (Gast & Rothermund, 2011). Pilot Study 2 revealed that backward CS contrast effects could in fact be obtained in a picture-picture paradigm using these instructions, regardless of whether explicit valence ratings and affective priming were presented before the learning phase. These results suggest that an 'evaluative mindset' was not responsible for the lack of backward CS contrast effects observed in Pilot Study 1.

2.4 Main Experiment

The primary aim of the main experiment was to determine whether affective relief/disappointment and US predictability without instructions would be sufficient to elicit backward CS contrast effects. A secondary aim was to test whether the 'valence-agency' instructions adopted from Moran et al. (2016) in Pilot Study 2 drive backward CS contrast effects when compared with the 'observe instructions' used in Pilot Study 1. If this were the case, it would suggest that an emphasis on 'valence' and 'agency' is essential for backward CS contrast effects in the picture-picture paradigm of Pilot Study 2. Another secondary aim was to determine whether the lack of backward CS contrast effects in both groups in Pilot Study 1 was due to features of the 'Mallan paradigm' not supporting backward CS contrast learning, regardless of instructions. It is possible that overlap between the US and the backward CS may

assist in highlighting the fact that the backward CS controls the offset of the US, as there is generally overlap between stimuli when one of them is responsible for stopping an event (i.e. a good Samaritan intervening to bring resolution to an altercation between two parties). Without this, backward CS contrast learning may not be possible. In addition to this, the variability of the US may increase the affective relief/disappointment experienced at the offset of the US, because varying the US duration makes it difficult to predict when the US is going to end. Thus, without US offset being unpredictable, backward CS contrast learning may be less likely.

2.4.1 Method

Participants and design. Participants were recruited through M-Turk using TurkPrime after the Curtin University Human Research Ethics Committee approved this research protocol (Litman, Robinson, & Abberbock, 2017). The sample comprised 194 participants after duplicates and those failing to complete the experiment were removed (n = 33). The sample size was based on previous research to detect the within-subjects interaction of interest for each group (Moran & Bar-Anan, 2013; Moran et al., 2016). Moran and Bar-Anan (2013) and Moran et al. (2016) had sample sizes ranging from 32 to 68 participants. In these studies, the within-subjects interaction of interest yielded large effects sizes between η_p^2 = .15 and η_p^2 = .60. Based on these effect sizes, we anticipated that approximately 50 participants per group would provide sufficient power to detect the effects of interest. The 'observe instructions, Mallan paradigm' group consisted of 49 participants (27 female), the 'observe instructions, Moran paradigm' group comprised 48 (24 female), the 'valence-agency instructions, Mallan paradigm' group comprised 50 (28 female), and the 'valence-agency instructions, Moran paradigm' group comprised 47 participants (21 female). Five participants failed to provide demographic information. The mean age of the 189 participants who provided demographic information was 36.35, SD = 11.085. Groups did not differ on gender, χ^2 (3) = 2.220, p = .528, ethnicity, χ^2 (15) = 13.029, p = .600, or age, F (1, 185) = .012, p = .913, η_p^2 = .000, BF_{incl} = 0.23.

Explicit valence ratings. In the 'Mallan paradigm' groups, each CS was presented one-by-one and participants were asked to rate how pleasant they found the stimulus on a 9-point scale ranging from 1 (*unpleasant*) to 9 (*pleasant*). In the 'Moran paradigm' groups, each CS family was presented alone and participants were

asked “Based on your very first emotional response, how much do you like the creatures in the picture? Click the appropriate answer below: dislike strongly, dislike moderately, dislike slightly, like slightly, like moderately, like strongly”.

Affective priming task. In the ‘Mallan paradigm’ groups, each of the four CSs were presented once with 10 positive target words and 10 negative target words for a total of 80 trials. In the ‘Moran paradigm’ groups, two creatures from each family were presented with positive and negative words twice, and two creatures from each family were presented with positive and negative words three times, for a total of 10 positive and 10 negative word pairings per family. This resulted in 80 trials. For both groups, a fixation cross was presented for 500ms, followed by the CS prime for 200ms, and then the target word until the participant provided their response. Participants were instructed to press the *I* key if the target word was positive and the *E* key if the target word was negative. Target words were taken from Hu et al. (2017a, 2017b). The positive words were *pleasant, good, outstanding, beautiful, magnificent, marvellous, excellent, appealing, delightful,* and *nice*. The negative words were *unpleasant, bad, horrible, miserable, hideous, dreadful, painful, repulsive, awful,* and *ugly.*

Recollective Memory Test. For exploratory purposes, the current study also included measures of recollective memory. In the ‘Mallan paradigm’ groups, participants were shown each CS and asked: “Circle the appropriate answer below. Was this picture presented: Together with pleasant pictures, together with unpleasant pictures, together with pleasant and unpleasant pictures, I did not see this picture, I could not tell?” In the ‘Moran paradigm’ group, participants were shown each CS and asked: “Circle the appropriate answer below. What is the role of this creature: To start pleasant pictures, to stop pleasant pictures, to start unpleasant pictures, to stop unpleasant pictures?” Using the sum of correct responses on the memory test, accuracy scores on the test could range from zero to four. Both groups were also presented with each US and each CS, and asked to indicate which CS came before or after each US. This procedure resulted in an accuracy score ranging from 0 to 16. Participants were classified as remembering the CS-US contingencies only if they scored 100% on both memory tests. In the ‘Moran paradigm’ groups, each CS family was presented alone and participants were asked “In the game, what was the role of the creatures in the picture? Click the appropriate answer below: Starting gold, starting garbage, stopping gold, stopping garbage?” The analysis of the recollective

memory data did not add substantially to the current report, and is available in the Supplementary Materials.

Demographics questionnaire. Participants were asked to report their age, gender, and ethnicity, and to provide information about the environment in which they completed the task, and if they had any comments.

Apparatus/stimuli. In the 'Mallan paradigm' groups, four images of aliens, one from each of the four families of alien creatures created by Moran and Bar-Anan (2013), were used as CSs (see below; materials from Moran and Bar-Anan, 2013, available at https://osf.io/cqsnj/). Each alien differed in colour and head shape. Four positive and four negative pictures from the International Affective Picture System (IAPS; CSEA, 1999) were used as USs (1050, 1300, 1440, 1710, 5833, 6313, 6560, and 8190). In the 'Moran paradigm' groups, CSs and USs were those used by Moran et al. (2016; available at https://osf.io/v2trw/). CSs were four families of alien creatures, with each family comprising four creatures for a total of 16 CSs. The positive US was a picture of puppies, gold bars, and a baby, presented next to each other as a single image, and the negative US was a picture of an aggressive dog, garbage, and a crying child presented next to each other as a single image. Inquisit 4 Web by Millisecond Software ™ (2016) was used to run the experiment and to record responses in all tasks.

Procedure. Participants selected the HIT (human intelligence task) on M-Turk and read the description of the study. When participants began the study, they were presented with an information sheet outlining the tasks, informed that they could withdraw at any time by pressing 'ctrl + q', and then prompted to press 'continue' if they consented to participate. Informed consent was implied if participants pressed 'continue'. In all groups, the first explicit valence ratings and affective priming task was presented followed by the training phase. In the 'Mallan paradigm' groups, the training phase comprised 12 positive and 12 negative trials presented pseudo-randomly, with inter-trial intervals of 4, 6, and 8 seconds. Each trial consisted of a forward CS, followed by a positive or negative US, followed by a backward CS. This CS-US-CS paradigm was adapted from Moran and Bar-Anan (2013), with some modifications based on Mallan et al. (2008). We used one CS from each of the four alien families, four positive and four negative pictures as USs, and each stimulus was presented for 4 seconds with onset and offsets coinciding (i.e., no overlap between CSs and USs). CSs were counter-balanced using a Latin square

resulting in four CS orders, with each CS occurring in each role equally. In the 'Moran paradigm' groups, the training phase comprised 12 positive and 12 negative trials randomly presented with inter-trial intervals of 2 seconds. Each trial consisted of a forward CS, followed by a positive or negative US, followed by a backward CS. This CS-US-CS paradigm was an exact replication of Moran et al. (2016). CSs were presented for 1.5 seconds and USs were presented in blocks of 1s flashes with a 200ms break between each flash for a total of 3 or 5 seconds of total US presentation time. Onset of the US coincided with offset of the forward CS, and onset of the backward CS occurred 200ms after the last US appearance. One group in each of the 'Mallan paradigm' and 'Moran paradigm' groups received the *valence-agency* instructions and one group received the *observe* instructions.

In the *valence-agency* instructions groups, participants received the following instructions from Moran et al. (2016) before the training phase:

> *In the next game, you will get piles of shiny gold bars, but also some stinky garbage piles. Getting gold bars is a happy event, whereas getting garbage piles is a sad event. In the game, four families of creatures control whether happy or sad events happen to you. These are the four families. One family of creatures will always start the gold bars coming your way. A second family of creatures will always stop the gold bars. A third family of creatures will always start garbage piles coming your way. A fourth family of creatures will always stop the garbage piles. Your goal in this game is to learn which family of creatures starts the gold, which family stops the gold, which family starts the garbage, and which family stops the garbage. We will test your learning later in the game, so please pay close attention. If you read and understood the instructions, hit the spare bar to continue. Please pay close attention to the images on the screen. Make sure you learn and remember which family does each of the four actions (start gold, stop gold, start garbage, stop garbage). Press space to start the game.*

After 12 trials, the following instructions were presented:

> *Do you know by now which family starts the gold, which family stops the gold, which family starts the garbage, and which family stops the garbage? Try to memorize what each family does for a later test. Press space for a few more rounds to help you remember the roles of the families better.*

In the *observe* instructions group, participants received the following instructions adapted from Mallan et al. (2008):

> *In this task you will be presented with a series of pictures. Please pay attention to which pictures follow each other as you will be tested on this at the end of the experiment.*

After the training phase, the second explicit valence ratings and affective priming task was presented, followed by the memory test and demographics questionnaire. Participants then received a completion code to receive their compensation, and were thanked for their participation. The experiment took approximately 20 minutes on average to complete, and participants were compensated US-$5.70.

Statistical analyses. Frequentist analyses were performed using IBM SPSS Statistics 25. We also report the results Bayesian analyses conducted in JASP 0.10.0.0 to supplement the frequentist analyses. BF_{10} values from the model comparison are reported for main effects, and $BF_{inclusion}$ (BF_{incl}) values from the effects analysis (across matched models) are reported for interactions. The $BF_{inclusion}$ (across matched models) compares models that contain the effect of interest with equivalent models that have had the effect of interest removed. The result is a model that provides only the effect of the interaction of interest without contributions from lower order effects (known as the Baws approach; see Mathôt, 2017, for a discussion).

The explicit valence ratings in the 'Moran paradigm' groups were transformed from a 6-point scale to a 9-point scale ([X – 1] * 1.6 + 1), so that ratings could be compared with the 'Mallan paradigm' groups. EC scores were calculated as the difference between ratings of CSs paired with positive USs and ratings of CSs paired with negative USs. EC scores were calculated separately for forward vs. backward conditioning and for pre-training vs. post-training. Positive EC scores represent an assimilation effect and negative EC scores represent a contrast effect. In the affective priming task, trials on which target words were categorised incorrectly were scored as error trials. Trials on which reaction times were shorter than 300ms and longer than 1000ms were categorised as outliers, as they were deemed to be outside the window of a valid response (see Koppehele-Gossel, Hoffmann, Banse, & Gawronski, in press). Participants with a percentage of invalid trials greater than 25% on the affective priming task were excluded from the priming analyses ('observe instructions, Mallan paradigm', n = 12, 'observe instructions, Moran

paradigm', n = 11, 'valence-agency instructions, Mallan paradigm', n = 5, 'valence-agency instructions, Moran paradigm', n = 7). In the final sample at pre-test, 7.47% of trials were incorrect categorisations of target words and 6.85% of trials were outliers. At post-test, 7.82% of trials were incorrect categorisations of target words and 8.60% of trials were outliers. For the 'Moran paradigm' groups, responses following CSs within the same family in the affective priming task were averaged to provide overall means for each family. Priming scores of EC effects were calculated as the difference in response times between incongruent and congruent trials: (CSs paired with positive USs/negative target words + CSs paired with negative USs/positive target words) – (CSs paired with positive USs/positive target words + CSs paired with negative USs/negative target words). Priming scores were calculated separately for forward vs. backward conditioning and for pre-training and post-training. Positive priming scores suggest an assimilation effect, while negative scores suggest a contrast effect. EC scores from explicit valence ratings and affective priming scores were subjected to separate Frequentist and Bayesian 2 (Instructions: observe vs. valence-agency instructions; between-participants) × 2 (Paradigm: Mallan paradigm vs. Moran paradigm; between-participants) × 2 (Conditioning Type: forward vs. backward; within-participants) × 2 (Time: pre-test vs. post-test; within-participants) mixed ANOVAs. Significant interactions from the Frequentist analyses were followed-up with pairwise comparisons and one sample *t*-tests where appropriate. Pillai's trace values of the multivariate solution are reported for main effects and interactions ($\alpha = .05$). These analyses were also performed as two-sided paired and independent samples Bayesian *t*-tests using the default settings in the BayesFactor package in R. BF_{10} values are reported for all Bayesian follow-up analyses. The reliability of the priming task was $\alpha = .11$ at pre-test and $\alpha = .21$ at post-test. The analyses of the error data from the affective priming task did not add substantially to the current report and are available in the Supplementary Materials.

2.4.2 Results

Explicit valence ratings. Figure 1 shows mean EC scores based on explicit valence ratings for forward and backward conditioning measured pre-training and post-training as a function of Instructions and Paradigm. The figure suggests assimilation effects for forward conditioning at post-training for all groups, and contrast effects for backward conditioning at post-training for the valence-agency instructions group only, regardless of paradigm. The ANOVA revealed significant

main effects of Conditioning Type, F (1, 190) = 119.82, p < .001, η_p^2 = .387, BF_{10} = 3.30 × 10^{17}, and Time, F (1, 190) = 93.24, p < .001, η_p^2 = .329, BF_{incl} = 1.29 × 10^{10}, a significant two-way interaction between Instructions and Conditioning Type, F (1, 190) = 60.53, p < .001, η_p^2 = .242, BF_{incl} = 3.28 × 10^{12}, a significant two-way interaction between Conditioning Type and Time, F (1, 190) = 186.36, p < .001, η_p^2 = .495, BF_{incl} = 2.67 × 10^{33}, and a significant two-way interaction between Paradigm and Time, F (1, 190) = 6.41, p = .012, η_p^2 = .033, BF_{incl} = 2.15, which were qualified by a significant three-way interaction between Instructions, Conditioning Type, and Time, F (1, 190) = 65.84, p < .001, η_p^2 = .257, BF_{incl} = 1.41 × 10^{13}. The four-way interaction between Instructions, Paradigm, Conditioning Type, and Time was not significant, F (1, 190) = 0.10, p = .749, η_p^2 = .001, BF_{incl} = 0.12. Decomposing the three-way interaction, follow-up analyses revealed that, for forward conditioning, EC scores for valence-agency instructions were significantly larger than EC scores for observe instructions at post-training, F (1, 190) = 53.94, p < .001, η_p^2 = .221, BF_{10} = 9.43 × 10^{8}, but not pre-training, F (1, 190) = 0.01, p = .924, η_p^2 < .001, BF_{10} = 0.16. In contrast, for backward conditioning, EC scores for valence-agency instructions were significantly smaller than EC scores for observe instructions at post-training, F (1, 190) = 41.51, p < .001, η_p^2 = .179, BF_{10} = 11.71 × 10^{6}, but not pre-training, F (1, 190) = 0.02, p = .879, η_p^2 < .001, BF_{10} = 0.16. One-sample t-tests further indicated that post-training EC scores in the observe instruction groups were significantly larger than zero for forward conditioning, t (96) = 6.96, p < .001, d = 0.71, BF_{10} = 22.96 × 10^{6}, and backward conditioning, t (96) = 2.66, p = .009, d = 0.27, BF_{10} = 3.13. In contrast, post-training EC scores in the valence-agency instruction groups were larger than zero for forward conditioning, t (96) = 19.23, p < .001, d = 1.95, BF_{10} = 2.47 × 10^{31}, and significantly smaller than zero for backward conditioning, t (96) = 6.11, p < .001, d = 0.62, BF_{10} = 53.77 × 10^{4}. EC scores for forward and backward conditioning did not significantly differ from zero at pre-training for any of the four groups, all ts < 1.30, all ps > .196, all ds < 0.13, BF_{10}s < 0.25. Decomposing the significant two-way interaction between paradigm and time, follow-up analyses showed that EC scores for the Mallan paradigm tended to be larger than EC scores for the Moran paradigm at post-training, F (1, 190) = 3.32, p = .070, η_p^2 = .017, BF_{10} = 0.23, but not pre-training, F (1, 190) = 2.69, p = .103, η_p^2 = .014, BF_{10} = 0.57.

Affective priming. Figure 2 shows mean EC scores based on affective priming for forward and backward conditioning measured pre-training and post-training as a function of instructions and paradigm. The figure suggests an assimilation effect at post-training in the valence-agency instructions/Mallan paradigm group and a baseline priming score larger than zero in the observe instructions/Moran paradigm group. A marginal main effect of Time, F (1, 153) = 3.80, p = .053, η_p^2 = .024, BF_{10} = 1.11, was qualified by a marginal interaction between Instructions and Time, F (1, 153) = 3.82, p = .053, η_p^2 = .024, BF_{incl} = 0.92. The four-way interaction between Instructions, Paradigm, Conditioning Type, and Time was not significant, F (1, 153) = 0.20, p = .888, η_p^2 < .001, BF_{incl} = 0.22. Follow-up analyses revealed that priming scores in the valence-agency instructions group were significantly larger at post-training compared to pre-training, F (1, 153) = 8.41, p = .004, η_p^2 = .052, BF_{10} = 5.68. There was no significant difference between pre-training and post-training in the observe instructions group, F (1, 153) < 0.01, p = .998, η_p^2 < .001, BF_{10} = 0.09. One sample t-tests revealed that priming scores in the valence-agency instructions group were significantly larger than zero at post-training, t (85) = 3.15, p = .002, d = 0.34, BF_{10} = 13.95, but not pre-training, t (85) = 0.48, p = .633, d = 0.05, BF_{10} = 0.10. Moreover, priming scores in the observe instructions group were significantly larger than zero at pre-training, t (70) = 2.06, p = .044, d = 0.24, BF_{10} = 0.94, and marginally larger at post-training, t (70) = 1.73, p = .088, d = 0.21, BF_{10} = 0.49.

2.4.3 Discussion

The primary aim of this experiment was to determine whether affective relief/disappointment at the offset of an aversive/pleasant US when that US was predictable (CS-US-CS) could elicit backward CS contrast effects without the need for an instructional manipulation. Our secondary aim was to assess whether the absence of 'valence' and 'agency' components of the instructional manipulation could be responsible for the lack of backward CS contrast effects in our Pilot Study 1. Finally, we investigated whether the offset of the US was required to be unpredictable in order for backward CS contrast effects to emerge (assessed by comparing paradigms with and without US offset predictability).

For forward conditioning, explicit valence ratings revealed assimilation effects regardless of instructions. In contrast, for backward conditioning, explicit valence ratings showed assimilation effects for 'observe instructions' and contrast

effects for ‘valence-agency instructions.’ Moderate evidence in support of the null hypothesis for the four-way interaction including ‘paradigm’ supports the conclusion that this pattern emerged in both the ‘Moran paradigm’ and ‘Mallan paradigm’ groups. Unexpected baseline differences in measures from the affective priming task make their interpretation difficult, although it seems that assimilation effects occurred regardless of conditioning type and more strongly in the ‘valence-agency instructions’ group. However, this conclusion should be regarded with caution, especially considering the Bayes factor for the interaction between instructions and time was inconclusive.

Overall, these findings clearly demonstrate that affective relief/disappointment at the offset of an aversive/pleasant stimulus when the US is predictable is not sufficient to elicit backward CS contrast effects without an instructional manipulation. Moreover, these findings show that backward CS contrast effects using picture USs occur when the instructions emphasize ‘valence’ and ‘agency’ (Moran et al., 2016), suggesting that the different results in Pilot Studies 1 and 2 are driven by differences in instructions. Finally, these findings demonstrate that backward CS contrast effects can occur regardless of whether there is overlap between CSs and USs, or variability in US duration. This suggests that unpredictable US offset and presenting stimuli in a manner that appears as if the CSs are controlling US onset and offset is not necessary to observe backward CS contrast effects in presence of contrastive instructive instructions that emphasize valence and agency.

As a caveat to these conclusions, we would like to note that post-hoc power analyses revealed that our main experiment was underpowered to detect potential higher-order interactions. Explicit valence ratings showed backward CS contrast effects for the valence-agency instruction groups and assimilation effects for the observe instruction groups. Greater statistical power may have permitted the detection of a significant four-way interaction involving the factor ‘paradigm’, given that the difference between forward and backward conditioning under observe instructions at post-test was somewhat smaller in the Moran paradigm group compared with the Mallan paradigm group. However, this difference would have been driven by smaller forward conditioning rather than a difference in backward conditioning in the observe instructions groups. Because one of our aims was to determine whether US offset needed to be predictable in the valence-agency

instructions groups for backward CS contrast effects to emerge and additional power seems unlikely to change the pattern of backward conditioning effects observed here, our conclusions appear valid despite the fact that the experiment was underpowered for the detection of a significant four-way interaction. Moreover, Bayesian analysis of the four-way interaction provided moderate support for the null hypothesis, suggesting that even with higher power this interaction would not be meaningful.

Low statistical power may have also contributed to the affective priming task yielding less reliable results than expected. Cronbach's α ranged between .11 and .21, which is lower than expected for an affective priming task using this outlier treatment (Koppehele-Gossel et al., in press). The large confidence intervals of the mean suggest a lack of sensitivity, consistent with concerns that affective priming is more susceptible to measurement error than other implicit measures (Gawronski & De Houwer, 2014). In retrospect, the task may have benefited from more trials, especially in the 'Moran paradigm' conditions that involved the presentation of multiple exemplars of each CS family during conditioning. However, because affective priming scores largely followed the results of Moran and Bar-Anan (2013) and Hu et al. (2017a), it seems unlikely that our conclusions would have been different if stronger priming results had been observed. Nevertheless, future research with larger samples and greater trial numbers in the affective priming task may help to corroborate our conclusions.

The hypothesis that affective relief/disappointment may result in backward CS contrast effects was not supported. Moreover, the presence of the forward CS making the US predictable did not lead to a greater contrast between the valence of the forward and backward CSs, which may have resulted in a backward CS contrast effect. To explain why the predicted backward CS contrast effect did not occur, we turn to findings from the pain relief literature and a study on elaborated encoding in EC by Fiedler and Unkelbach (2011). Studies on pain relief and relief learning suggest that the more intense or aversive the pain eliciting stimulus, the greater the amount of pain relief experienced at its offset (Andreatta et al., 2010; Bitar, Marchand, & Potvin, 2018). Moreover, Fiedler and Unkelbach (2011) showed that increasing the relevance of the relational qualifier to the participant resulted in more elaborate encoding of the propositional information about the relation between the CS and the US, which in turn led to contrast effects. Taken together, these findings suggest that USs of higher intensity may lead to deeper processing and more

substantial encoding of CS-US relations, thus leading to larger backward CS contrast effects. While there was no relational qualifier present in the 'observe instructions' groups, it is plausible that the higher the intensity of the US, the higher the relevance and encoding of the US and its offset. This assumption suggests that if a threshold level of processing is not met (either due to a low US intensity or a lack of elaborate encoding of the US offset), contrast effects may not occur. By this reasoning, it stands that the USs employed in the main experiment may not have been intense enough to elicit affective relief/disappointment and/or may not have been relevant enough for participants to encode the offset of the US as an important event that would trigger an emotional response such as relief or disappointment.

The backward CS contrast effects observed on explicit valence ratings confirmed that the difference in results between Pilot Studies 1 and 2 were a result of including 'valence' and 'agency' components in the instructions. It is possible that the propositional information in Pilot Study 1 did not lend itself to sufficiently deep encoding to drive backward CS contrast effects, because the instructions lacked personal relevance to the participants. In the 'valence-agency' instructions, participants were told that the aliens controlled whether 'happy' or 'sad' events would happen specifically to them. This information is of greater personal relevance to participants, thereby increasing the salience of the propositional relation and, presumably, the depth at which this information is encoded. The instructions from Pilot Study 1 did not contain this level of specificity toward the participant. Therefore, the combination of low intensity pictorial USs (as compared to the auditory USs used by Moran & Bar-Anan, 2013) and the instructional manipulation used may explain the lack of contrast effects in Pilot Study 1. Furthermore, combining these same USs with an instructional manipulation that highlights personal relevance and emphasizes 'valence' and 'agency' as in Pilot Study 2 did result in backward CS contrast effects.

The overlap of CSs with USs in the 'Moran paradigm' that made the CSs look like they had control over starting and stopping the USs was shown not to be a requirement for backward CS contrast effects. Moreover, varying the duration of the USs appeared to have no effect on backward CS learning. This further supports the notion that the results from Moran and Bar-Anan (2013) and Moran et al. (2016) are purely driven by the instructional manipulation, not by the appearance that CSs control US onset and offset, or because the offset of the US was unpredictable.

Future research could investigate the importance of these parameters with more intense USs. It is possible that these parameters do not influence backward CS learning when a propositional mechanism is at play. However, it is possible that these parameters are important when affective relief/disappointment is engaged, as in the Andreatta et al. (2010; 2013) studies.

In summary, the current findings suggest that in a picture-picture paradigm affective relief/disappointment at the offset of an aversive/pleasant stimulus in the presence of a predictable US is not sufficient for backward CS contrast effects to occur. Rather, instructions determined whether backward CS valence ratings showed an assimilation or a contrast effect. These instructional manipulations are likely to interact with the properties of the US to influence CS evaluations during backward conditioning, possibly due to different levels of processing and encoding of propositional information about stimulus relations. The findings reported here clarify the effects of instructional manipulations and affective relief/disappointment in backward EC utilising picture-picture paradigms.

2.4.4 Figures

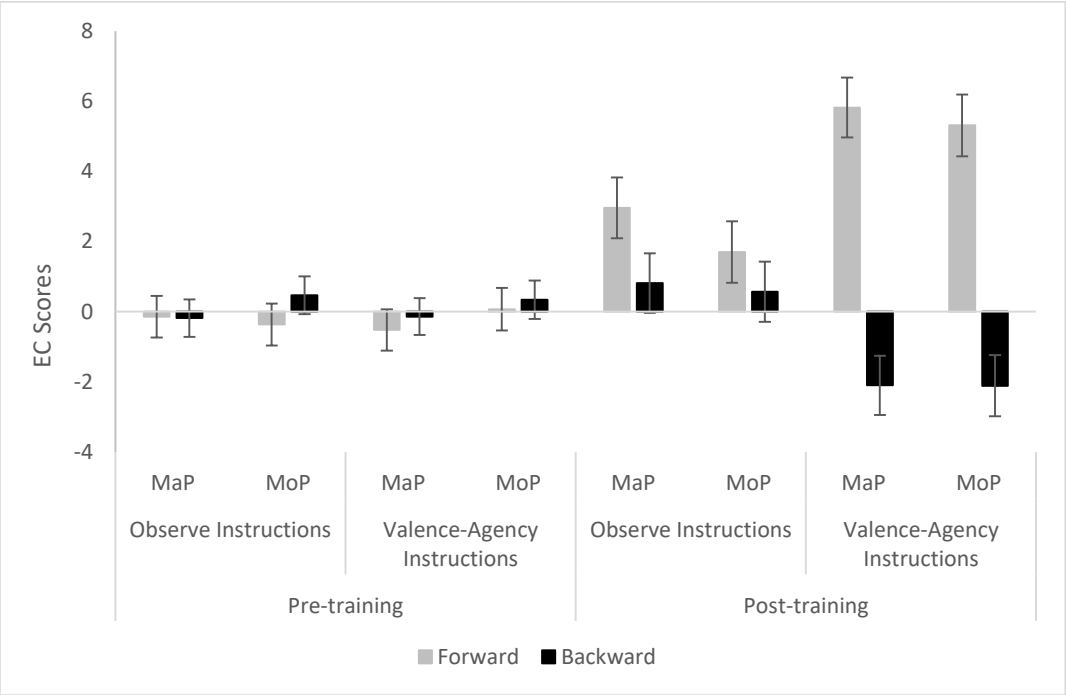

Figure 1. EC scores on explicit valence ratings for forward and backward conditioning measured pre-training and post-training as a function of instructions ('observe instructions' and 'valence-agency instructions') and paradigm ('Mallan paradigm (MaP)' and 'Moran paradigm (MoP)'). Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.

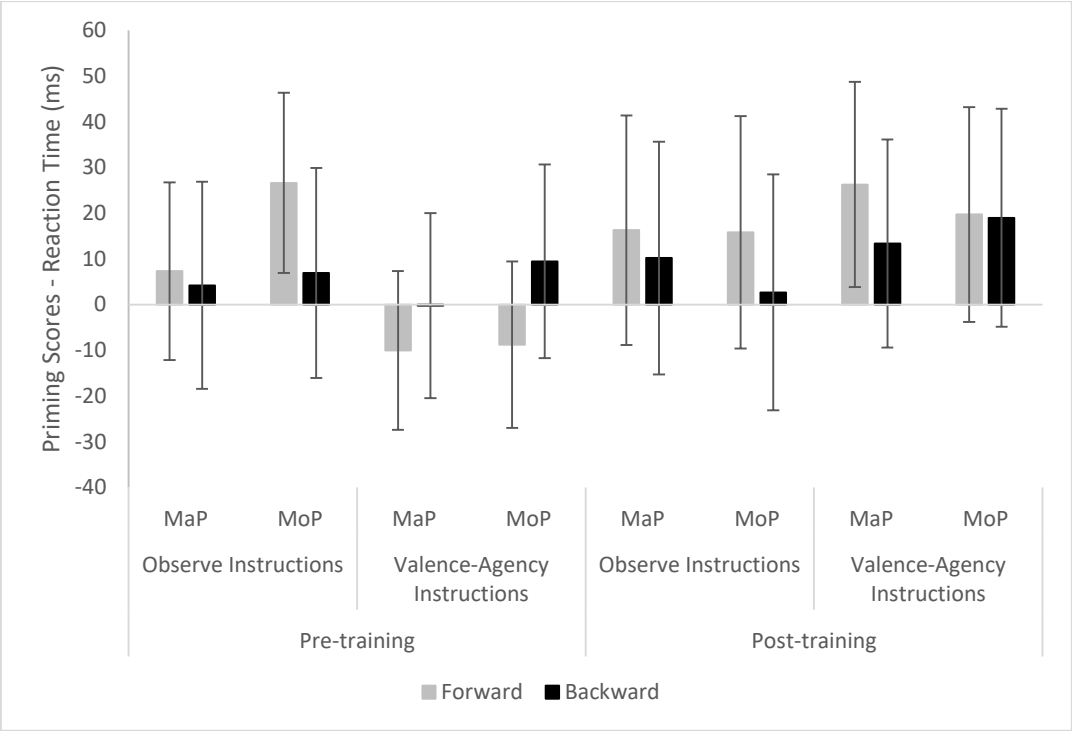

Figure 2. EC scores on affective priming for forward and backward conditioning measured pre-training and post-training as a function of instructions ('observe instructions' and 'valence-agency instructions') and paradigm ('Mallan paradigm (MaP)' and 'Moran paradigm (MoP)'). Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.

2.4.5 References

Andreatta, M., Mühlberger, A., Glotzbach-Schoon, E., & Pauli, P. (2013). Pain predictability reverses valence ratings of a relief-associated stimulus. *Frontiers in Systems Neuroscience, 7,* 1-12.

Andreatta, M., Mühlberger, A., Yarali, A., Gerber, B., & Pauli, P. (2010). A rift between implicit and explicit conditioned valence in human pain relief learning. *Proceedings of the Royal Society of London B: Biological Sciences, 277,* 2411-2416.

Andreatta, M., & Pauli, P. (2017). Learning mechanisms underlying threat absence and threat relief: Influences of trait anxiety. *Neurobiology of Learning and Memory, 145,* 105-113.

Bading, K., Stahl, C., & Rothermund, K. (2019). Why a standard IAT effect cannot provide evidence for association formation: The role of similarity construction. *Cognition and Emotion.* doi:10.1080/02699931.2019.1604322.

Bitar, N., Marchand, S., & Potvin, S. (2018). Pleasant pain relief and inhibitory conditioned pain modulation: A psychophysical study. *Pain Research and Management, 1935056.*

Center for the Study of Emotion and Attention [CSEA – NIHM] (1999). *International affective picture system: Digitized photographs*. The Center for Research in Psychophysiology, University of Florida.

Corneille, O., & Stahl, C. (2019). Associative attitudes learning: A closer look at evidence and how it relates to attitude models. *Personality and Social Psychology Review, 23,* 161-189.

De Houwer, J. (2007). A conceptual and theoretical analysis of evaluative conditioning. *The Spanish Journal of Psychology, 10,* 230-241.

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

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,* 853-869.

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,* 1013-1027.

Fiedler, K., & Unkelbach, C. (2011). Evaluative conditioning depends on higher order encoding processes. *Cognition and Emotion, 25,* 639-656.

Galdi, S., Arcuri, L., & Gawronski, B. (2008). Automatic mental associations predict future choices of undecided decision makers. *Science, 321,* 1100-1102.

Gast, A, & Rothermund, K. (2011). What you see is what will change: Evaluative conditioning effects depend on a focus on valence. *Cognition and Emotion, 25,* 89-110.

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

Gawronski, B., & Bodenhausen, G. V. (2011). The associative-propositional evaluation model: Theory, evidence, and open questions. *Advances in Experimental Social Psychology, 44,* 59-127.

Gawronski, B., & Bodenhausen, G. V. (2018). Evaluative conditioning from the perspective of the associative-propositional evaluation model. *Social Psychological Bulletin, 13*(3), e28024.

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.

Gerber, B., Yarali, A., Diegelmann, S., Wotjak, C. T., Pauli, P., & Fendt, M. (2014). Pain relief learning in flies, rats, and man: Basic research and applied perspectives. *Learning and Memory, 21*, 232-252.

Gibson, B. (2008). Can evaluative conditioning change attitudes toward mature brands? New evidence from the Implicit Association Test. *Journal of Consumer Research, 35,* 178-188.

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

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

Hu, X., Gawronski, B., & Balas, R. (2017a). 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,* 17-32.

Hu, X., Gawronski, B., & Balas, R. (2017b). Propositional versus dual-process accounts of evaluative conditioning: II. The effectiveness of counter-conditioning and counter-instructions in changing implicit and explicit evaluations. *Social Psychological and Personality Science, 8,* 858-866.

Inquisit 4 [Computer software]. (2016). Retrieved from https://www.millisecond.com.

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,* 61-68.

Koppehele-Gossel, J., Hoffmann, L., Banse, R., & Gawronski, B. (in press). Evaluative priming as an implicit measure of evaluation: An examination of

outlier-treatments for evaluative priming scores. *Journal of Experimental Social Psychology.*

LeBel, E. P., & Campbell, L. (2009). Implicit partner affect, relationship satisfaction, and the prediction of romantic breakup. *Journal of Experimental Social Psychology, 45,* 1291-1294.

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 49,* 433-442.

Luck, C. C., & Lipp, O. V. (2017). Startle modulation and explicit valence evaluations dissociate during backward fear conditioning. *Psychophysiology, 54,* 673-683.

Mallan, K. M., Lipp, O. V., & Libera, M. (2008). Affect, attention, or anticipatory arousal? Human blink startle modulation in forward and backward affective conditioning. *International Journal of Psychophysiology, 69.* 9-17.

Mathôt, S. (2017) Bayes like a Baws: Interpreting Bayesian repeated measures in JASP. Retrieved from https://www.cogsci.nl/blog/interpreting-bayesian-repeated-measures-in-jasp

Moran, T., & Bar-Anan, Y. (2013). The effect of object-valence relations on automatic evaluation. *Cognition and Emotion, 27,* 743-752.

Moran, T., & Bar-Anan, Y. (2019). The effect of co-occurrence and relational information on speeded evaluation. *Cognition and Emotion*. doi:10.1080/02699931.2019.1604321

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2016). The assimilative effect of co-occurrence on evaluation above and beyond the effect of relational qualifiers. *Social Cognition, 34.* 435-461.

Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science, 12,* 413-147.

2.4.6 Footnotes

[1] ©American Psychological Association, [2019]. This paper is not the copy of record and may not exactly replicate the authoritative document published in the APA journal. Please do not copy or cite without author's permission. The final article is available, upon publication, at https://dx.doi.org/10.1037/emo0000701

[2] All materials, data, and analysis files are available at https://osf.io/ur5kd/.

2.5 Supplementary Materials:

2.5.1 Pilot Study 1

To investigate whether affective relief/disappointment alone would elicit backward CS contrast effects in a within-subjects forward and backward conditioning procedure, we combined the concurrent forward and backward conditioning procedure and stimuli from Moran and Bar-Anan (2013) with the stimulus presentation parameters from Mallan et al. (2008). In addition, we manipulated between groups whether participants received the instructions used by Moran and Bar-Anan (2013; termed the *start-stop instructions* group) or the instructions used by Mallan et al. (2008; termed the *observe instructions* group). We expected to observe contrast effects for backward CSs and assimilation effects for forward CSs in both groups on explicit valence ratings. Based on the recent findings of Bading, Stahl, and Rothermund (2019), and Hu, Gawronski, and Balas (2017a), we expected to observe assimilation effects on implicit measures for forward and backward conditioning in both groups.

2.5.1.1 Method

Participants and design. Participants were recruited through Amazon Mechanical Turk (M-Turk) through TurkPrime (Litman, Robinson, & Abberbock, 2016). The sample comprised 94 participants (mean age = 35.67, *SD* = 9.84) after duplicates and participants failing to complete the experiment were removed (*n* = 19). The sample size was based on Moran and Bar-Anan (2013) and Moran, Bar-Anan, and Nosek (2016) who had sample sizes ranging from 32 to 68 participants. In these studies, the within-subjects interaction of interest yielded large effects sizes between $\eta_p^2 = .15$ and $\eta_p^2 = .60$. Based on these effect sizes, we anticipated that approximately 50 participants per group would provide sufficient power to detect the effects of interest. The start-stop instructions group consisted of 42 participants (17 female, 25 male); the observe instructions group included 52 participants (24 female, 28 male). Groups did not differ on gender, $\chi^2(1) = 0.31$, $p = .581$, ethnicity, $\chi^2(4) = 6.86$, $p = .144$, or age, $t(92) = 1.35$, $p = .179$, $d = 0.28$, $BF_{10} = 0.49$. A 2 (Instructions: start-stop vs. observe; between-participants) × 2 (Conditioning Type: forward vs. backward; within-participants) × 2 (Time: pre-test vs post-test; within-participants) × 2 (US Valence: positive vs. negative; within-participants) mixed design was used to

determine the effect of instruction type on CS valence after a forward and backward conditioning procedure.

Explicit valence ratings. Each CS was presented one-by-one and participants were asked to rate how pleasant they found the stimulus on a 9-point scale ranging from 1 (*unpleasant*) to 9 (*pleasant*).

Affective priming task. Each of the four CSs was presented once with 10 positive target words and 10 negative target words for a total of 80 trials. A fixation cross was presented for 500ms, followed by the CS prime for 200ms, and then the target word until the participant provided their response. Participants were instructed to press the *I* key if the target word was positive and the *E* key if the target word was negative. Target words were taken from Hu et al. (2017a, 2017b). The positive words were *pleasant, good, outstanding, beautiful, magnificent, marvellous, excellent, appealing, delightful,* and *nice*. The negative words were *unpleasant, bad, horrible, miserable, hideous, dreadful, painful, repulsive, awful,* and *ugly.*

Recollective Memory Test. For exploratory purposes, the current study also included measures of recollective memory. In the observe instructions group, participants were shown each CS and asked: "Circle the appropriate answer below. Was this picture presented: Together with pleasant pictures, together with unpleasant pictures, together with pleasant and unpleasant pictures, I did not see this picture, I could not tell?" In the start-stop instructions group, participants were shown each CS and asked: "Circle the appropriate answer below. What is the role of this creature: To start pleasant pictures, to stop pleasant pictures, to start unpleasant pictures, to stop unpleasant pictures?" Using the sum of correct responses on the memory test, accuracy scores on the test could range from zero to four. Both groups were also presented with each US and each CS, and asked to indicate which CS came before or after each US. This procedure resulted in an accuracy score ranging from 0 to 16. Participants were classified as remembering the CS-US contingencies only if they scored 100% on both memory tests. The analyses of the recollective memory data for Pilot Study 1 and all subsequent experiments did not add substantially to the current report and are available in the additional analyses section below.

Demographics questionnaire. Participants were asked to report their age, gender, and ethnicity, and to provide information about the environment in which they completed the task, and if they had any comments.

Apparatus/stimuli. Four images of aliens, one from each of the four families of alien creatures created by Moran and Bar-Anan (2013), were used as CSs (see below; materials from Moran and Bar-Anan, 2013, available at https://osf.io/cqsnj/). Each alien differed in colour and head shape. Four positive and four negative pictures from the International Affective Picture System (IAPS; CSEA, 1999) were used as USs (1050, 1300, 1440, 1710, 5833, 6313, 6560, and 8190). Inquisit 4 Web by Millisecond Software ™ (2016) was used to run the experiment and to record responses in all tasks.

Procedure. Participants selected the HIT (human intelligence task) on M-Turk and read the description of the study. When participants began the study, they were presented with an information sheet outlining the tasks, informed that they could withdraw at any time by pressing 'ctrl + q', and then prompted to press 'continue' if they consented to participate. Informed consent was implied if participants pressed 'continue'. Next, the first explicit valence ratings and affective priming task was presented followed by the training phase. The training phase comprised 12 positive and 12 negative trials presented pseudo-randomly, with inter-trial intervals of 4, 6, and 8 seconds. Each trial consisted of a forward CS, followed by a positive or negative US, followed by a backward CS. This CS-US-CS paradigm was adapted from Moran and Bar-Anan (2013), with some modifications based on Mallan et al. (2008). We used one CS from each of the four alien families, four positive and four negative pictures as USs, and each stimulus was presented for 4 seconds with onset and offsets coinciding. CSs were counter-balanced using a Latin square resulting in four CS orders, with each CS occurring in each role equally.

In the *start-stop instructions* group, participants received the following instructions adapted from Moran and Bar-Anan (2013) before the training phase:

> *In this task you are going to see two types of pictures. Pleasant pictures: Including pictures of scenery and animals. Unpleasant pictures: Including pictures of threat of violence and aggressive animals. These pictures will be shown multiple times. Before and after the pleasant and unpleasant pictures, four different types of creatures will appear on the screen. Each creature will have a fixed role out of four possible roles. Your task is to learn which role each creature plays. At the end of the experiment we will examine your memory of the role played by each creature. The four possible roles are: After the appearance of one creature, a pleasant picture appears. On the*

appearance of one creature, a pleasant picture ends. After the appearance of one creature, an unpleasant picture appears. On the appearance of one creature, an unpleasant picture ends. The four creatures that will start and stop the pictures are:

The experiment is about to begin. Remember, you must learn the role of each of the creatures: Which creature starts the pleasant pictures? Which creature stops the pleasant pictures? Which creature starts the unpleasant pictures? Which creature stops the unpleasant pictures? At the end of the experiment we will examine what you have learned. To start the experiment, press the space bar.

In the *observe instructions* group, participants received the following instructions adapted from Mallan et al. (2008):

In this task you will be presented with a series of pictures. Please pay attention to which pictures follow each other as you will be tested on this at the end of the experiment.

After the training phase, the second explicit valence ratings and affective priming task was presented, followed by the memory test and demographics questionnaire. Participants then received a completion code to receive their compensation, and were thanked for their participation. The experiment took approximately 20 minutes on average to complete, and participants were compensated US-$4.80.

Statistical analyses. Frequentist analyses were performed using IBM SPSS Statistics 25 and Bayesian analyses were performed using JASP 0.10.0.0. Bayesian follow-up analyses were performed using the BayesFactor package in R. For explicit valence ratings, EC scores were calculated as the difference between ratings of CSs paired with positive USs and ratings of CSs paired with negative USs. EC scores were calculated separately for forward vs. backward conditioning and for pre-training vs. post-training. Positive EC scores represent an assimilation effect and negative EC scores represent a contrast effect. In the affective priming task, trials on which target words were categorised incorrectly were scored as error trials. Trials on which reaction times were shorter than 300ms and longer than 1000ms were categorised as outliers, as they were deemed to be outside the window of a valid response (see Koppehele-Gossel, Hoffmann, Banse, & Gawronski, in press). Participants with a percentage of invalid trials larger than 25% were removed from

the analyses (n = 5 in start-stop instructions group, n = 8 in observe instructions group). In the final sample at pre-test, 5.43% of trials were errors from incorrect categorisation of target words and 4.77% of trials were outliers. At post-test, 5.58% of trials were errors from incorrect categorisation of target words and 6.33% of trials were outliers. Priming scores were calculated as the difference in response times between incongruent and congruent trials: (CSs paired with positive USs/negative target words + CSs paired with negative USs/positive target words) – (CSs paired with positive USs/positive target words + CSs paired with negative USs/negative target words). Priming scores were calculated separately for forward vs. backward conditioning and for pre-training and post-training. Positive priming scores suggest an assimilation effect, while negative scores suggest a contrast effect. EC scores and priming scores from reaction time data were subjected to 2 (Instructions: start-stop vs. observe; between-participants) × 2 (Conditioning Type: forward vs. backward; within-participants) × 2 (Time: pre-test vs post-test; within-participants) mixed ANOVAs, and significant interactions were followed-up with pairwise comparisons and one sample *t*-tests where appropriate. Pillai's trace values of the multivariate solution are reported for main effects and interactions ($\alpha = .05$). The reliability of the priming task was $\alpha = -.62$ at pre-test, and $\alpha = .16$ at post-test. The analyses of the error data from the affective priming task for Experiment 1 and all subsequent experiments did not add substantially to the current report and are available in the additional analyses section below.

2.5.1.2 Results

Explicit valence ratings. Mean EC scores are depicted in Figure S1. The ANOVA revealed significant main effects of Conditioning Type, F (1, 92) = 24.73, p < .001, η_p^2 = .212, BF_{10} = 618.85, and Time, F (1, 92) = 56.80, p < .001, η_p^2 = .382, BF_{10} = 5.41 × 10^9, which were qualified by a significant two-way interaction between Conditioning Type and Time, F (1, 92) = 31.21, p < .001, η_p^2 = .253, BF_{10} = 16794.55, and a significant two-way interaction between Conditioning Type and Instructions, F (1, 92) = 8.93, p = .004, η_p^2 = .088, BF_{10} = 13.65. The three-way interaction between Instructions, Conditioning Type, and Time was not significant, F (1, 92) = 2.14, p = .146, η_p^2 = .023, BF_{incl} = 0.52. Follow-up analyses for the Conditioning Type × Time interaction revealed no difference between the two conditioning type conditions at pre-training, F (1, 92) = 0.01, p = .912, η_p^2 = .000, BF_{10} = 0.12, and a significantly larger EC score for forward conditioning than

backward at post-training, F (1, 92) = 37.35, $p < .001$, $\eta_p^2 = .289$, $BF_{10} = 85889.24$. One sample *t*-tests confirmed that EC scores were significantly larger than zero for both forward conditioning, t (93) = 10.24, $p < .001$, $d = 1.06$, $BF_{10} = 9.50 \times 10^{13}$, and backward conditioning, t (93) = 2.51, $p = .014$, $d = 0.26$, $BF_{10} = 2.19$, at post-training, but not pre-training, t (93) = 0.31, $p = .759$, $d = 0.03$, $BF_{10} = 0.12$, and, t (93) = 0.73, $p = .470$, $d = 0.07$, $BF_{10} = 0.15$, respectively. Follow-up analyses for the Conditioning Type × Instructions interaction revealed that, averaged across pre- and post-test, EC scores for forward conditioning were significantly larger than EC scores for backward conditioning in the start-stop instructions group, F (1, 92) = 28.65, $p < .001$, $\eta_p^2 = .237$, $BF_{10} = 860.85$, but not the observe instructions group, F (1, 92) = 2.20, $p = .141$, $\eta_p^2 = .023$, $BF_{10} = 0.44$.

Affective priming. Mean affective priming scores are depicted in Figure S2. The ANOVA revealed only a significant main effect of Time, F (1, 79) = 6.29, $p = .014$, $\eta_p^2 = .074$, $BF_{10} = 2.79$, indicating that priming scores were significantly larger at post-training than pre-training. Follow-up analyses revealed that priming scores were significantly larger than zero at post-training, t (80) = 3.02, $p = .003$, $d = 0.34$, $BF_{10} = 5.43$, but not pre-training, t (80) = 0.33, $p = .746$, $d = 0.04$, $BF_{10} = 0.09$. The three-way interaction between Instructions, Conditioning Type, and Time, was not significant, F (1, 79) = 0.11, $p = .739$, $\eta_p^2 = .001$, $BF_{incl} = 0.26$.

2.5.1.3 Discussion

Explicit valence ratings and reaction time priming scores revealed assimilation effects for forward and backward conditioning. Bayesian analyses provided weak to moderate evidence for the null hypothesis, suggesting that no differences on EC scores between instruction groups were present. Thus, the hypothesis that both groups would show backward CS contrast effects was not supported. It is possible that the 'start-stop instructions' are not sufficient to produce backward CS contrast effects without additional procedural details of Moran and Bar-Anan's (2013) paradigm. The paradigm we employed was different from Moran and Bar-Anan's, because we used multiple USs, single CSs, and the same stimulus presentation timing as Mallan et al. (2008). For example, it is possible that removing the overlap between US offset and CS onset and removing the varying US durations that made US offset predictable rendered the instructions less effective. We also measured explicit valence ratings and presented the affective priming task before and after the training phase, whereas Moran and Bar-Anan (2013) only presented their

measures of CS valence after the training phase. This may have resulted in an 'evaluative mindset' that led participants to evaluate stimuli differently than in Moran and Bar-Anan (2013; Gast & Rothermund, 2011). Furthermore, it could be that the 'start-stop instructions' produce backward CS contrast effects only when using acoustic USs, or only when the USs are more intense and/or more salient than the picture USs used here. Consistent with this possibility, Moran et al. (2016) found backward CS contrast effects in a picture-picture paradigm using instructions that emphasised the valence of the US (i.e., "getting gold bars is a happy event, whereas getting garbage piles is a sad event") and the agency of the families in starting and stopping the US (i.e., "creatures control whether happy or sad events happen to you").

Before we can draw conclusions about our hypothesis regarding US predictability and affective relief/disappointment eliciting backward CS contrast effects, we need to ensure that we can elicit backward CS contrast effects with instructions in a picture-picture paradigm. Moreover, we need to assess whether putting participants in an 'evaluative mindset' by presenting valence ratings and affective priming before the learning phase is affecting backward CS learning. To achieve this we decided to replicate Moran et al.'s (2016) findings using their exact paradigm and instructions to determine whether US and CS overlap, US variability or an 'evaluative mindset' may have contributed to not observing backward CS contrast effects in the 'start-stop instructions' group of Experiment 1.

2.5.2 Pilot Study 2

Pilot Study 2 had two aims: (1) to replicate Experiment 1 from Moran et al. (2016), as we did not find contrast effects in our first experiment using a slightly different paradigm and using instructions that were based on an earlier study employing sound USs (Moran & Bar-Anan, 2013), and (2) to assess whether having participants complete explicit valence ratings and an affective priming task before conditioning puts them in an 'evaluative mindset' that results in a failure to find contrast effects in backward conditioning. If contrast effects emerge for backwardly conditioned CSs in both groups in Pilot Study 2, the lack of a contrast effect in Pilot Study 1 may be due to the differences in the paradigm or instructions. If contrast effects emerge for backwardly conditioned CSs in the no pre-measure group only,

the lack of contrast effects observed in Pilot Study 1 may be due to putting participants in an 'evaluative mindset' before the conditioning task.

2.5.2.1 Method

Participants and design. Participants were recruited through M-Turk. The sample comprised 95 participants, with a mean age of 36.57, *SD* = 10.504, after duplicates and those failing to complete the experiment were removed (*n* = 18). As in Pilot Study 1, the sample size was based on previous research (Moran & Bar-Anan, 2013; Moran et al., 2016). The 'no pre-measure' group consisted of 48 participants (19 female) and the 'pre-measure' group included 47 participants (23 female). Groups did not differ on gender, χ^2 (2) = 2.06, p = .356, ethnicity, χ^2 (4) = 0.42, p = .981, or age, t (93) = 0.81, p = .418, d = 0.17, BF_{10} = 0.29. A 2 (Pre-measure: pre-measure vs no pre-measure; between-participants) × 2 (Conditioning Type: forward vs backward; within-participants) × 2 (US Valence: positive vs negative; within-participants) mixed design was used to replicate Moran et al. (2016) and to determine whether presenting explicit valence rating and affective priming tasks before conditioning affects the pattern of responding on these tasks after conditioning. In the pre-measure group, participants completed an explicit valence rating and affective priming task before and after conditioning. In the no pre-measure group, which was a direct replication of the first experiment in Moran et al. (2016), participants completed a conditioning task, followed by explicit valence ratings and an affective priming task. Both groups then completed a recollective memory test and a demographics questionnaire.

Explicit valence ratings. Each CS family was presented alone and participants were asked "Based on your very first emotional response, how much do you like the creatures in the picture? Click the appropriate answer below: dislike strongly, dislike moderately, dislike slightly, like slightly, like moderately, like strongly".

Affective priming task. Two creatures from each family were presented with positive and negative words twice, and two creatures from each family were presented with positive and negative words three times, for a total of 10 positive and 10 negative word pairings per family. This resulted in 80 trials. All other details were the same as in Pilot Study 1.

Recollective memory test. Each CS family was presented alone and participants were asked "In the game, what was the role of the creatures in the

picture? Click the appropriate answer below: Starting gold, starting garbage, stopping gold, stopping garbage?"

Demographics questionnaire. The demographics questionnaire was identical to Experiment 1.

Apparatus/stimuli. CSs and USs were those used by Moran et al. (2016; available at https://osf.io/v2trw/). CSs were four families of alien creatures, with each family comprising four creatures for a total of 16 CSs. The positive US was a picture of puppies, gold bars, and a baby, presented next to each other as a single image, and the negative US was a picture of an aggressive dog, garbage, and a crying child presented next to each other as a single image. Inquisit 4 by Millisecond Software ™ (2016) was used to run the experiment and to record responses in all tasks.

Procedure. The 'pre-measure' group completed explicit valence ratings and an affective priming task before the training phase, whilst the 'no pre-measure' group went straight to the training phase. The training phase comprised 12 positive and 12 negative trials randomly presented with inter-trial intervals of 2s. Each trial consisted of a forward CS, followed by a positive or negative US, followed by a backward CS. This CS-US-CS paradigm was an exact replication of Moran et al. (2016). CSs were presented for 1.5 seconds and USs were presented in blocks of 1s flashes with a 200ms break between each flash for a total of 3 or 5s of total US presentation time. Onset of the US coincided with offset of the forward CS, and onset of the backward CS occurred 200ms after the last US appearance.

Prior to the training phase, we presented participants with the exact instructions used by Moran et al. (2016; termed 'valence-agency instructions'). These instructions differ slightly from those used by Moran and Bar-Anan (2013; termed 'start-stop instructions'), as they highlight valence and agency. The instructions read as follows:

> *In the next game, you will get piles of shiny gold bars, but also some stinky garbage piles. Getting gold bars is a happy event, whereas getting garbage piles is a sad event. In the game, four families of creatures control whether happy or sad events happen to you. These are the four families. One family of creatures will always start the gold bars coming your way. A second family of creatures will always stop the gold bars. A third family of creatures will always start garbage piles coming your way. A fourth family of creatures will*

always stop the garbage piles. Your goal in this game is to learn which family of creatures starts the gold, which family stops the gold, which family starts the garbage, and which family stops the garbage. We will test your learning later in the game, so please pay close attention. If you read and understood the instructions, hit the spare bar to continue. Please pay close attention to the images on the screen. Make sure you learn and remember which family does each of the four actions (start gold, stop gold, start garbage, stop garbage). Press space to start the game.

After 12 trials, the following instructions were presented:

Do you know by now which family starts the gold, which family stops the gold, which family starts the garbage, and which family stops the garbage? Try to memorize what each family does for a later test. Press space for a few more rounds to help you remember the roles of the families better.

After the training phase, the explicit valence ratings and affective priming tasks were presented, followed by the recollective memory test and demographics questionnaire. Participants then received a completion code to enter to receive their payment, and thanked for participating. The experiment took 13 minutes on average to complete, and participants were compensated US-$5.

Statistical analyses. Responses following CSs within the same family in the affective priming task were averaged to provide overall means for each family. EC scores were calculated following the procedures in Pilot Study 1. EC scores and priming scores were subjected to separate 2 (Pre-measure: pre-measure vs no pre-measure; between-participants) × 2 (Conditioning Type: forward vs backward; within-participants) mixed-model ANOVAs. Participants with a percentage of invalid trials larger than 25% were removed from the priming analyses (n = 4 in the 'pre-measure' group, n = 6 in the 'no pre-measure' group). In the final sample at pre-test, 6.62% of trials were incorrectly categorised and 7.47% of trials were outliers. At post-test, 7.50% of trials were incorrectly categorised and 6.33% of trials were outliers. The reliability of the priming task was $\alpha = .13$ at pre-test, and $\alpha = .45$ at post-test. All other details were the same as in Experiment 1.

2.5.2.2 Results

Explicit valence ratings. Figure S3 shows mean EC scores at post-test as a function of Conditioning Type and Pre-measure. The figure suggests assimilation effects for forward conditioning and contrast effects for backward conditioning in

both the ‘pre-measure’ and the ‘no-pre-measure’ groups. A main effect of Conditioning Type revealed that forward conditioning EC scores were significantly larger than backward conditioning EC scores, F (1, 93) = 223.65, $p < .001$, $\eta_p^2 = .706$, $BF_{10} = 1.81 \times 10^{40}$. One-sample t-tests further showed that forward conditioning EC scores were significantly larger than zero, t (94) = 18.67, $p < .001$, $d = 1.92$, $BF_{10} = 1.30 \times 10^{30}$, and backward conditioning EC scores were significantly smaller than zero, t (94) = 7.49, $p < .001$, $d = 0.77$, 2.45×10^8. The two-way interaction between Instructions and Conditioning Type was not significant, F (1, 93) = 1.21, $p = .274$, $\eta_p^2 = .013$, $BF_{incl} = 0.49$.

Affective priming. Figure S4 shows mean EC scores on affective priming at post-test as a function of Conditioning Type and Pre-measure. A marginal main effect of Conditioning Type suggests a larger priming score for forward conditioning than backward conditioning, F (1, 83) = 3.42, $p = .068$, $\eta_p^2 = .040$, $BF_{incl} = 0.99$. The two-way interaction between Instructions and Conditioning Type was not significant, F (1, 83) = 1.14, $p = .289$, $\eta_p^2 = .014$, $BF_{incl} = 0.41$. One sample t-tests showed that only priming scores for forward conditioning were significantly larger than zero, t (84) = 2.53, $p = .013$, $d = 0.27$, $BF_{10} = 2.39$. Priming scores for backward conditioning were not significantly different from zero, t (84) = 0.04, $p = .972$, $d < 0.01$, $BF_{10} = 0.12$.

2.5.2.3 Discussion

In the current study, explicit valence ratings showed assimilation effects for forward CSs and contrast effects for backward CSs in both groups, and this pattern emerged regardless of whether participants did or did not complete measures of CS valence before the training phase. In addition to successfully replicating the backward CS contrast effects observed by Moran et al. (2016), these findings suggest that the lack of a contrast effect for backwardly conditioned CSs in Pilot Study 1 was not due to participants being in a ‘evaluative mindset’. However, differing from Moran and Bar-Anan’s (2013) findings, affective priming scores revealed a significant but weak assimilation effect only for forward, but not for backward, CSs.

Although Pilot Study 2 rules out an ‘evaluative mindset’ as a potential explanation for the lack of a backward contrast effects on explicit valence ratings in Pilot Study 1, it is possible that the difference between the findings in Pilot Studies 1 and 2 is due to the procedural differences between the two studies mentioned in the Pilot Study 1 discussion. Whereas the conditioning procedure in Pilot Study 1

aligned more closely with Mallan et al.'s (2008) paradigm, the conditioning procedure in Pilot Study 2 directly replicated Moran et al.'s (2016) paradigm which included US/CS overlap and US variability. These parameters may increase the influence of the instructions and may also increase the amount of affective relief/disappointment that occurs at the offset of the US. On the other hand, the different results may have been due to differences between the instructions used in Pilot Study 1 (taken from Moran & Bar-Anan, 2013) and those used in Pilot Study 2 (taken from Moran et al., 2016), which different in the strength and perception of control the CSs possess ('valence' and 'agency' components). Finally, the observed backward CS contrast effect in Pilot Study 2 may have been artificially amplified by using a 6-point explicit valence rating scale instead of the 9-point scale used in Pilot Study 1. The lack of a midpoint on the scale forces participants to choose either 'dislike slightly', or 'like slightly', which may have pushed those who would have rated CSs as neutral to select a specific valence. This aspect of the rating scale may have resulted in a significant backward CS contrast effect in Pilot Study 2 that was not observed in Pilot Study 1, because participants had the option to rate stimuli as neutral.

No priming effects were found for backward CSs which deviates from previous research and a-priori predictions. It is possible that presenting multiple exemplars of the same set of CSs reduced an already small effect to a null effect as supported by the Bayesian analyses. It is also possible that competing assimilation and contrast effects cancelled each other out leading to a null effect. This may be plausible given the speeded nature of the task. However, results regarding the effects of speeded tasks (i.e., valence rating tasks that impose time limits on responding, as well as other implicit measures) on backward CS evaluations in this paradigm are currently inconclusive (see Moran & Bar-Anan, 2019).

The concerns regarding explicit valence ratings mentioned above, and potential moderators of the backward CS contrast effects present in Pilot Study 2, but not in Pilot Study 1, were addressed in the main experiment by contrasting paradigm ('Mallan paradigm' from Pilot Study 1 vs. 'Moran paradigm' from Pilot Study 2) and instructions ('observe instructions' from Pilot Study 1 vs. 'valence-agency instructions' from Pilot Study 2) in a mixed factorial design.

2.5.2.4 Figures from Pilot Study 1 and 2

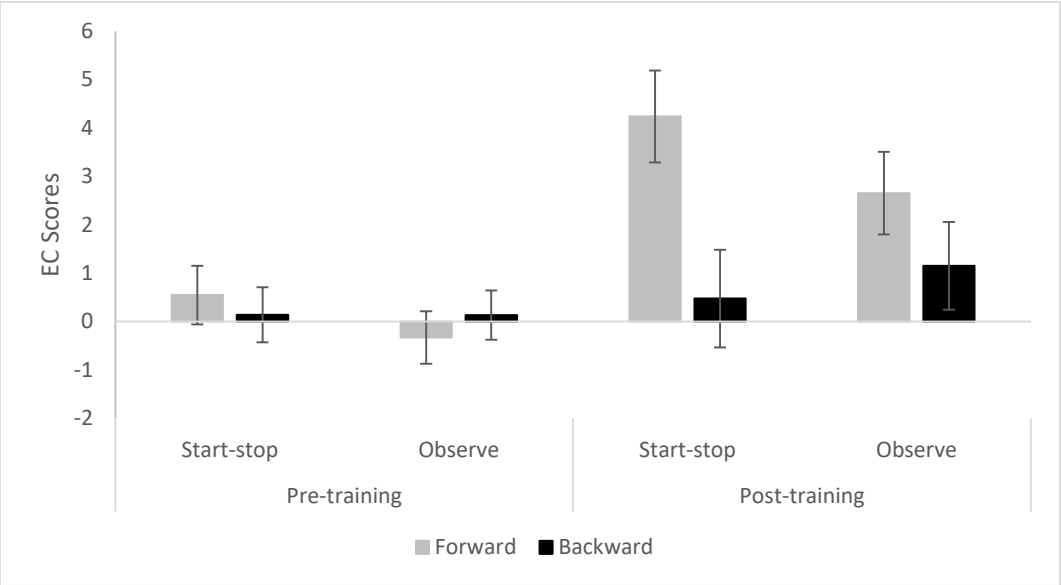

Figure S1. EC scores on explicit valence ratings for forward and backward conditioning measured pre-training and post-training for the start-stop instructions and observe instructions groups, Pilot Study 1. Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.

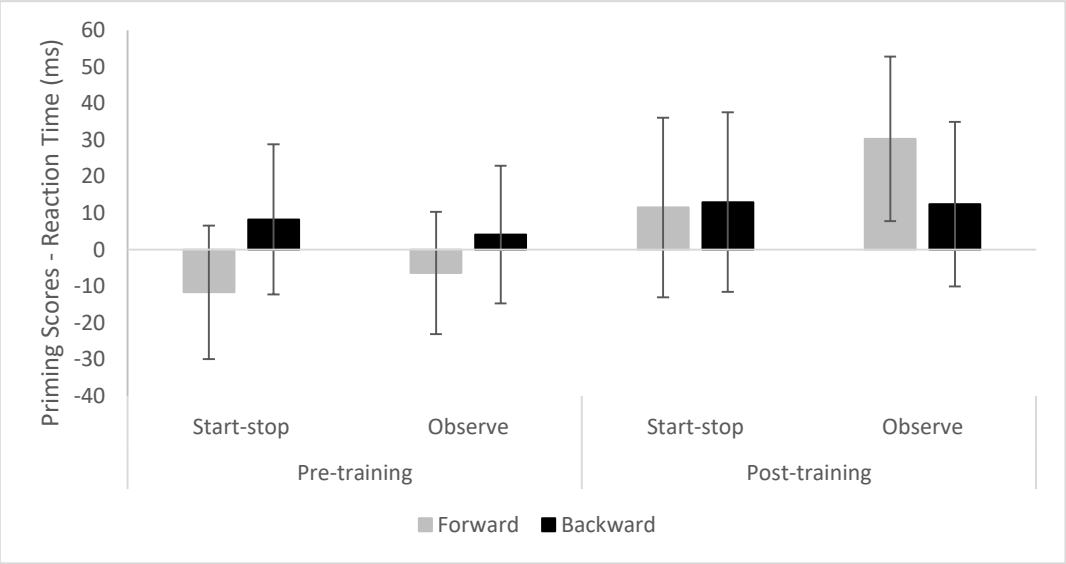

Figure S2. EC scores on affective priming for forward and backward conditioning, measured pre-training and post-training for the start-stop instructions and observe instructions groups, Pilot Study 1. Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.

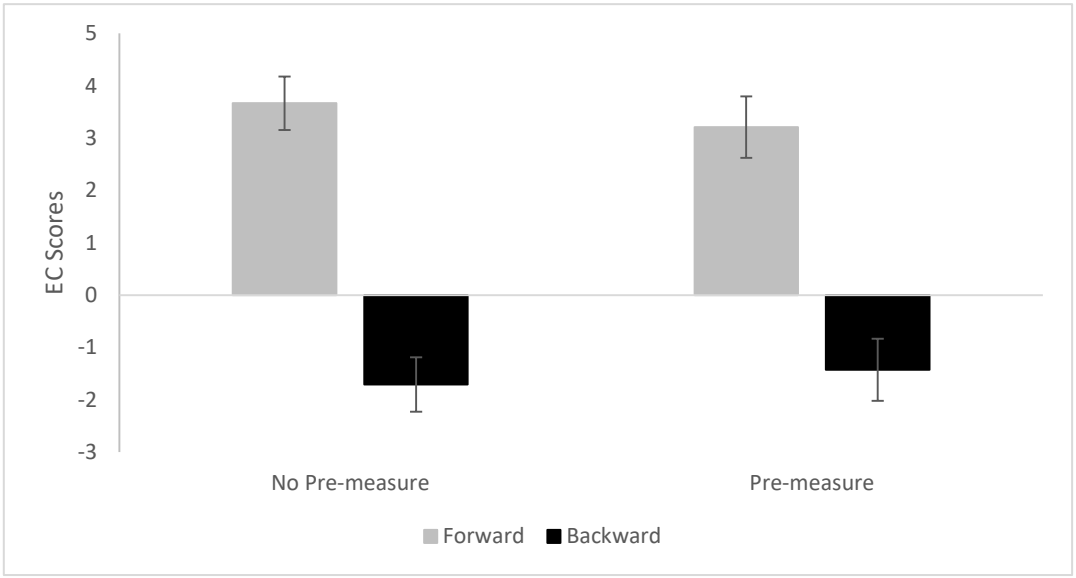

Figure S3. Post-test EC scores on explicit valence ratings for forward and backward conditioning for the 'no pre-measure' and 'pre-measure' groups, Pilot Study 2. Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.

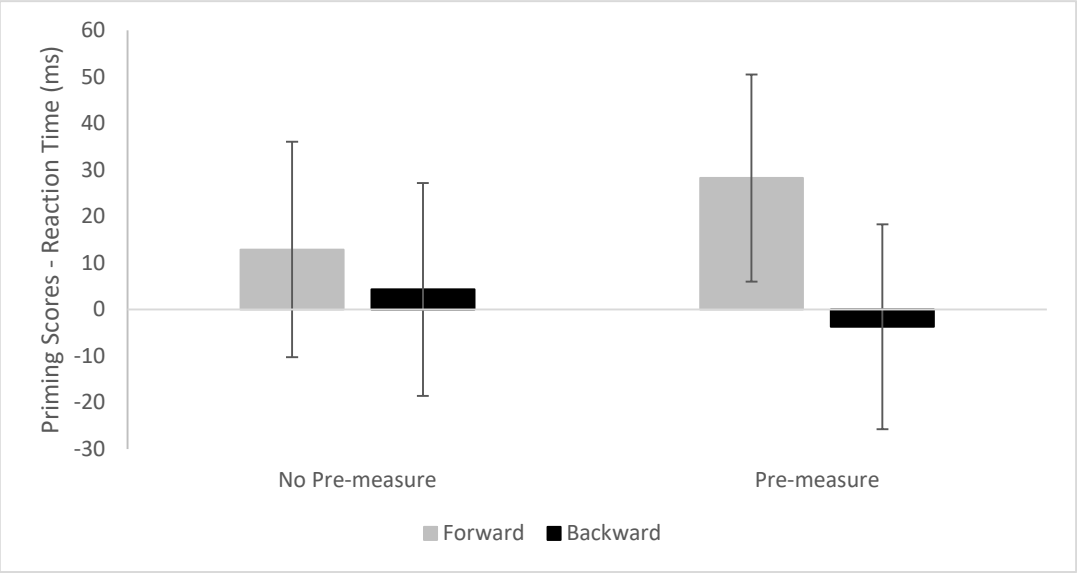

Figure S4. Post-test EC scores on affective priming for forward and backward conditioning for the 'no pre-measure' and 'pre-measure' groups, Pilot Study 2. Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.

2.5.2.5 References

Bading, K., Stahl, C., & Rothermund, K. (2019). Why a standard IAT effect cannot provide evidence for association formation: The role of similarity construction. *Cognition and Emotion.* doi:10.1080/02699931.2019.1604322.

Center for the Study of Emotion and Attention [CSEA – NIHM] (1999). *International affective picture system: Digitized photographs*. The Center for Research in Psychophysiology, University of Florida.

Gast, A, & Rothermund, K. (2011). What you see is what will change: Evaluative conditioning effects depend on a focus on valence. *Cognition and Emotion, 25,* 89-110.

Hu, X., Gawronski, B., & Balas, R. (2017a). 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,* 17-32.

Hu, X., Gawronski, B., & Balas, R. (2017b). Propositional versus dual-process accounts of evaluative conditioning: II. The effectiveness of counter-conditioning and counter-instructions in changing implicit and explicit evaluations. *Social Psychological and Personality Science, 8,* 858-866.

Inquisit 4 [Computer software]. (2016). Retrieved from https://www.millisecond.com.

Koppehele-Gossel, J., Hoffmann, L., Banse, R., & Gawronski, B. (in press). Evaluative priming as an implicit measure of evaluation: An examination of outlier-treatments for evaluative priming scores. *Journal of Experimental Social Psychology.*

Litman, L., Robinson, J., & Abberbock, T. (2016). TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. Behavior Research Methods, 1-10.

Mallan, K. M., Lipp, O. V., & Libera, M. (2008). Affect, attention, or anticipatory arousal? Human blink startle modulation in forward and backward affective conditioning. *International Journal of Psychophysiology, 69.* 9-17.

Moran, T., & Bar-Anan, Y. (2013). The effect of object-valence relations on automatic evaluation. *Cognition and Emotion, 27,* 743-752.

Moran, T., & Bar-Anan, Y. (2019). The effect of co-occurrence and relational information on speeded evaluation. *Cognition and Emotion*. doi:10.1080/02699931.2019.1604321

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2016). The assimilative effect of co-occurrence on evaluation above and beyond the effect of relational qualifiers. *Social Cognition, 34.* 435-461.

2.5.3 Additional Analyses: Pilot Study 1

Recollective Memory Test. Seven participants in the start-stop instructions group, and 43 participants in the observe instructions group failed to correctly verbalise the contingency. This difference between groups seems due to the very strict criterion (100% correct report) applied and is not indicative of learning without awareness in the observe instructions group.

Explicit valence ratings. The pattern of EC scores for participants who passed the recollective memory test was similar to the entire sample. The only difference was that the significant Conditioning Type × Time interaction from the entire sample was marginal, F (1, 42) = 4.015, p = .052, η_p^2 = .087. Follow-up analyses revealed the same pattern as the entire sample.

Affective priming – Reaction Time. When analysing only those who passed the recollective memory test, the significant main effect of time became marginal, F (1, 39) = 3.511, p = .068, η_p^2 = .083. As per the entire sample, EC scores were larger at post-training than pre-training.

Affective priming – Errors. Mean EC scores for errors on the affective priming measure are depicted in Figure S5. The ANOVA revealed only a significant main effect of Time, F (1, 79) = 3.991, p = .049, η_p^2 = .048, indicating that EC scores were significantly larger at post-training than pre-training. Follow-up analyses revealed that priming scores were not different zero at pre-training, t (80) = 1.218, p = .227, d = 0.14, or at post-training, t (80) = 1.832, p = .071, d = 0.20. When analysing data from participants who passed the recollective memory test, the main effect of time was not significant, F (1, 39) = 0.262, p = .612, η_p^2 = .007.

2.5.4 Additional Analyses: Pilot Study 2

Recollective Memory Test. Nine participants in each group failed to correctly verbalise the contingency.

Explicit valence ratings. The pattern of results was the same as the entire sample when analysing those who passed the recollective memory test.

Affective priming – Reaction Time. When analysing only those who passed the recollective memory test, the main effect of conditioning type became significant, F (1, 71) = 8.784, p = .004, η_p^2 = .110. Follow-up analyses revealed the same pattern as the entire sample.

Affective priming – Errors. Figure S6 shows mean EC scores for errors on affective priming at post-test as a function of Conditioning Type and Pre-measure. A main effect of Conditioning Type suggests a greater priming score for forward conditioning than backward conditioning, F (1, 83) = 4.278, p = .042, η_p^2 = .049. However, one sample t-tests showed that priming scores for forward conditioning, t (84) = 1.302, p = .197, d = 0.14, and backward conditioning were not significantly different from zero, t (84) = 1.565, p = .121, d = 0.17.

Figure S7 suggests a contrast effect for backward CSs in the pre-measure group only when analysing participants who passed the recollective memory test. The Conditioning Type × Group interaction was significant, F (1, 71) = 5.513, p = .022, η_p^2 = .072. Follow-up analyses revealed that EC error scores were smaller for backward conditioning than forward conditioning for the pre-measure group only, t (36) = 2.842, p = .006, d = 0.47, with no differences for the no pre-measure group t (35) = 0.494, p = .623, d = 0.08. One-sample t-tests showed that backward conditioning EC error scores in the pre-measure group were significantly below zero, t (36) = 2.412, p = .021, d = 0.40. No other comparisons differed from zero, all ts < 1.346, all ps > .181, all ds < 0.22.

Analyses of pre-measure only group across Time. EC and priming scores from the 'Pre-measure' group were also subjected to a 2 (Conditioning Type: forward vs backward; within-participants) × 2 (Time: pre-test vs post-test; within-participants) repeated measures ANOVA.

Explicit valence ratings. Figure S8 shows an assimilation effect for forward conditioning and a contrast effect for backward conditioning at post-training. Main effects of Conditioning Type, F (1, 46) = 97.713, p < .001, ηp^2 = .680, and Time, F (1, 46) = 21.192, p < .001, ηp^2 = .315, were qualified by a Conditioning Type × Time interaction, F (1, 46) = 69.436, p < .001, ηp^2 = .602. Follow-up analyses revealed that EC scores for forward conditioning were significantly larger than EC scores for backward conditioning at post-training, F (1, 46) = 106.963, p < .001, ηp^2 = .699, but no difference occurred at pre-training, F (1, 46) = 0.028, p = .868, ηp^2 = .001. One-sample t-tests showed forward conditioning EC scores were significantly larger than 0, t (46) = 11.965, p < .001, d = 1.75, and backward conditioning EC scores were significantly less than 0, t (46) = 4.893, p < .001, d = 0.71, at post-training. Forward and backward conditioning EC scores did not differ from 0 at pre-training, t (46) = 0.133, p = .894, d = 0.02, and t (46) = 0.116, p = .908, d = 0.02, respectively. The

pattern of results was the same when only analysing those who passed the recollective memory test.

Affective priming – Reaction time. Figure S9 shows an assimilation effect for forward conditioning at post-training. The Conditioning Type × Time interaction was significant, F (1, 39) = 9.895, p = .003, ηp^2 = .202. Forward conditioning priming scores were larger than backward conditioning at post-training, F (1, 39) = 6.844, p = .013, ηp^2 = .149, but not pre-training, F (1, 39) = 0.902, p = .348, ηp^2 = .023. One sample t-tests showed that forward conditioning priming scores at post-training were significantly larger than 0, t (39) = 2.908, p = .006, d's < 0.46, and that no other scores differed from 0, t's < 1.562, p's > .126, d's < 0.25. The pattern of results from those who passed the recollective memory did not differ from the entire sample.

Affective priming – Errors. Figure S10 shows mean EC scores for errors on affective priming at post-test as a function of conditioning type and Pre-measure. A main effect of Conditioning Type suggests a larger priming score for forward conditioning than backward conditioning, F (1, 39) = 9.823, p = .003, η_p^2 = .201. A one sample t-test showed that the priming score for forward conditioning did not differ from zero, t (39) = 1.657, p = .105, d = 0.26. For backward conditioning, the priming score was significantly less than zero, t (39) = 2.726, p = .010, d = 0.43. The pattern of results from those who passed the recollective memory did not differ from the entire sample.

2.5.5 Additional Analyses: Main Experiment

Recollective Memory Test. Thirty two participants in the 'Start-Stop instruction, Mallan paradigm' group, 33 participants in the 'Start-Stop instruction, Moran paradigm' group, eight participants in the 'Valence-Agency instruction, Mallan paradigm' group, and seven participants in the 'Valence-Agency instruction, Moran paradigm' group failed to verbalise the contingency.

Explicit valence ratings. When analysing participants who passed the recollective memory test, the Instructions × Time interaction became significant, F (1, 106) = 5.451, p = .021, η_p^2 = .049, and the Time × Paradigm interaction became marginal, F (1, 106) = 3.538, p = .063, η_p^2 = .032. Overall, the pattern of results did

not change as the Instructions × Conditioning Type × Time interaction remained significant, F (1, 106) = 20.693, p < .001, η_p^2 = .163.

Affective priming – Reaction time. When analysing participants who passed the recollective memory test the main effect of Time became significant, F (1, 95) = 4.142, p = .045, η_p^2 = .042, revealing the same pattern of results as the entire sample.

Affective priming – Errors. Figure S11 shows mean EC scores on affective priming for forward and backward conditioning measured pre-training and post-training as a function of Instructions and Paradigm. The figure suggests a contrast effect at post-training in the valence-agency instructions/Mallan paradigm group for backward conditioning. An Instructions × Time interaction, F (1, 153) = 6.124, p = .014, η_p^2 = .038, and a Paradigm × Conditioning Type × Time interaction were found. Following up the Instruction × Time interaction revealed that priming scores in the observe instructions group were marginally larger at post-training than pre-training, t (39) = 1.696, p = .092, d = 0.27. In the valence-agency instructions group, priming scores did not differ between pre- and post-training, t (153) = 0.321, p = .749, d = 0.05. Following up the Paradigm × Conditioning Type × Time interaction revealed larger priming scores in the Moran paradigm group at post-training when compared with pre-training for backward conditioning, t (153) = 2.402, p = .018, d = 0.38. No other comparisons were significant, ts < 1.552, ps > .123, ds < 0.25. When analysing participants who passed the recollective memory test, no main effects or interactions were significant, Fs < 2.959, ps > .089, η_p^2s < .030.

2.5.5.1 Figures from Additional Analyses

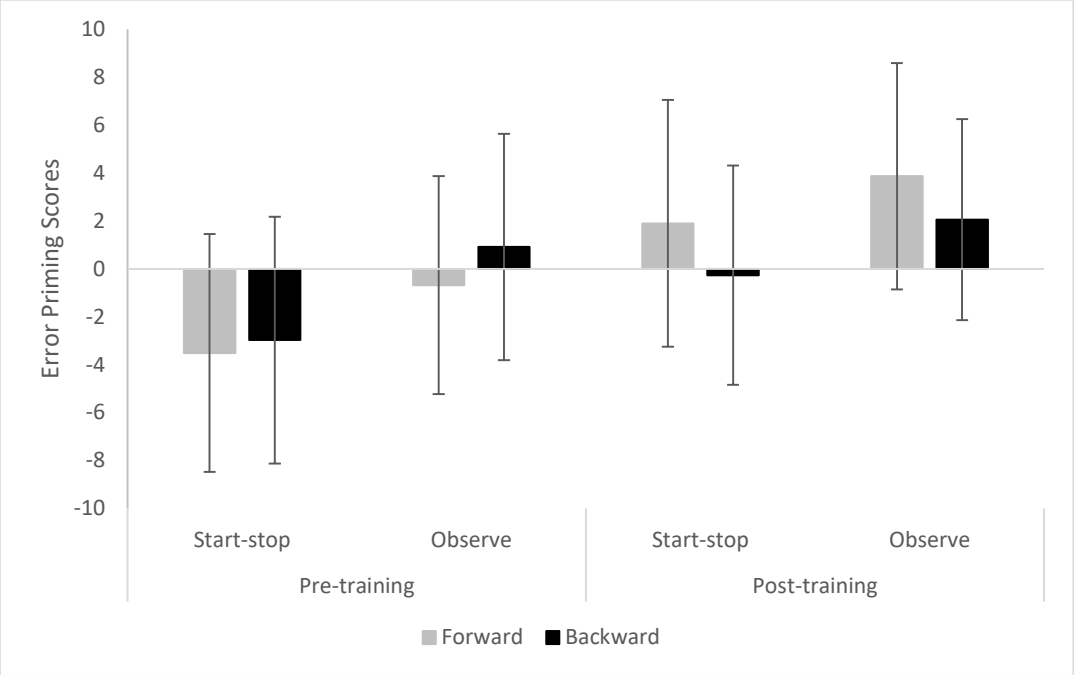

Figure S5. EC scores for errors on affective priming for forward and backward conditioning, measured pre-training and post-training for start-stop instructions and observe instructions, Pilot Study 1. Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.

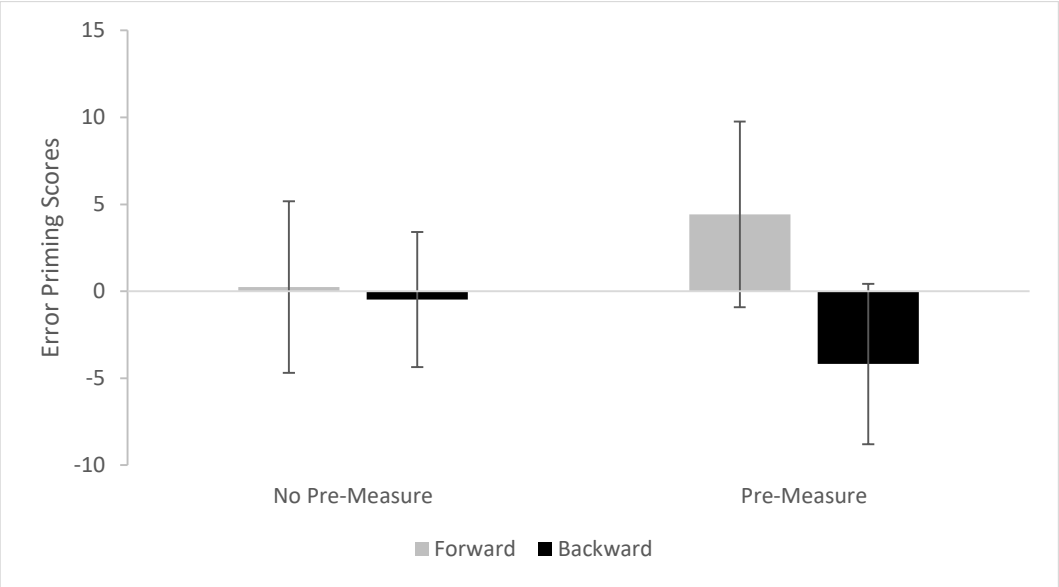

Figure S6. EC scores for errors on affective priming for forward and backward conditioning, measured post-training for no pre-measure and pre-measure groups, Pilot Study 2. Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.

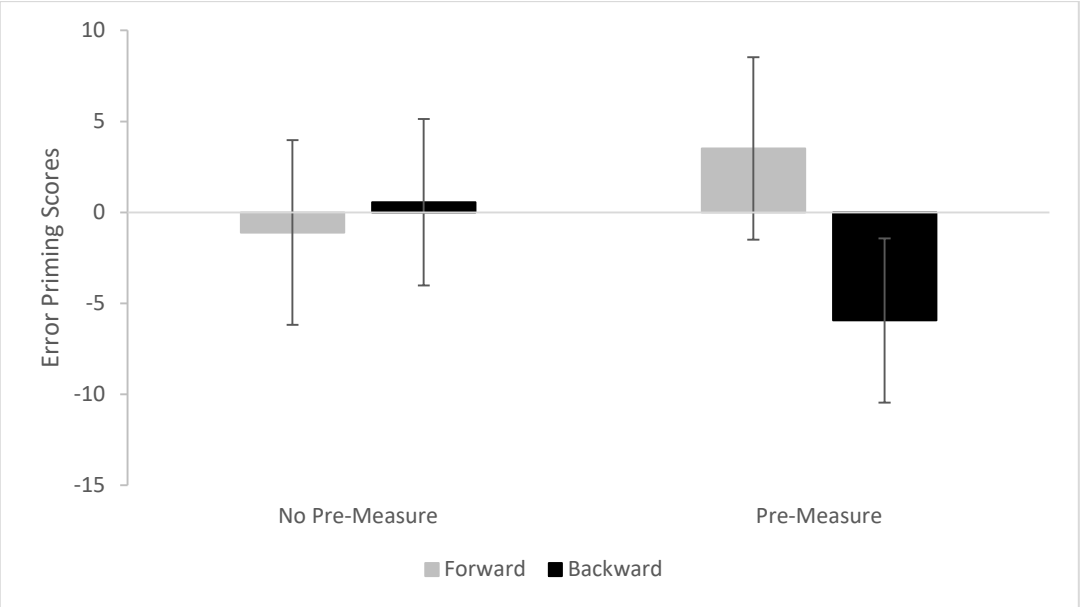

Figure S7. EC scores from participants who passed the recollective memory test for errors on affective priming for forward and backward conditioning, measured post-training for no pre-measure and premeasure groups, Pilot Study 2. Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.

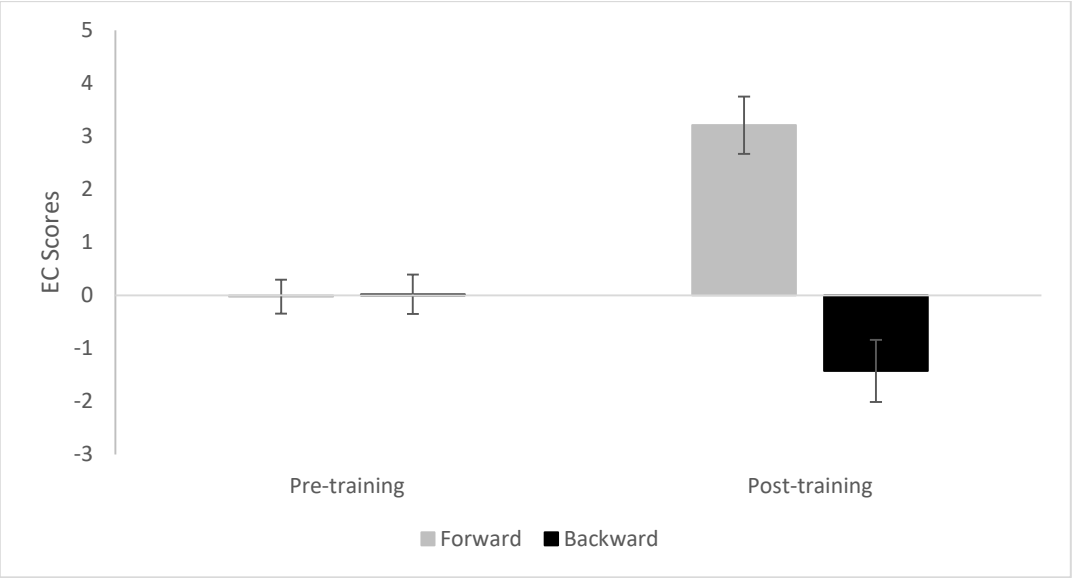

Figure S8. EC scores for forward and backward conditioning measured pre and post-training for the pre-measure group only, Pilot Study 2. Positive scores suggest assimilation effects, negative scores suggest contrast effects. Error bars show 95% confidence intervals of the mean.

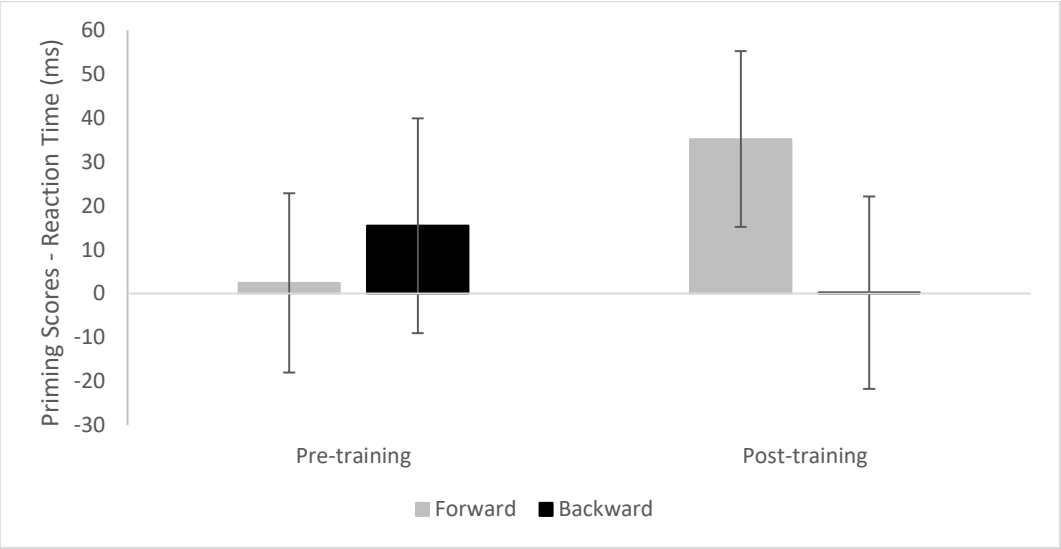

Figure S9. Priming scores from reaction times for forward and backward conditioning, measured pre and post-training for the pre-measure group only, Pilot Study 2. Positive scores suggest assimilation effects, negative scores suggest contrast effects. Error bars show 95% confidence intervals of the mean.

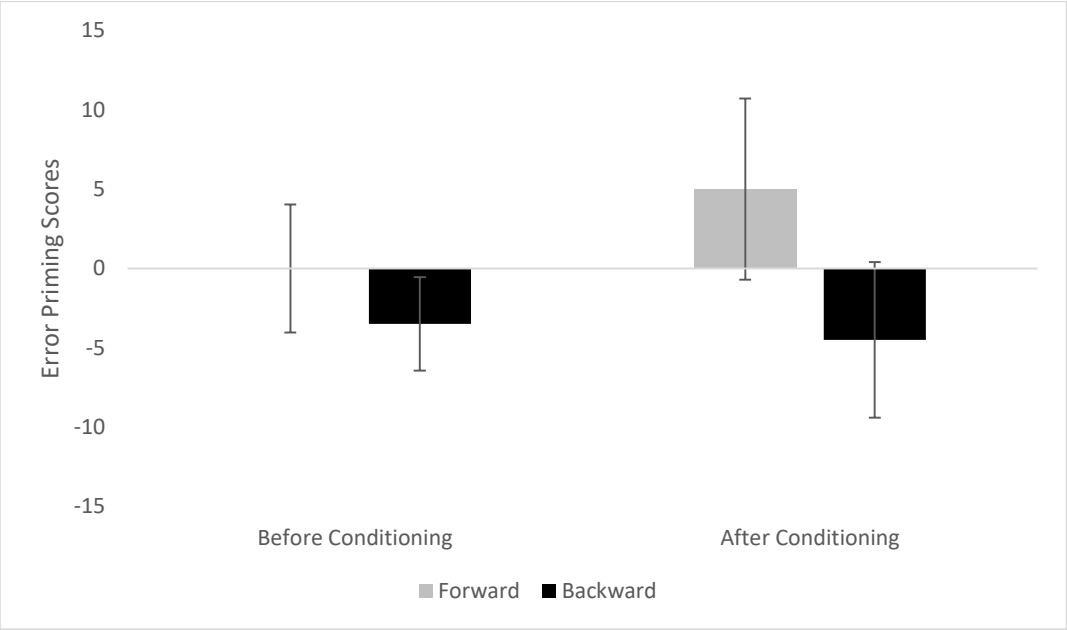

Figure S10. EC scores EC scores for errors on affective priming for forward and backward conditioning measured pre and post-training for the pre-measure group only, Pilot Study 2. Positive scores suggest assimilation effects, negative scores suggest contrast effects. Error bars show 95% confidence intervals of the mean.

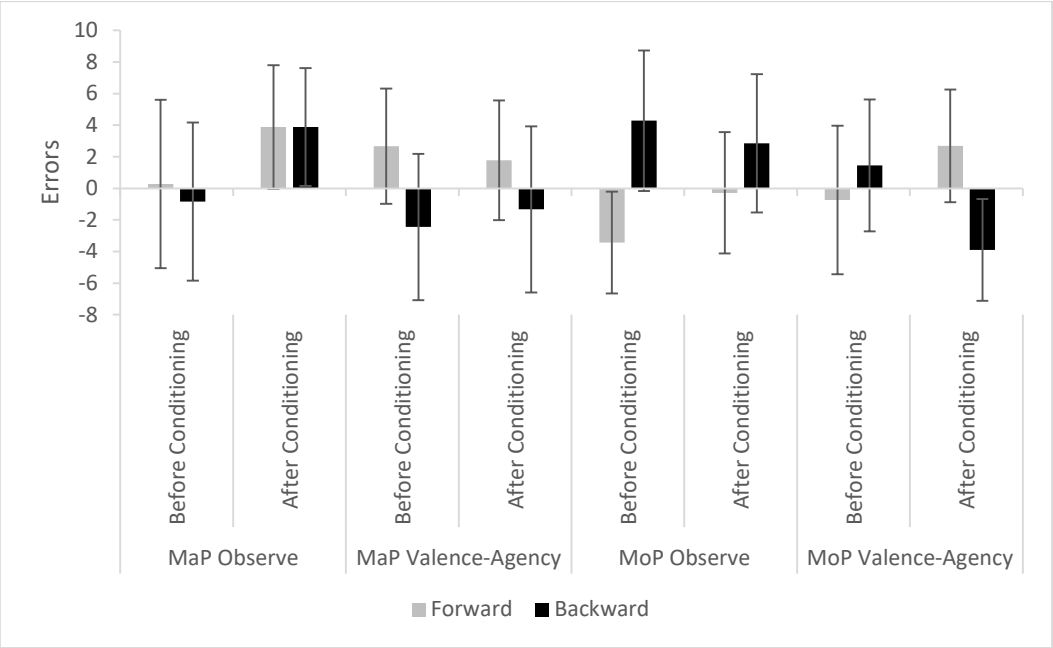

Figure S11. EC scores for errors on affective priming for forward and backward conditioning measured pre-training and post-training as a function of instructions ('observe instructions' and 'valence-agency instructions') and paradigm ('Mallan paradigm (MaP)' and 'Moran paradigm (MoP)'), Main Experiment. Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.

Chapter 3: How disappointing: Startle modulation reveals conditional stimuli presented after pleasant unconditional stimuli acquire negative valence

Luke J. S. Green
Curtin University

Camilla C. Luck
Curtin University

Ottmar V. Lipp
Curtin University

Author Notes

This work was supported by an Australian Government Research Training Program Scholarship to Luke Green and grants DP180111869 and SR120300015 from the Australian Research Council to Ottmar Lipp.

The authors have no conflicts of interest to declare.

Correspondence concerning this article should be sent to: Luke J S Green, School of Psychology, Curtin University, GPO Box U1987 Perth WA 6845, Australia. Email: luke.green2@postgrad.curtin.edu.au.

3.1 Abstract

Past research on backward conditioning in evaluative and fear conditioning yielded inconsistent results in that self-report measures suggest that conditional stimuli (CS) acquired the valence of the US in fear conditioning (assimilation effects), but the opposite valence in evaluative conditioning (contrast effects). Conversely, implicit measure of CS valence suggest assimilation effects in evaluative backward conditioning whereas startle modulation indicates contrast effects in backward fear conditioning. Experiment 1 investigated whether US intensity could account for the dissociation on implicit measures between fear and evaluative conditioning. Self-report measures of evaluative learning indicated assimilation effects for forward conditioning, whereas backward contrast effects were observed with intense USs only. Blink startle modulation indicated assimilation effects in forward conditioning and contrast effects in backward conditioning, regardless of US intensity. Experiment 2 included a neutral US in order to assess whether the offset of the positive US elicits an opponent emotional response that mirrors relief (disappointment), which is thought to mediate the reduction in startle seen during backward CSs in fear conditioning. This opponent emotional response was evident as startle magnitude during backward CSs increased linearly with increasing US pleasantness. Omission of the forward CSs led to an assimilation effect in self-report measures. The current results extend our understanding of emotional learning to stimuli encountered after salient emotional events. Startle reflects the emotion prevailing after US offset, relief or disappointment, whereas self-report measures seem more attuned to factors such as US predictability and intensity.

Keywords: Evaluative conditioning, associative learning, backward conditioning, startle modulation, propositional learning

3.2 Introduction

Evaluations, i.e. how positive or negative a stimulus or event is, can influence many aspects of our lives, including career choice, voting and consumer behaviour, and our relationships with other people (see Galdi, Arcuri, & Gawronski, 2008; Gibson, 2008; LeBel & Campbell, 2009). These evaluations can be manipulated through *evaluative conditioning* (De Houwer, Thomas, & Baeyens, 2001; Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010), a process in which the valence of a neutral conditional stimulus (CS) can be changed by pairing it with a positively or negatively valenced unconditional stimulus (US; De Houwer, 2007). Advertising campaigns often exploit EC by, for instance, presenting a product (the CS) with a popular celebrity (the US), resulting in positive evaluations of the product. As we all continually encounter stimuli of differing valence that co-occur in different temporal and spatial arrangements, the study of evaluative conditioning has immense importance as it is relevant for many facets of psychology and life in general.

Evaluative conditioning can be studied in the laboratory using a picture-picture paradigm. In this paradigm, neutral pictures (CS) are paired with positive or negative pictures (USs), which results in CSs paired with positive USs becoming more pleasant, and CSs paired with negative USs becoming more unpleasant (Hofmann et al., 2010; Levey & Martin, 1975; Mallan, Lipp, & Libera, 2008). These changes in CS valence can be tracked with explicit, i.e., self-report, and implicit measures, i.e., affective priming or Implicit Association Tests (Fazio & Olson, 2003). A similar evaluative change occurs in differential fear conditioning, as pairing a neutral picture (CS+) with an aversive electro-tactile stimulus (US) results in explicit negative evaluations of the CS+ in comparison to a neutral picture that was not paired with the US (CS-; Lipp, 2006). Moreover, negative CS+ valence acquired during differential fear conditioning can be measured implicitly using the startle blink reflex, as startle blink magnitude is larger during negative stimuli and smaller during positive stimuli when compared to neutral stimuli (Bradley, Cuthbert, & Lang, 1990; Vrana, Spence, & Lang, 1988; but see Lipp, Siddle, & Dall, 2003).

Changes in stimulus evaluation such that the CS acquires the valence of the US are known as assimilation effects. Assimilation effects have been demonstrated in evaluative and fear conditioning utilising forward conditioning procedures (CS-US), and in evaluative conditioning utilising simultaneous (CS+US) and backward

conditioning procedures (US-CS; Mallan et al., 2008; Hoffman et al., 2010). Assimilation effects have been shown using explicit valence ratings, the startle blink reflex, and reaction time based implicit measures of CS valence (Mallan et al., 2008; Olson & Fazio, 2001). However, contrast effects, the CS acquiring valence that is opposite to that of the US, have also been observed. Contrast effects have been shown on explicit valence ratings in evaluative conditioning employing instructions that emphasise CS agency (start/stop) after both forward and backward CS and US pairings (Hu, Gawronski, & Balas, 2017a; Moran & Bar-Anan, 2013; Moran, Bar-Anan, & Nosek, 2016; Unkelbach & Fiedler, 2016). They have also been observed for backward conditioning on the startle blink reflex and explicit valence ratings in fear conditioning (Andreatta, Mühlberger, Yarali, Gerber, & Pauli, 2010; Andreatta, Mühlberger, Glotzbach-Schoon, & Pauli, 2013; Luck & Lipp, 2017; see also Mühlberger et al., 2011 for evidence of contrast effects during videos of faces changing from neutral to happy and angry and vice versa).

3.2.1 Contrast Effects in Evaluative Conditioning

Moran and Bar-Anan (2013) demonstrated contrast effects for backward CSs in a concurrent forward and backward evaluative conditioning procedure. In these within-subjects experiments, a pleasant melody (positive US) and an unpleasant human scream (negative US) were paired with CSs drawn from four different families of alien creatures. On positive trials, one CS (forward CSpos; F-CSpos) was presented before the pleasant melody (positive US) and a second CS (backward CSpos; B-CSpos) was presented after the pleasant melody. On negative trials, a third CS (forward CSneg; F-CSneg) preceded the unpleasant human scream (negative US) and a fourth CS (backward CSneg; B-CSneg) followed the unpleasant human scream. Before conditioning, participants were informed that each CS family had a different role to play; that one would start the positive US, one would stop the positive US, one would start the negative US, and one would stop the negative US, and that they needed to learn the role of each family for a later memory test. When assessing CS valence with explicit (ratings) and implicit measures (response time based tasks), a dissociation emerged. On implicit measures, assimilation effects were demonstrated for forward and backward conditioning, as CSs paired with the positive US were evaluated more positively than CSs paired with the negative US. On explicit measures, an assimilation effect was demonstrated for forward conditioning only; CSs preceding the positive US (F-CSpos) were evaluated as more pleasant than

CSs preceding the negative US (F-CSneg). For backward conditioning, however, a contrast effect was demonstrated; CSs following the positive US (B-CSpos) were rated as more negative than CSs following the negative US (B-CSneg). This contrast effect has also been replicated using picture USs but more explicit instructions highlighting the agency of the CSs in starting and stopping 'happy' and 'sad' events were required to obtain this result (Green, Luck, Gawronski, & Lipp, 2019; Moran et al., 2016).

3.2.2 Contrast Effects in Fear Conditioning (Relief Learning)

In fear conditioning, assimilation effects are found on explicit valence ratings and the startle blink reflex after forward conditioning, while backward conditioning leads to assimilation effects on explicit measures and contrast effects for the startle blink reflex (Andreatta et al., 2010; Andreatta, Mühlberger, & Pauli, 2016). This dissociation between explicit and implicit measures of CS valence in backward fear conditioning was initially demonstrated by Andreatta et al. (2010). Pictures of geometric shapes were presented either alone (CS-) or paired with an aversive electro-tactile stimulus (CS+) in a forward conditioning group (CS+-US/CS-), a backward conditioning group (US-6s gap-CS+/CS-), and a control group (CS+-6s gap-US/CS-). The CS+ was rated more negatively after forward and backward conditioning than before conditioning. Compared with the mean of all responses, startle blink magnitude elicited during CS+ was larger in the forward conditioning group (suggesting negative valence) and smaller during the CS+ in the backward conditioning group (suggesting positive valence). Moreover, startle blink magnitude during the CS+ in the backward conditioning group was smaller than during the CS- (suggesting the CS+ had acquired positive valence relative to the safety signal). These startle results are due to the 'relief' experienced at the offset of the aversive electro-tactile stimulus being conditioned to the backward CS (relief learning), and have been replicated using different timings between US offset and CS onset (Andreatta, et al., 2016; Luck & Lipp, 2017; see also Gerber et al., 2014 and Deutsch, Smith, Kordts-Freudinger, & Reichardt, 2015 for reviews on relief learning).

3.2.3 Explaining Opposite Patterns of Dissociations

Moran and Bar-Anan (2013) and Andreatta et al. (2010) found different patterns of dissociations between their explicit and respective implicit measures for

backward conditioning. Moran and Bar-Anan (2013) found contrast effects on explicit valence ratings, while Andreatta et al. (2010) found an assimilation effect. On the other hand, Moran and Bar-Anan (2013) found assimilation effects on implicit measures of CS valence, while Andreatta et al. (2010) found a contrast effect on the startle blink reflex. There are several differences between these studies that could explain the contrasting patterns of dissociations, such as using different task instructions, presenting forward and backward conditioning within-subjects concurrently instead of comparing forward and backward conditioning between-subjects separately, presenting CSs and USs for different durations and with different inter-stimulus intervals, and using USs of differing intensities.

The different pattern of explicit valence ratings reported by Andreatta et al. (2010) and Moran and Bar-Anan (2013) can be explained by the task instructions. Green, et al. (2019) showed that presenting relational information highlighting the role of the CSs in a within-subjects concurrent forward and backward conditioning procedure produces contrast effects, and that without this information assimilation effects emerge. The different pattern on implicit measures, however, remains unexplained. The different procedures used (presenting forward and backward conditioning within-subjects as compared to between-subjects) cannot account for this as Andreatta and Pauli (2017) using a within-subjects procedure found the same results as Andreatta et al. (2010) using a between-subjects procedure. The differences in CS and US presentation duration and different inter-stimulus intervals cannot explain the difference either. Luck and Lipp (2017) found the same pattern as Andreatta et al. (2010) when using a 100ms gap instead of a 6s gap between US offset and backward CS onset. Moreover, Green et al. (2019) found that small paradigmatic differences such as CS and US duration, using multiple or single CSs and USs, and overlapping the CS and US presentations, do not influence backward evaluative conditioning on explicit or implicit measures in a within-subjects concurrent forward and backward conditioning procedure. Differences in US intensity on the other hand may explain the different patterns of results on implicit measures, as the shock US used by Andreatta et al. (2010) is arguably more unpleasant than the auditory USs used by Moran and Bar-Anan (2013)[1]. US intensity is likely to affect relief learning as Bitar, Marchard, and Potvin (2018) found that

[1] Moran and Bar-Anan (2013) did not report the intensity to which their sound USs were set, but attempts to recreate them following their description suggest that they were below 90dBA.

pain relief was positively correlated with pain level, such that higher levels of pain led to greater pain relief (see also Leknes, Brooks, Wiech, & Tracey, 2008). Moreover, stronger learning tends to occur with stronger USs (Annau & Kamin, 1961; Pavlov, 1927). Therefore, assuming that implicit measures are less sensitive than explicit measures to contrast effects, then the lower intensity USs used in Moran and Bar-Anan (2013) may not have been sufficient to drive these effects, in contrast to the higher intensity USs used in Andreatta et al. (2010).

In Experiment 1, we aimed to use Moran and Bar-Anan's (2013) procedure to replicate their backward CS contrast effects on explicit ratings, while, measuring the startle blink reflex and manipulating US intensity between groups in a 2 (Group: low vs high intensity; between-participants) × 2 (Conditioning Type: forward vs backward; within-participants) × 2 (US Valence: positive vs negative; within-participants) mixed design. We hypothesised that contrast effects on explicit valence ratings would emerge for backward CSs in both groups, with a larger effect in the high intensity group. As Andreatta et al. (2010) demonstrated that startle blink magnitude was inhibited during a CS following a shock US (suggesting positive valence), we hypothesised that the startle blink reflex would show backward CS contrast effects for both groups, with a larger difference in the high intensity group. Moreover, we hypothesised that US intensity would influence the pattern of responding to backward CSs on an implicit behavioural measure, such that assimilation effects would occur in the low intensity group, and contrast effects would occur in the high intensity group. Finally, assimilation effects were expected on all measures for forward CSs[2].

3.3 Experiment 1

3.3.1 Method

Participants. Following ethical approval for this research protocol from the Curtin University Human Research Ethics Committee, 66 undergraduate students from the School of Psychology at Curtin University participated in exchange for course credit. Two participants were excluded for having participated in an earlier study employing the same conditioning task. The final sample comprised 64 students (50 female), *M* age = 21.63, *SD* = 6.46, with 32 participants per group. The sample

[2] All materials, data, analysis files, and supplementary materials, are available at https://osf.io/q46mp/.

size was based on Andreatta et al. (2013), who employed 28 participants and found an effect size of $\eta p^2 = 0.316$ for the within-subjects comparison of conditioning type. Moreover, Andreatta et al. (2010) employed 33-34 participants per group and found a significant Conditioning Type × CS interaction. Based on this, we determined that 32 participants per group would provide sufficient power to detect the effects we were interested in. Five participants in the low intensity group and four in the high intensity group failed the recollective memory test. To pass this test participants needed to correctly identify the role of each of the four CSs with 100% accuracy. Analyses were run with and without these participants. Results are reported for the entire sample with those from participants who passed the recollective memory test added only if they provide additional clarification.

Apparatus/Stimuli. Four families of four aliens from Moran and Bar-Anan (2013) were used as CSs (examples shown in Figure 1). The CS families differed in head shape and colour. Also taken from Moran and Bar-Anan (2013), the positive US was a pleasant guitar melody (the start of 'The Shape of My Heart' by Sting), and the negative US was a human scream (all stimuli can be found at https://osf.io/cqsnj/). A pilot study (n = 20) was performed to match the perceived intensities of the USs, which were manipulated by altering the volume (see Table 1 for dBA values of USs used in the low and high intensity groups). US intensity ratings were submitted to a 2 × 2 repeated measures ANOVA revealing only main effects for valence, F (1, 17) = 15.02, $p < .001$, $\eta_p^2 = .442$, and intensity, F (3, 17) = 30.37, $p < .001$, $\eta_p^2 = .843$. The perceived intensity of the positive and negative USs used within each group was comparable, however, the dynamics of each US differed which led to different dBA readings (pilot study reported at https://osf.io/q46mp/).

Orbicularis Oculi electromyogram (EMG), skin conductance, and respiration were recorded using a Biopac MP150 system with AcqKnowledge Version 4.1 at a sampling rate of 1000Hz. Orbicularis Oculi EMG was measured using two 4 mm Ag/AgCl electrodes filled with electrode gel and attached using double-sided adhesive electrode collars. The first electrode was placed directly under the pupil of the left eye, and the second under the corner of the left eye. Impedance was assessed to confirm electrode contact, though no threshold criterion was employed. A custom built noise-generator was used to present a 105dBA white noise burst lasting 50ms with a near instantaneous rise time as the startle eliciting stimulus. The EMG signal was amplified by a Biopac EMG100C amplifier at a gain of 5000 and high and low

pass filtered at 10 and 500 Hz. Electrodermal responding was measured using two self-adhesive isotonic Biopac EL507 electrodes attached to the thenar and hypothenar eminences of the non-dominant hand. A Biopac EDA100C amplifier was used to DC amplify responses at a gain of 5 µSiemens per volt. A chest gauge was used to measure respiration to control for respiration or movement related artefacts in electrodermal responding. For the affective priming task, the CSs were used as primes and 10 positive words (*pleasant, good, outstanding, beautiful, magnificent, marvellous, excellent, appealing, delightful,* and *nice*) and 10 negative words (*unpleasant, bad, horrible, miserable, hideous, dreadful, painful, repulsive, awful,* and *ugly*) taken from Hu et al. (2017a; Hu, Gawronski, & Balas, 2017b) served as target stimuli. DMDX (Forster & Forster, 2003) was used to control stimulus presentations and markers and to present and record responses from the explicit valence ratings task, the affective priming task, and the memory test. Sennheiser HD-25-1 headphones were used to present auditory USs and startle probes.

Procedure. Participants read the information sheet and were played each US from their condition for 30 seconds before providing informed consent. After signing the consent form, participants washed and dried the area under their left eye and their hands. The recording equipment was attached, and three habituation startles were presented (timing controlled manually by the experimenter to avoid coinciding with deliberate blinks, laughter, and fidgeting etc.) followed by a three minute baseline recording of skin conductance. The startles probes during habituation were not controlled by the software to be presented as specific time intervals. The experimenter pressed the shift key to present startle probes to the participant after they had recovered from the previous probe (i.e. no laughter, fidgeting, closing the eyes etc., and the EMG recording had returned to baseline). The researcher then told participants to learn which family of aliens started and stopped the positive and negative sounds and that they would be tested on this at the end. The researcher started the script and the instructions were presented again on the screen. Participants were then presented with the CS-US-CS procedure comprising 10 positive US and 10 negative US trials (see Figure 1 for a depiction of a positive US trial). CSs were presented for 8s each, with a 2s overlap between CS-US and US-CS. Two CSs from each set of CSs were presented 3 times and 2 were presented twice, totalling 10 trials per set (10 F-CSpos, 10 B-CSpos, 10 F-CSneg, and 10 B-CSneg). USs were presented for 10, 15, 20, 25, or 30s. Startle probes were presented at 4.5 or 5.5s after

forward CS onset, and 6.5 and 7.5s after backward CS onset (i.e., 4.5 or 5.5s after US offset; see Figure 2 for a depiction of startle probe timing). Startle probes were presented on six trials for each CS set, for a total of 24 probes. These probes were assigned randomly within forward and backward CSs separately. The inter-trial intervals were 12, 14, or 16s (randomly dispersed) and startle probes were presented half-way through half of the inter-trial intervals for a total of 10 startle probes. After conditioning, the experimenter informed participants they would now be asked to rate how much they liked each family of aliens. Participants were shown each set of CSs (four CSs per set) separately and asked to provide a rating of how much they liked each family on a scale from 1 = don't like at all, to 9 = like a lot. After providing ratings, the experimenter explained the affective priming task to the participants. The affective priming task comprised 80 trials where each set of CSs were presented with 10 positive words and 10 negative words. Two CSs from each set were each presented with all positive words once, while the other two CSs were presented with all negative words once. During a trial, a fixation cross was presented for 500ms, followed by the CS prime for 200ms, and then the target word until the participant responded by pressing the right 'SHIFT' key if the target word was positive, and the left 'SHIFT' key if the target word was negative. After a 20 trial affective priming practice task, the experimenter told participants they would now do the main affective priming task. Following the main affective priming task, the experimenter told participants it was time for the memory test. Participants were shown each family separately, and asked:

> *What was the role of the creatures in this picture? 1. Started the human sound. 2. Stopped the human sound. 3. Started the musical sound. 4. Stopped the musical sound.*

After the memory test, participants were told to pay attention to the screen and follow any instructions that appeared. Participants were told 'the experiment will now continue', and an extinction phase was presented. During extinction, each member of each CS set was presented twice for a total of 32 trials. Startles were presented at 4.5 or 5.5s after CS onset on six of the eight presentations of each CS set, totalling 24 startle probes. The inter-trial intervals were 12, 14, or 16s, and startle probes were presented half-way through the interval on half of the trials for a total of 16 startle probes. After this, participants were disconnected from the recording equipment and asked to fill out a demographics questionnaire. Participants were

asked their age, gender, and ethnicity, as well as how pleasant/unpleasant the human sound, musical sound, and loud noises were on a 7-point scale ranging from -3 = very unpleasant to 3 = very pleasant. Participants were also asked how intense the human and musical sounds were on a 7-point scale from 0 = not at all to 6 = very intense, and how startling the loud noises (startle probes) were on a 7-point scale from 0 = not at all to 6 = very startling. Participants were then debriefed, and thanked for their time. The entire experiment took approximately 1 hour.

Scoring, response definition, and statistical analyses. Data were analysed using mixed-model ANOVAs in IBM SPSS Statistics 25. Significant interactions ($\alpha = .05$, Pillai's trace statistics of the multivariate solution reported) were followed-up with pairwise comparisons. The current report is focussed on explicit valence ratings, startle magnitude during acquisition, and reaction times from the affective priming task. Affective priming error data, startle blink latency data from acquisition and extinction, and startle blink magnitude data from extinction and electrodermal responses to forward CS and US onset were analysed, and results are included in the supplementary materials at https://osf.io/q46mp/. SCRs to backward CSs and SCRs during extinction were not analysed.

Startle blink magnitude. The raw EMG signal was notched at 50 Hz, high and low pass filtered at 30 and 500 Hz, and rectified and smoothed by using 5 consecutive measurement points to calculate a moving average. Startle blink magnitude was defined as the largest response that occurred within 120ms of the startle probe, beginning 20-60ms after startle probe onset (Blumenthal et al., 2005). A non-response trial was defined as a trial where response onset could not be visually identified within this window. Non-response trials were scored as zeros and included in the analysis. A trial was defined as missing if the response could not be visually differentiated from background EMG activity, or if a blink occurred between the startle probe onset and the response window onset (Experiment 1 = 2.73% and Experiment 2 = 1.69%). Individual differences and variation across individual trials were controlled for by blocking trials and transforming raw data into *T*-scores. Blocks of two trials were created for each US valence for forward and backward conditioning separately, resulting in three blocks per condition. *T*-scores were then subjected to a 2 (Group: low vs high intensity; between-participants) × 2 (Conditioning Type: forward vs backward; within-participants) × 2 (US valence: positive vs negative; within-participants) × 3 (Block: 1, 2, 3; within-participants)

mixed model ANOVA. One participant from each group was excluded for failing to respond to more than 50% of the startle probes resulting in 62 participants being included in the startle blink magnitude analyses.

Explicit valence ratings. Participants rated each family of CSs on how much they liked them. The higher the rating, the more positive the valence of the family. These data were subjected to a 2 (Group: low vs high intensity; between-participants) × 2 (Conditioning Type: forward vs backward; within-participants) × 2 (US Valence: positive vs negative; within-participants) mixed model ANOVA.

Affective priming. Participants categorised positive or negative target words following the presentation of the CS primes. Incorrect categorisation of target words were scored as errors. Responses faster than 300ms and slower than 1000ms were also scored as errors, as they were deemed to be outside the window of a response suggestive of task adherence (Koppehele-Gossel, Hoffmann, Banse, & Gawronski, 2020). Participants who made more than 25% errors were removed from the analyses, leaving a total of 59 participants (low intensity group, $n = 29$). Reaction times to each target word following CSs from the same set were averaged to provide mean reaction times, resulting in means for each CS set for positive target words and each CS set for negative target words. Percentage of errors was also calculated for each set for each target word. These means were then used to calculate priming scores (incongruent trials [CSs paired with positive USs/negative target words + CSs paired with negative USs/positive target words] – congruent trials [CSs paired with positive USs/positive target words + CSs paired with negative USs/negative target words]), which were subjected to separate 2 (Group: low vs high intensity; between-participants) × 2 (Conditioning Type: forward vs backward; within-participants) mixed model ANOVAs.

Manipulation checks. Groups did not differ on gender, $\chi 2$ (1, $N = 64$) = < .001, $p > .999$, ethnicity, $\chi 2$ (4, $N = 64$) = 2.1, $p = .718$, or age, t (62) = 0.88, $p = .378$, $d = 0.23$. Post-experimental valence and intensity ratings of the USs and startle probe were subjected to separate 2 (Group: low vs high intensity; between-participants) × 3 (Valence: positive US vs negative US vs startle probe; within-participants) mixed model ANOVAs. For the valence ratings, a main effect of US valence, F (2, 61) = 586.63, $p < .001$, $\eta_p^2 = .95$, showed that the positive US was rated more positively than the negative US, ($M = 2.30$, $SD = 0.63$ vs $M = -2.20$, $SD = 0.80$), t (62) = 33.83, $p < .001$, $d = 4.23$, and the startle probe, ($M = 2.30$, $SD = 0.63$

vs *M* = -1.42, *SD* = 1.11), *t* (62) = 22.40, *p* < .001, *d* = 2.80, and that the startle probe was rated more positively than the negative US, (*M* = -1.42, *SD* = 1.11 vs *M* = -2.20, *SD* = 0.80), *t* (62) = 5.10, *p* < .001, *d* = 0.64. For the intensity ratings, a main effect of US valence, *F* (2, 61) = 53.14, *p* < .001, η_p^2 = .635, was qualified by a Group × Valence interaction, *F* (2, 61) = 3.27, *p* = .045, η_p^2 = .097, suggesting that the negative US was more intense in the high intensity group than the low intensity group, (*M* = 4.72, *SD* = 1.02 vs *M* = 3.72, *SD* = 1.37), *t* (62) = 3.30, *p* = .002, *d* = 0.84. In the low intensity group, the positive US, (*M* = 1.62, *SD* = 1.52), was rated as less intense than the negative US, (*M* = 3.72, *SD* = 1.37), *t* (61) = 6.20, *p* < .001, *d* = 1.10, and the startle probe, (*M* = 4.09, *SD* = 1.28), *t* (61) = 6.75, *p* < .001, *d* = 1.19, while, the negative US and startle probe did not differ, (*M* = 3.72, *SD* = 1.37 vs *M* = 4.09, *SD* = 1.28), *t* (61) = 1.49, *p* = .141, *d* = 0.26. In the high intensity group, the negative US, (*M* = 4.72, *SD* = 1.02), was rated as more intense than the startle probe, (*M* = 4.22, *SD* = 1.43), *t* (61) = 1.98, *p* = .051, *d* = 0.35, and the positive US, (*M* = 1.97, *SD* = 1.45), *t* (61) = 8.14, *p* < .001, *d* = 1.44, and the startle probe was rated as more intense than the positive US, (*M* = 4.22, *SD* = 1.43 vs *M* = 1.97, *SD* = 1.45), *t* (61) = 6.15, *p* < .001, *d* = 1.09. The startle magnitude during the inter-trial intervals did not differ between groups for acquisition, *t* (60) = 1.27, *p* = .209, *d* = 0.32.

3.3.2 Results

Startle blink magnitude – Acquisition. Figure 3 suggests larger startle responses during CSs presented before negative USs compared to CSs presented before positive USs in both groups, suggesting an assimilation effect. Startle responses during backward CSs following positive USs appear larger compared to responses during backward CSs following negative USs, suggesting a contrast effect. Main effects of conditioning type, *F* (1, 60) = 158.93, *p* < .001, η_p^2 = .726, and block, *F* (2, 59) = 120.33, *p* < .001, η_p^2 = .803, were qualified by a Conditioning Type × Block interaction, *F* (2, 59) = 14.59, *p* < .001, η_p^2 = .331, a Group × Conditioning Type interaction, *F* (1, 60) = 22.29, *p* < .001, η_p^2 = .271, and a Conditioning Type × US Valence interaction, *F* (1, 60) = 10.96, *p* = .002, η_p^2 = .154. As all follow-up analyses for the Conditioning Type × Block interaction were significant, difference scores between blocks 1 and 2, 2 and 3, and 1 and 3, were calculated for forward and backward conditioning, and subjected to paired sample *t*-tests comparing forward and backward conditioning for each difference score. Forward conditioning showed a larger difference than backward conditioning between blocks 1 and 2, *t* (61) = 5.20, *p*

< .001, d = 0.66, and blocks 1 and 3, t (61) = 4.91, p < .001, d = 0.62, with no difference appearing between blocks 2 and 3, t (61) = 0.53, p = .60, d = 0.07. The Group × Conditioning Type interaction showed there was no difference in startle blink magnitude between the low intensity and high intensity groups during forward CSs, F (1, 60) = 0.88, p = .352, η_p^2 = .014, and that startle blink magnitude was smaller in the high intensity group, than the low intensity group, F (1, 60) = 37.73, p < .001, η_p^2 = .386, during backward CSs, indicative of greater relief at US offset following high intensity stimuli. The Conditioning Type × US Valence interaction showed blink magnitude was larger during positive backward CSs than negative backward CSs, F (1, 60) = 10.56, p = .002, η_p^2 = .150, which is indicative of a contrast effect. No differences between valence for forward conditioning was found, F (1, 60) = 2.36, p = .129, η_p^2 = .038. However, when only participants who passed the memory test were included in the analysis, startle blink magnitude during negative forward CSs was larger than during positive forward CSs, F (1, 51) = 4.81, p = .033, η_p^2 = .086.

Explicit valence ratings. Figure 4 suggests assimilation effects for forward conditioning as CSs paired with positive USs are rated as more pleasant than CSs paired with negative USs. Contrast effects seem to be present for backward CSs for both groups, as CSs paired with positive USs are rated as less pleasant than CSs paired with negative USs. Main effects of conditioning type, F (1, 62) = 27.13, p < .001, η_p^2 = .304, and US valence, F (1, 62) = 130.41, p < .001, η_p^2 = .678, and a Conditioning Type × US Valence interaction, F (1, 62) = 212.47, p < .001, η_p^2 = .774, were qualified by a Group × Conditioning Type × US Valence interaction, F (1, 62) = 8.17, p = .006, η_p^2 = .116. Follow up analyses revealed a contrast effect in the high intensity group as backward CSs paired with positive USs were rated as less pleasant than backward CSs paired with negative USs, F (1, 62) = 29.59, p < .001, η_p^2 = .323, but not in the low intensity group, F (1, 62) = 0.80, p = .375, η_p^2 = .013. Moreover, assimilation effects were found for forward conditioning in both groups, as CSs paired with positive USs were rated as more positive than CSs paired with negative USs; low intensity: F (1, 62) = 141.10, p < .001, η_p^2 = .695, high intensity: F (1, 62) = 188.84, p < .001, η_p^2 = .753.

Affective priming – Reaction times. As shown in Figure 5, assimilation effects are suggested for forward conditioning, regardless of group. A main effect of conditioning type showed that the priming score for forward conditioning was

significantly larger than that for backward conditioning, F (1, 59) = 8.48, p = .005, η_p^2 = .126. Moreover, the forward conditioning priming score was significantly larger than 0, t (60) = 5.10, p < .001, d = 0.64, while the backward conditioning priming score was not, t (60) = 1.84, p = .070, d = 0.24.

3.3.3 Discussion

Experiment 1 aimed to replicate the backward CS contrast effects shown in Moran and Bar-Anan (2013) on explicit valence ratings while measuring the startle blink reflex and to determine whether US intensity could account for the differing dissociations between implicit and explicit measures reported by Moran and Bar-Anan (2013) and Andreatta et al. (2010). Blink magnitude data revealed a contrast effect for backward CSs in both groups, and an assimilation effect for forward CSs which was significant only in participants who recalled the contingencies. Startle blink responses were smaller during backward CSs in the high intensity group than the low intensity group. This suggests greater relief at the offset of the high intensity USs compared to the low intensity USs, which was expected only for negative USs. Increasing the volume of the USs may have resulted in both USs becoming less pleasant, rendering the positive US slightly unpleasant. Furthermore, this would explain the lack of differentiation between forward CSs paired with positive and negative USs, as the positive US becoming slightly more negative in the high intensity group would wash out any effects of differential US valence.

We found backward CS contrast effects on explicit valence ratings for the high intensity group only and assimilation effects for forward CSs in both groups. Thus, the intensity manipulation functioned as expected, resulting in larger backward CS contrast effects in the high intensity group than the low intensity group. However, unexpectedly, the backward CS contrast effect in the low intensity group was not significant. This may be due to the fact that our low intensity USs were less intense as those used in Moran and Bar-Anan (2013), which supports the idea that US intensity does in fact moderate backward CS contrast effects. It is also possible that presenting startle probes during acquisition made the USs seem less intense, therefore requiring more intense USs to produce a contrast effect. This explanation is supported by the fact that a backward CS contrast effect occurred in the high intensity group, as startle probes were also present during acquisition in this group.

Priming scores provide support for assimilation effects for forward CSs regardless of group, indicating a more negative evaluation of CSs preceding negative

USs. Priming scores for backward CSs did not reveal any acquisition of differential valence as a function of US valence. Thus, there was no support for the hypothesis that for backward conditioned CSs contrast effects would appear in the high intensity condition and assimilation effects in the low intensity condition in the affective priming task. This pattern of results, which is in contrast to that seen for explicit evaluations, may reflect a difference in sensitivity between explicit and implicit measures.

In summary, we were able to partially replicate the findings from Moran and Bar-Anan (2013) while measuring physiology and manipulating US intensity. We showed that startle blinks elicited during backward CSs largely follow the pattern of explicit valence ratings in this paradigm. Our findings also demonstrate that US intensity cannot account for the differing dissociations between implicit and explicit measures reported by Moran and Bar-Anan (2013) and Andreatta et al. (2010).

3.4 Experiment 2

The findings from Experiment 1 and work by Andreatta and colleagues indicate that backward CSs following aversive USs have acquired positive valence, as startles elicited during CSs presented after aversive stimuli (CS+/CSneg) are smaller than startles elicited during CSs presented alone (CS-) or CSs presented after positive USs (CSpos; Andreatta et al., 2010; Andreatta et al., 2013). This effect is known as ‘relief learning’, as the positive effect that occurs after the offset of an aversive stimulus elicits feelings of relief (Deutsch et al., 2015; Gerber et al., 2014). It has been proposed that a similar valence reversal would be observed after positive stimuli, i.e., that the offset of a positive stimulus would elicit negative feelings, which would result in stimuli presented after a positive US acquiring negative valence (B-CSpos; see Felsenberg et al., 2014 for demonstration in honeybees; Gerber et al., 2014). However, in absence of a neutral baseline condition, it is difficult to determine whether the relative difference in backward CS valence observed in Experiment 1 reflects positive valence for B-CSneg and negative valence for B-CSpos. If this were the case, we would expect startle modulation during backward CSs paired with negative, neutral, and positive USs to follow a linear trend. This would be shown by smaller startles during the B-CSneg (suggesting positive valence) compared to during a CS paired with a neutral US (B-CSneut), and

larger startles during the B-CSpos (suggestive of negative valence) compared to during the B-CSneut. On the other hand, if valence acquisition during pleasant and aversive backward conditioning are qualitatively different, and negative valence does not occur at the offset of a positive stimulus, a quadratic trend would be expected. In this case, startle responses during the B-CSneut would be larger than during both the B-CSpos and B-CSneg. This would suggest positive valence for both the B-CSpos and B-CSneg, regardless of any observed difference between them (such as larger startle inhibition during the B-CSneg relative to the B-CSpos). This would mean that the backward CS contrast effects observed on the startle blink reflex in Experiment 1 and in Andreatta and colleagues' work would be the sole result of startle inhibition during the B-CSneg, with no opponent process occurring for backward conditioning with the positive US (Andreatta et al., 2010; Andreatta et al., 2013).

To investigate whether a linear or quadratic trend best represents startle blink magnitude during backward conditioning, we added trials with a neutral US to the paradigm used in Experiment 1. The addition of these neutral US trials would have required participants to learn six contingencies; start positive US, stop positive US, start neutral US, stop neutral US, start negative US, and stop negative US. In order to make the task less challenging, we decided to remove the forward CSs from the procedure. This meant that participants only had to learn three contingencies (stop positive US, stop neutral US, and stop negative US), which increased the likelihood of correct contingency recall. Removing the forward CS from a CS-US-CS paradigm has been shown to have no effect on the startle blink reflex during backward conditioning (Andreatta et al., 2010; Andreatta et al., 2013). For explicit valence ratings, however, Andreatta et al. (2013) found that a concurrent forward and backward conditioning design (CS-US-CS) led to backward CS contrast effects, while simple backward conditioning (US-CS) led to backward CS assimilation effects. As we retained the instructions from Experiment 1 and presented them in a backward conditioning paradigm (US-CS), we were afforded the opportunity to test whether a backward CS contrast effect as predicted by the instructions, or a backward CS assimilation effect as predicted by the paradigm, would occur. No explicit hypothesis was proposed.

Experiment 2 was designed to assess whether startle modulation during backward conditioning with positive, neutral and aversive USs would reveal a linear or quadratic pattern. We also wanted to assess whether backward CS contrast effects

would still be observed on explicit valence ratings when no forward CSs were presented due to the instructions highlighting the role of the backward CSs. To determine this, participants were told to learn which CSs stopped the pleasant, neutral, and aversive USs (backward conditioning: US-CS). It was hypothesised that startle blink modulation would follow a significant linear trend indicative of backward CS contrast effects. Largest responses were expected during CSs following positive USs, and smallest responses during CSs following negative USs.

3.4.1 Method

Participants. Thirty-eight undergraduate students (25 female) from the School of Psychology at Curtin University participated in this experiment for course credit, *M* age = 22, *SD* = 6.89. As in Experiment 1, sample size was based on previous research (Andreatta et al., 2013; Andreatta et al., 2010). Two participants failed the recollective memory test. Analyses were run with and without these participants and the pattern of results do not differ, hence results from the full sample are reported.

Apparatus/Stimuli. The apparatus and stimuli were the same as for Experiment 1, except only three of the alien families were used as CS sets (yellow, purple, and red), USs from the low intensity group were used (positive US: 47 dBA; negative US: 72 dBA), and a neutrally valenced auditory US was added. Low intensity USs were used as they showed a clearer pattern of startle modulation in Experiment 1 (despite resulting in non-significant backward CS contrast effects on explicit valence ratings). The neutral US was selected by asking participants in a pilot study (n = 20) to provide valence and intensity ratings for 9 neutral stimuli chosen from the International Affective Digitized Sounds (2nd Edition; IADS-2) database. These stimuli were matched in volume to the positive and negative USs from the low intensity group in Experiment 1. The stimulus that was rated as the most neutral on valence and that participants could accurately describe was chosen as the neutral US. This stimulus was the sound of a train passing a train station (sound #425 from IADS-2, 56 dBA; pilot study reported at https://osf.io/q46mp/).

Procedure. The procedure was the same as in Experiment 1, except that during acquisition, only a backward conditioning procedure (US-CS) was used. Participants were presented with eight negative, eight positive, and eight neutral trials. Trials were presented in a pseudo random order, with no more than two consecutive trials of the same valence. On each trial, USs were presented for 10, 15,

20, 25, or 30 seconds, and CSs were presented for 8 seconds, beginning 2 seconds before US offset (see Figure 6 for a depiction of a positive US trial) . This resulted in USs and CSs overlapping for 2 seconds. Startle probes were assigned randomly and presented on 18 of the 24 trials, with six probes occurring in each valence condition (see Figure 7 for a depiction of startle probe timing). Half of the ITIs were probed, totalling 12 startle probes. During extinction, each CS from each of the three CS sets was presented once, and two CSs from each set were presented twice for a total of six presentations per set. CSs were shown for 8s each, for a total of 18 presentations. Four of the six presentations of each CS set were probed, totalling 18 startle probes. Half of the ITIs were probed, totalling nine startle probes. The instructions used were the same as in Experiment 1, except participants were told each of the three families would stop one of the sound USs[3]. In the affective priming task sixty trials were presented as only three CS sets were used in the conditioning task. Each of the three sets was presented once with positive and negative words. All other details of the affective priming task were the same as in Experiment 1. In the memory test, participants were shown each family separately, and asked:

What was the role of the creatures in this picture? 1. Stopped the human sound. 2. Stopped the musical sound. 3. Stopped the metropolitan sound.

The post-experimental questionnaire was the same as in Experiment 1, with the addition of asking for valence and intensity ratings of the metropolitan sound.

Scoring, response definition, and statistical analyses. Data were analysed using repeated measures ANOVAs. All other details are the same as in Experiment 1 unless noted below.

Startle. *T*-scores were subjected to a 3 (US Valence: Positive vs neutral vs negative) × 3 (Block: 1, 2, 3) repeated measures ANOVA with a subsequent trend analysis. No participants were excluded.

Explicit valence ratings. Ratings were subjected to a repeated measures ANOVA and subsequent trend analysis comparing (US valence: Positive vs neutral vs negative). No participants were excluded.

[3] An error in the instructions was spotted by the 27th participant. Instead of saying "The three families of creatures:" it said "The four families of creatures:", and then showed only three families. Participants after this were asked if they noticed anything about the instructions, and then if they noticed if it said "four families" at any point upon completion of the experiment. Of the 11 participants asked, four of them noticed. Three participants said they thought it was a typo, and one of them thought it was referring to the fact that there were four aliens in each family.

Affective priming. As neutrally valenced USs were presented, scores based on the difference between positive and negative target words for each prime valence were calculated for reaction times and errors. Assimilation effects are represented by negative scores for positive CS primes and by positive scores for negative CS primes. These scores were subjected to a repeated measures ANOVA trend analysis (US valence: Positive vs neutral vs negative). No participants were excluded. Analysis of the error data did not add substantially to the current report and is reported in the supplementary materials at https://osf.io/q46mp/.

Manipulation checks. Post-experimental valence and intensity ratings of the USs and startle probe were subjected to separate repeated measures ANOVAs (positive US vs neutral US vs negative US vs startle probe; within-participants). A linear relationship was found for US valence ratings, F (3, 35) = 146.87, $p < .001$, $\eta_p^2 = .926$. The positive US (M = 2.40, SD = 0.72) was rated significantly more positively than the neutral US (M = 0.03, SD = 1.05), t (37) = 11.20, $p < .001$, d = 1.82, the negative US (M = -1.82, SD = 0.80), t (37) = 19.10, $p < .001$, d = 3.10, and the startle probe (M = -1.84, SD = 0.97), t (37) = 19.14, $p < .001$, d = 3.10. The neutral US (M = 0.03, SD = 1.05) was rated significantly more positively than the negative US (M = -1.82, SD = 0.80), t (37) = 8.19, $p < .001$, d = 1.33, and the startle probe (M = -1.84, SD = 0.97), t (37) = 8.35, $p < .001$, d = 1.35. There were no differences between the negative US (M = -1.82, SD = 0.80) and the startle probe (M = -1.84, SD = 0.97), t (37) = 0.13, p = .898, d = 0.02. A linear relationship was also observed for US intensity ratings, F (3, 35) = 72.74, $p < .001$, $\eta_p^2 = .862$. The startle probe (M = 4.55, SD = 1.06) was rated as significantly more intense than the negative US (M = 3.76, SD = 1.34), t (37) = 3.53, p = .001, d = 0.57, the neutral US (M = 2.18, SD = 1.33), t (37) = 10.26, $p < .001$, d = 1.66, and the positive US (M = 1.34, SD = 1.36), t (37) = 14.79, $p < .001$, d = 2.40. The negative US (M = 3.76, SD = 1.34) was rated as significantly more intense than the neutral US (M = 2.18, SD = 1.33), t (37) = 5.29, $p < .001$, d = 0.86, and the positive US (M = 1.34, SD = 1.36), t (37) = 9.30, $p < .001$, d = 1.51. The neutral US (M = 2.18, SD = 1.33) was rated as significantly more intense than the positive US (M = 1.34, SD = 1.36), t (37) = 3.75, p = .001, d = 0.61.

3.4.2 Results

Startle blink magnitude – Acquisition. Figure 8 shows larger responses during CSs following positive USs than during CSs following neutral and negative

USs which decreased across blocks. This was confirmed by main effects of US valence, F (2, 36) = 8.77, p = .001, η_p^2 = .328, and block, F (2, 36) = 23.58, p < .001, η_p^2 = .567. Responses during CSs following the positive US were marginally larger than responses during CSs following the neutral US, t (36) = 1.97, p = .057, d = 0.32, and significantly larger than responses during CSs following the negative US, t (36) = 4.23, p < .001, d = 0.69. Responses during CSs paired with the neutral US were significantly larger than responses during CSs paired with the negative US, t (36) = 2.19, p = .034, d = 0.36. Responses at block 1 were larger than blocks 2, t (36) = 3.40, p = .002, d = 0.55, and 3, t (36) = 6.80, p < .001, d = 1.10, and responses at block 2 were larger than block 3, t (36) = 3.89, p < .001, d = 0.63. Tests of within-subject contrasts showed only linear trends for US valence, F (1, 37) = 17.90, p < .001, η_p^2 = .326, and block, F (1, 37) = 46.29, p < .001, η_p^2 = .556.

Explicit valence ratings. Figure 9 shows a linear trend for US valence suggestive of an assimilation effect, as CSs following the positive US were rated more positively than CSs following the neutral US, and CSs following the neutral US were rated as more positive than CSs following the negative US. This was confirmed by a significant one-way repeated measures ANOVA, F (2, 36) = 9.16, p = .001, η_p^2 = .337. CSs paired with the positive US were rated as significantly more pleasant than CSs paired with the neutral US, t (36) = 2.40, p = .022, d = 0.39, and the negative US, t (36) = 4.31, p < .001, d = 0.70, and CSs paired with the neutral US were rated as significantly more pleasant than CSs paired with the negative US, t (36) = 3.46, p = .001, d = 0.56. Tests of within-subject contrasts showed only a significant linear trend for US valence, F (1, 37) = 18.61, p < .001, η_p^2 = .335.

Affective priming – Reaction times. Figure 10 shows a linear trend suggestive of an assimilation effect, although the main effect for US valence was not significant, F (2, 36) = 1.65, p = .206, η_p^2 = .084, and the trend for US valence was only marginal, F (1, 37) = 3.14, p = .085, η_p^2 = .078.

3.4.3 Discussion

Experiment 2 aimed to determine whether both positive and negative USs lead to opposing emotional responses at their offset (shown by linear startle modulation), and whether instructions highlighting the role of the backward CSs exert their effect in a backward conditioning only paradigm. Linear trends for explicit valence ratings and affective priming (only trending) suggest that backward CS assimilation effects occurred, despite presenting instructions that should support

backward CS contrast effects. This provides evidence that the backward CS contrast effects driven by instructional manipulations, as reported for instance in Experiment 1, were not due to demand characteristics, as the same pattern of results should have emerged here. Removing the forward CS appeared to have no impact on startle blink magnitude and inclusion of the neutral US pairing showed that startle blink modulation also followed a linear trend. This trend was suggestive of a contrast effect as startle blink magnitude was larger during CSs following the positive US than during CSs following neutral and negative USs. This confirmed that an opponent process mirroring relief occurs at the offset of positive stimuli, which to our knowledge is the first demonstration that the offset of both positive and negative stimuli elicits an opposing emotional reaction in humans which can be indexed by startle blink reflexes.

While emotional responses elicited at the offset of valenced stimuli seem the most plausible explanation for the pattern of startle modulation, it is also possible that the instructional manipulation highlights the role of the CSs and therefore affects startle modulation. This is because the same pattern of startle modulation is expected from the instructions, i.e. CSs stopping the negative US should become positive, and CSs stopping the positive US should become negative. Even though the instructions did not affect explicit valence ratings, we cannot rule out this conclusion because explicit valence ratings and the startle blink reflex have been shown to dissociate in backward conditioning only designs (US-CS; Andreatta et al., 2013; Andreatta et al., 2010). Future research should confirm that the inverse 'relief learning' process occurring at the offset of the positive US was not due to the instructions.

3.5 General Discussion

The current experiments assessed whether US intensity could explain the different patterns of dissociations between explicit and implicit measures of backward CS valence reported in studies of evaluative and fear conditioning (Experiment 1), and whether a linear pattern of startle modulation suggestive of opposing emotional responses after the offset of positive and negative USs would emerge during backward conditioning (Experiment 2). Moreover, Experiment 1 assessed whether due to the instructional manipulation used, startle modulation and explicit valence ratings would reveal backward CS contrast effects, and Experiment

2 tested whether presenting similar instructions in a backward conditioning only procedure would lead to assimilation or contrast effects. The current results suggest that backward conditioning leads to the same pattern of startle modulation regardless of whether forward and backward conditioning are trained concurrently or only backward conditioning is assessed, and whether a neutral US is included in the backward conditioning procedure. The intensity effect observed on the startle response in Experiment 1 shows that responses overall were smaller in the high intensity group, not that the pattern of startle modulation differs as a function of US valence or conditioning type. Hence, US intensity does not moderate backward CS contrast effects on the startle response. On the other hand, explicit valence ratings appear sensitive to US intensity and the conditioning procedure. Explicit valence ratings revealed backward CS contrast effects during concurrent forward and backward conditioning in Experiment 1 in the high intensity condition only, and a backward CS assimilation effect during backward conditioning in Experiment 2. Moreover, the instructional manipulation highlighting the role of the CSs appeared to have no effect on explicit valence ratings without concurrent forward conditioning in Experiment 2.

The backward conditioning results on explicit valence ratings in Experiments 1 and 2 replicate earlier work on relief learning that yielded assimilation effects when the US was unpredictable (US-CS), and contrast effects when the US was predictable (CS-US-CS; Andreatta et al., 2013). However, unlike Andreatta and colleagues, we informed participants that the backward CS would stop the US which was expected to support backward CS contrast effects on explicit valence ratings, as observed in Experiment 1 and in other studies involving relational instructions (Moran & Bar-Anan, 2013; Moran et al., 2016). Below we offer three explanations as to why contrast effects did not occur in Experiment 2.

Firstly, it is possible that the instructions were less salient in the US-CS design than in the CS-US-CS design, as participants only had to learn one relation for each of the USs. Less focus on the instructions may have resulted in weaker encoding of the proposition that families stop the USs, therefore rendering the instructions ineffective. However, the fact that only two participants failed the recollective memory test suggests this was not the case, as poor encoding of the instructions should also result in poor performance on the recollective memory test.

Second, it may be that the onset of the US is more salient in the US-CS design because there is no forward CS. This may render the US more aversive, which may then overpower the effect of the instructions leading to an assimilation effect. If this was to occur, then more intense USs should also lead to assimilation effects. Experiment 1 shows this was not the case, as the high intensity USs led to larger contrast effects than the low intensity USs.

Finally, the lack of a backward CS contrast effect may be due to temporal overshadowing. In the CS-US-CS trials the forward CS could be considered the most salient CS as it predicts the onset of the US. If overshadowing can occur across stimuli in a temporal arrangement, then the forward CS may have overshadowed the association between the backward CS and the valence of the US. This may permit the association between the backward CS and the emotional response elicited by US offset to become apparent, and/or for the instructional manipulation to take effect. In absence of the forward CS no overshadowing occurred which permitted the development of an association between US valence and the backward CS. This explanation can also account for the findings of Andreatta et al. (2013), who showed backward CS contrast effects in the CS-US-CS design and assimilation effects in the US-CS design, in absence of any instructional manipulation. If we consider that the forward CS overshadows valence transfer from the US to the backward CS, then backward CS valence may be influenced by feelings of relief after US offset, resulting in positive ratings of a backward CS following an aversive shock. While intriguing, the temporal overshadowing account holds only if we assume that the startle reflex and valence ratings reflect different learning mechanisms, as startle modulation showed the same pattern of results in both the CS-US-CS and US-CS designs.

The studies we have presented confirm that the offset of a positive US leads to negative emotion in humans (disappointment). Startle responses were larger during CSs presented at the offset of the positive US in comparison to CSs presented after neutral and negative USs. This process of disappointment learning mirrors that of relief learning which occurs during backward conditioning at the offset of a positive US. While it is early to speculate on potential clinical implications of disappointment learning, this phenomenon may hold explanatory value for those with affective disorders who tend to avoid situations in which they may experience pleasure.

In the case of avoiding pleasure, the pleasurable experience could be considered a positive US, and the offset of this pleasurable experience may result in disappointment. In terms of disappointment learning, any stimulus that is present at the offset of this pleasurable experience may acquire negative valence. This means that if any component of the pleasurable experience persists once pleasure is no longer being experienced, this component may become negative. The result would be a pleasurable experience that is now remembered as disappointing. In addition to this, the disappointment experienced at the offset of the pleasurable experience may also serve as a punisher that reduces the likelihood that an individual will partake in the pleasurable experience again. It is also possible that the disappointment experienced at the offset of the pleasurable experience is more intense and/or more salient than the pleasure itself. If so, individuals may avoid pleasant experiences all together in order to avoid the possibility of having to experience disappointment.

Another situation in which our findings may provide insight is substance abuse. Substance abuse may be motivated by the reduction of negative experiences through substance use which results in a pleasant experience. The offset of this pleasant experience may then result in disappointment, which instead of resulting in disappointment learning that may dissuade future substance abuse, leads to substance use again to provide relief from disappointment. In this scenario, disappointment could be considered a negative US and substance use itself could serve as a stimulus that initially provides relief from the negative US. It then becomes a backward CS that elicits positive feelings from being paired with the offset of an aversive event (relief learning). The end result is a vicious cycle where disappointment, relief, and relief learning serve to perpetuate substance use. Future research should investigate whether the concepts of disappointment and disappointment learning can be used to further our understanding of affective and substance use disorders. A limitation worth considering is that using low intensity USs in Experiment 2 reduced the chance to find backward CS contrast effects on explicit valence ratings, as no backward CS contrast effect was observed in the low intensity group in Experiment 1. However, such an effect of the low intensity USs seems unlikely as a clear assimilation effect was found in Experiment 2 and because startle blink magnitude shows that low intensity USs were sufficiently intense to elicit relief learning in both experiments. Moreover, these findings replicate that of Andreatta et al. (2013),

suggesting that using low intensity USs did not preclude the observation of a backward CS contrast effect.

In summary, the current experiments show that US intensity does not moderate backward CS contrast effects to the point of attaining different patterns of startle modulation. Moreover, it shows that removing concurrent forward conditioning and adding a neutral US does not affect startle modulation during backward CSs and that both pleasant and aversive stimuli lead to contrasting emotional responses at their offset. Finally, assimilation effects were observed on explicit valence ratings in a backward conditioning paradigm (US-CS), even when relational instructions that should support backward CS contrast effects were presented. This suggests that backward conditioning is affected by simultaneous forward conditioning and that these events have a larger effect on learning than relational instructions that emphasise a-priori propositions about the relationships between stimuli.

3.6 References

Andreatta, M., Mühlberger, A., Glotzbach-Schoon, E., & Pauli, P. (2013). Pain predictability reverses valence ratings of a relief-associated stimulus. *Frontiers in Systems Neuroscience, 7*(53), 1-12, https://doi.org/10.3389/fnsys.2013.00053

Andreatta, M., Mühlberger, A., & Pauli, P. (2016). When does pleasure start after the end of pain? The time course of relief. *Journal of Comparative Neurology, 524*, 1653-1667. https://doi.org/10.1002/cne.23872

Andreatta, M., Mühlberger, A., Yarali, A., Gerber, B., & Pauli, P. (2010). A rift between implicit and explicit conditioned valence in human pain relief learning. *Proceedings of the Royal Society of London B: Biological Sciences*. https://doi.org/10.1098/rspb.2010.0103

Andreatta, M., & Pauli, P. (2017). Learning mechanisms underlying threat absence and threat relief: Influences of trait anxiety. *Neurobiology of Learning and Memory, 145*, 105-113. https://doi.org/10.1016/j.nlm.2017.09.005

Annau, Z., & Kamin, L. J. (1961). The conditioned emotional response as a function of intensity of the US. *Journal of Comparative and Physiological Psychology, 54, 428-432*. https://doi.org/10.1037/h0042199

Bitar, N., Marchand, S., & Potvin, S. (2018). Pleasant pain relief and inhibitory conditioned pain modulation: A psychophysical study. *Pain Research and Management, 2018*, 1-8. https://doi.org/10.1155/2018/1935056

Blumenthal, T. D., Cuthbert, B. N., Filion, D. L., Hackley, S., Lipp, O. V., & Van Boxtel, A. (2005). Committee report: Guidelines for human startle eyeblink electromyographic studies. *Psychophysiology, 42*, 1-14. https://doi.org/10.1111/j.1469-8986.2005.00271.x

Bradley, M. M., Cuthbert, B. N., & Lang, P. J. (1990). Startle reflex modification: Emotion of attention? *Psychophysiology, 27*, 513-522. https://doi.org/10.1111/j.1469-8986.1990.tb01966.x

Bradley, M. M., & Lang, P. J. (2007). *The International Affective Digitized Sounds (2nd Edition; IADS- 2): Affective ratings of sounds and instruction manual. Technical report B-3*. University of Florida, Gainesville, Fl.

De Houwer, J. (2007). A conceptual and theoretical analysis of evaluative conditioning. *The Spanish Journal of Psychology, 10,* 230-241. https://doi.org/10.1017/S1138741600006491

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,* 853- 869. https://doi.org/10.1037//0033-2909.127.6.853

Deutsch, R., Smith, K. J. M., Kordts-Freudinger, R., & Reichardt, R. (2015). How absent negativity relates to affect and motivation: An integrative relief model. *Frontiers in Psychology, 6(152),* 1-23. https://doi.org/10.3389/fpsyg.2015.00152

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

Felsenberg, J., Plath, J. A., Lorang, S., Morgenstern, L., & Eisenhardt. D. (2014). Short- and long-term memories formed upon backward conditioning in honeybees (Apis mellifera). *Learning and Memory, 21,* 37-45. https://doi.org/10.1101/lm.031765.113

Galdi, S., Arcuri, L., & Gawronski, B. (2008). Automatic mental associations predict future choices of undecided decision makers. *Science, 321,* 1100-1102. https://doi.org/10.1126/science.1160769

Gibson, B. (2008). Can evaluative conditioning change attitudes toward mature brands? New evidence from the Implicit Association Test. *Journal of Consumer Research, 35,* 178-188. https://doi.org/10.1086/527341

Gerber, B., Yarali, A., Diegelmann, S., Wotjak, C. T., Pauli, P., & Fendt, M. (2014). Pain- relief learning in flies, rats, and man: Basic research and applied

perspectives. *Learning and Memory, 21*, 232-252. https://doi.org/10.1101/lm.032995.113

Green, L. J. S., Luck, C., Gawronski, B., & Lipp, O. V. (2019). Contrast effects in backward evaluative conditioning: Exploring effects of affective relief/disappointment versus instructional information. *Emotion*. Advance online publication. https://doi.org/10.1037/emo0000701

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. *Psychological Bulletin, 136*, 390-421. https://doi.org/10.1037/a0018916

Hu, X., Gawronski, B., & Balas, R. (2017a). 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*, 17-32. https://doi.org/10.1177/0146167216673351

Hu, X., Gawronski, B., & Balas, R. (2017b). Propositional versus dual-process accounts of evaluative conditioning: II. The effectiveness of counter-conditioning and counter-instructions in changing implicit and explicit evaluations. *Social Psychological and Personality Science, 8*, 858-866.

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

LeBel, E. P., & Campbell, L. (2009). Implicit partner affect, relationship satisfaction, and the prediction of romantic breakup. *Journal of Experimental Social Psychology, 45*, 1291-1294. https://doi.org/10.1016/j.jesp.2009.07.003

Leknes, S., Brooks, J. C. W., Wiech, K., & Tracey, I. (2008). Pain relief as an opponent process: A psychophysical investigation. *European Journal of Neuroscience, 28*, 794-810. https://doi.org/10.1111/j.1460-9568.2008.06380.x

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

Lipp, O.V., Siddle, D.A.T., & Dall, P.J. (2003). The effects of unconditional stimulus valence and conditioning paradigm on verbal, skeletal, and autonomic indices of Pavlovian conditioning. *Learning and Motivation, 34*, 32-51. https://doi.org/10.1016/S0023-9690(02)00507-6

Lipp, O. V. (2006). Human fear learning: Contemporary procedures and measurement. In M. G. Craske, D. Hermans & D. Vansteenwegen (Eds.), (2006). Fear and learning: From basic processes to clinical implications (pp. 37-52). Washington: APA Books.

Luck. C. C., & Lipp, O. V. (2017). Startle modulation and explicit valence evaluations dissociate during backward fear conditioning. *Psychophysiology, 54,* 673-683. https://doi.org/10.1111/psyp.12834

Mallan, K. M., Lipp, O. V., & Libera, M. (2008). Affect, attention, or anticipatory arousal? Human blink startle modulation in forward and backward affective conditioning. *International Journal of Psychophysiology, 69.* 9-17. https://doi.org/10.1016/j.ijpsycho.2008.02.005

Moran, T., and Bar-Anan, Y. (2013). The effect of object-valence relations on automatic evaluation. *Cognition and Emotion. 27.* 743-752. https://doi.org/10.1080/02699931.2012.732040

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2016). The assimilative effect of co-occurrence on evaluation above and beyond the effect of relational qualifiers. *Social Cognition, 34.* 435-461. https://doi.org/101521soco2016345435

Mühlberger, A., Wieser, M. J., Gerdes, A. B. M., Frey, M. C. M., Weyers, P., & Pauli, P. (2015). Stop looking angry and smile, please: Start and stop of the very same facial expression differentially activate threat- and reward-related brain networks. *Social Cognitive and Affective Neuroscience, 6,* 321-329. https://doi.org/10.1093/scan/nsq039

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

Pavlov, I. P. (1927). Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex. Oxford, England: Oxford University Press.

Unkelbach, C., & Fiedler, K. (2016). Contrastive CS-US relations reverse evaluative conditioning effects. *Social Cognition, 34,* 413-434. https://doi.org/101521soco2016345413

Vrana, S. R., Spence, E. L., & Lang, P. J. (1988). The startle probe response: A new measure of emotion? *Journal of Abnormal Psychology, 97,* 487-491. https://doi.org/10.1037/0021-843X.97.4.487

3.7 Footnotes

[1] Moran and Bar-Anan (2013) did not report the intensity to which their sound USs were set, but attempts to recreate them following their description suggest that they were below 90dBA.

[2] All materials, data, analysis files, and supplementary materials, are available at https://osf.io/q46mp/.

[3] An error in the instructions was spotted by the 27th participant. Instead of saying “The three families of creatures:” it said “The four families of creatures:”, and then showed only three families. Participants after this were asked if they noticed anything about the instructions, and then if they noticed if it said “four families” at any point upon completion of the experiment. Of the 11 participants asked, four of them noticed. Three participants said they thought it was a typo, and one of them thought it was referring to the fact that there were four aliens in each family.

3.8 Figures and Tables

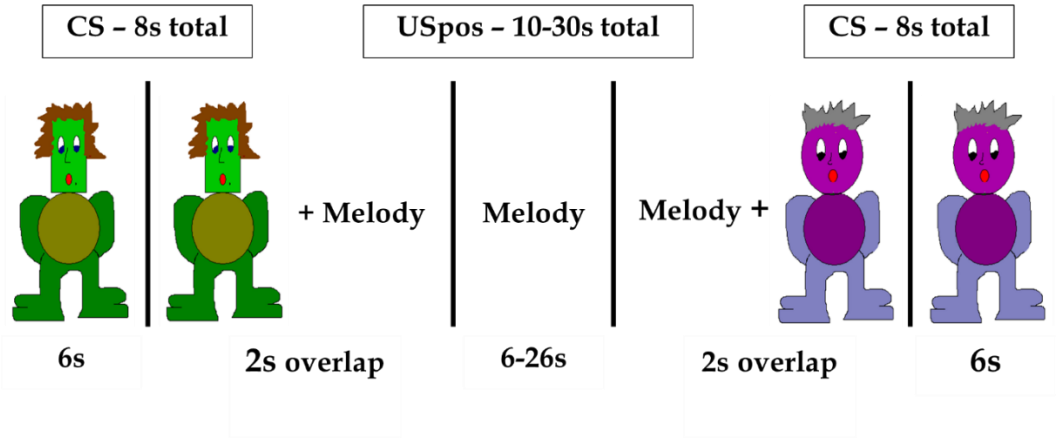

Figure 1. Example of a positive US trial in Experiment 1. Forward and backward CSs were presented alone for 6 seconds and overlapping with the US for 2 seconds (8 seconds of total CS presentation). USs varied in duration for 10, 15, 20, 25, or 30 seconds. CS = Conditional stimulus, USpos = Positive unconditional stimulus.

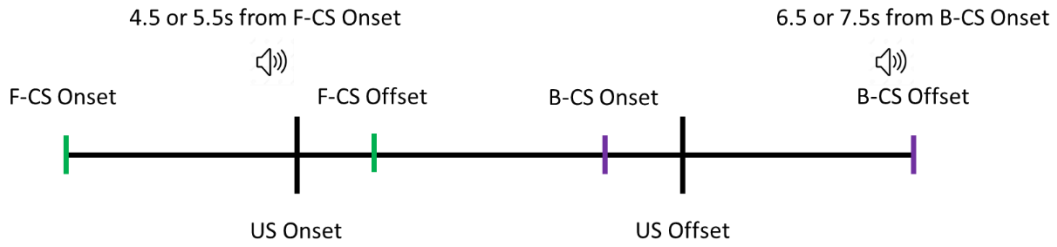

Figure 2. Example of startle probe timing relative to F-CS, US, and B-CS onset and offset in Experiment 1. Startle probes were presented at 4.5 or 5.5 seconds after F-CS onset and 6.5 or 7.5 seconds after B-CS onset. F-CS = Forward conditional stimulus, US = Unconditional stimulus, B-CS = Backward conditional stimulus, speaker picture represents startle probe.

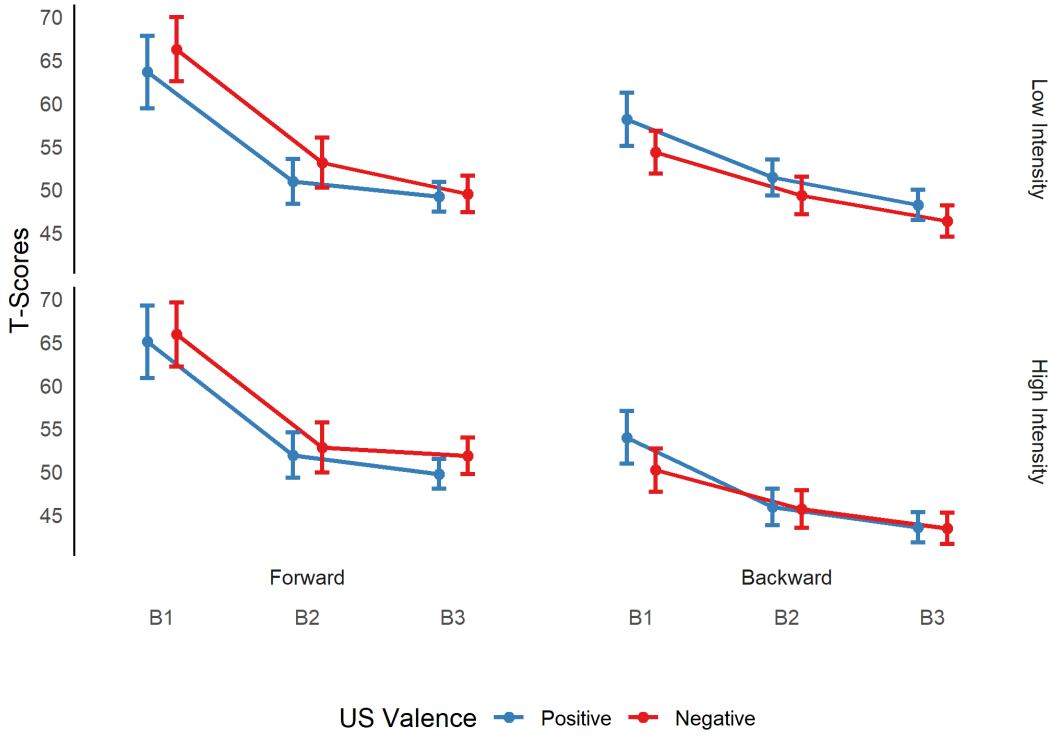

Figure 3. Startle blink magnitude (*T*-scores) by block (1, 2, and 3) for forward and backward CSs as a function of US valence (positive vs. negative) and US intensity (low vs high). Error bars represent 95% confidence intervals of the mean.

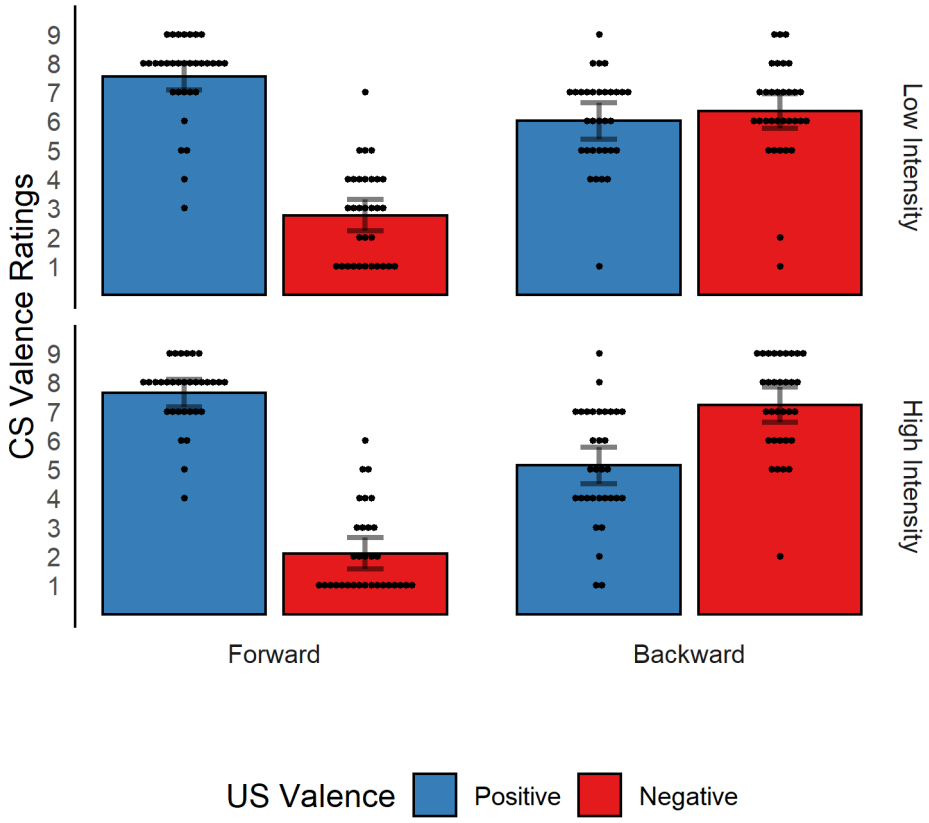

Figure 4. Mean explicit valence ratings with individual participant values plotted for forward and backward CSs as a function of US valence (positive vs. negative) and US intensity (low vs high). Error bars represent 95% confidence intervals of the mean.

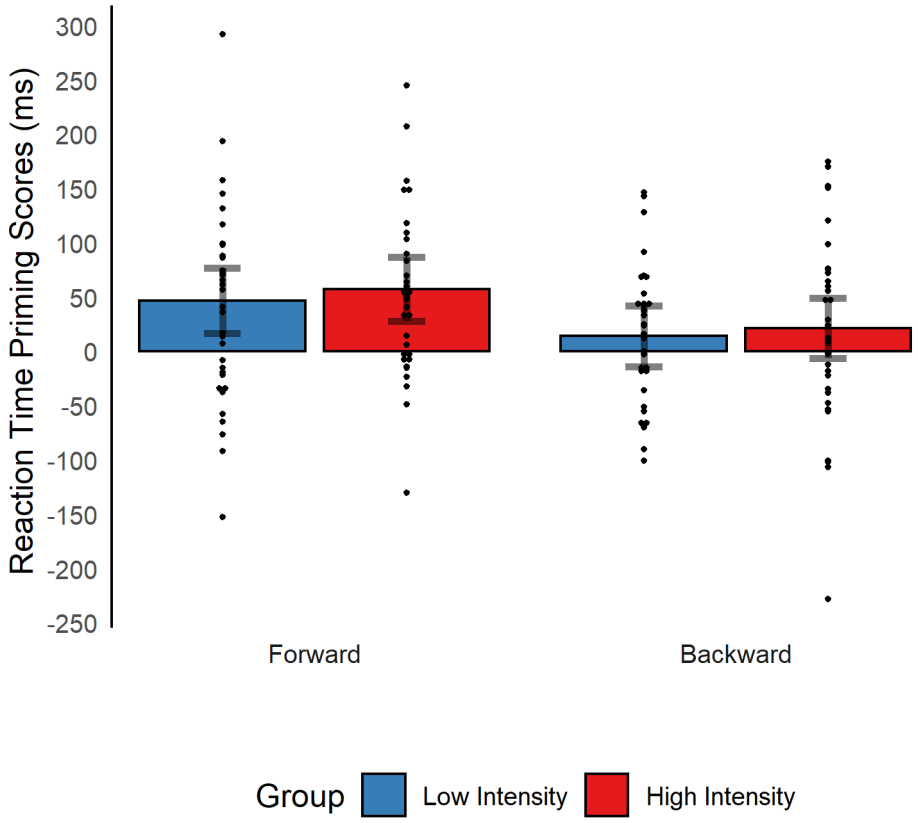

Figure 5. Mean priming scores (RT based) with individual participant values plotted from the affective priming task for forward and backward CSs as a function of US intensity. Positive scores suggest assimilation effects. Error bars show 95% confidence intervals of the mean.

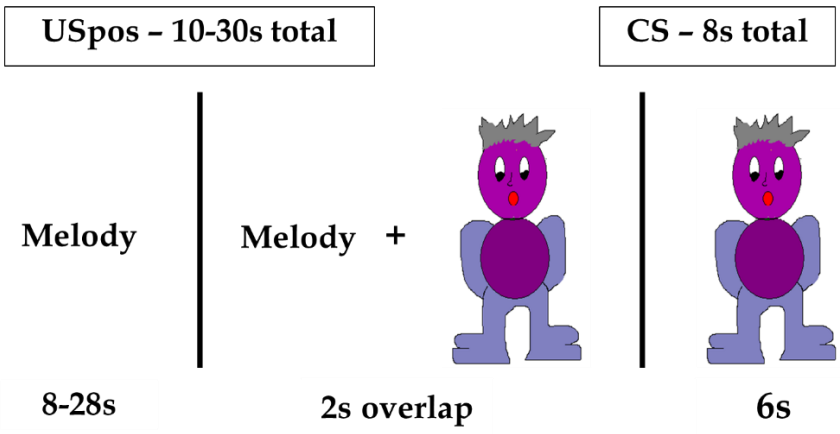

Figure 6. Example of a positive US trial in Experiment 2. Backward CSs were presented alone for 6 seconds and overlapping with the US for 2 seconds (8 seconds of total CS presentation). USs varied in duration for 10, 15, 20, 25, or 30 seconds. CS = Conditional stimulus, USpos = Positive unconditional stimulus.

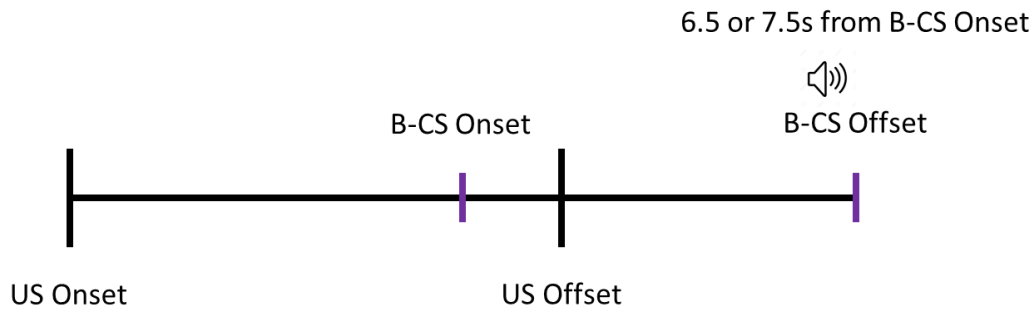

Figure 7. Example of startle probe timing relative to US and B-CS onset and offset in Experiment 2. Startle probes were presented at 6.5 or 7.5 seconds after B-CS onset. US = Unconditional stimulus, B-CS = Backward conditional stimulus, speaker picture represents startle probe.

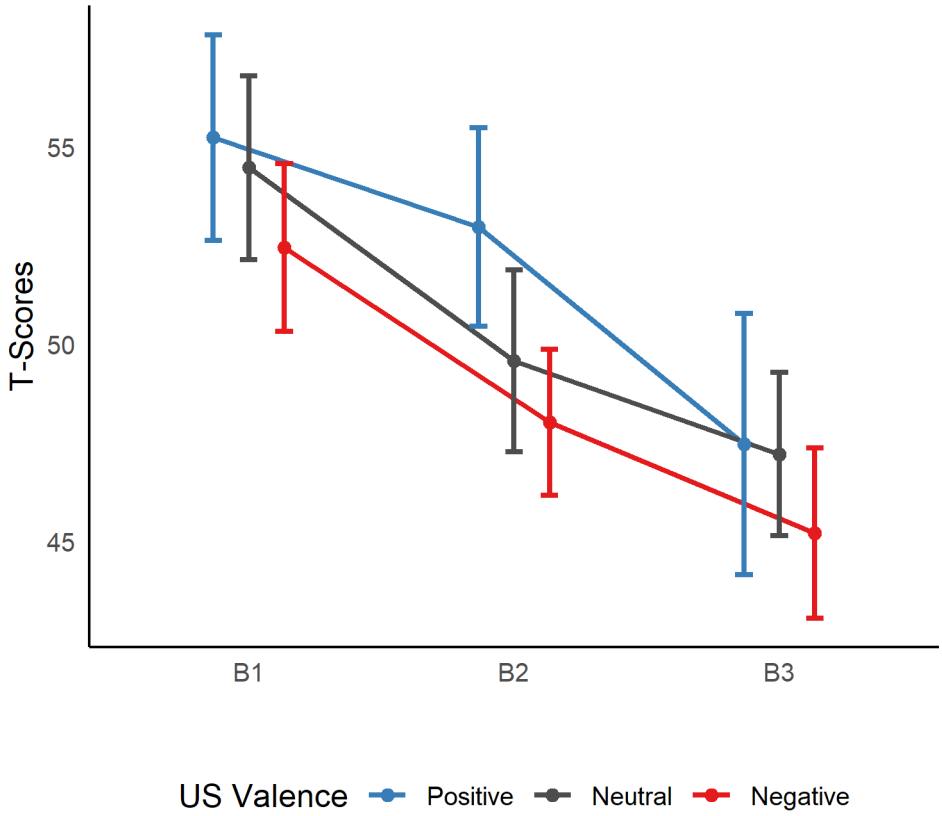

Figure 8. Startle blink magnitude (*T*-scores) by block (1, 2, and 3) for backward CS as a function of US valence (positive vs. neutral vs. negative). Error bars represent 95% confidence intervals of the mean.

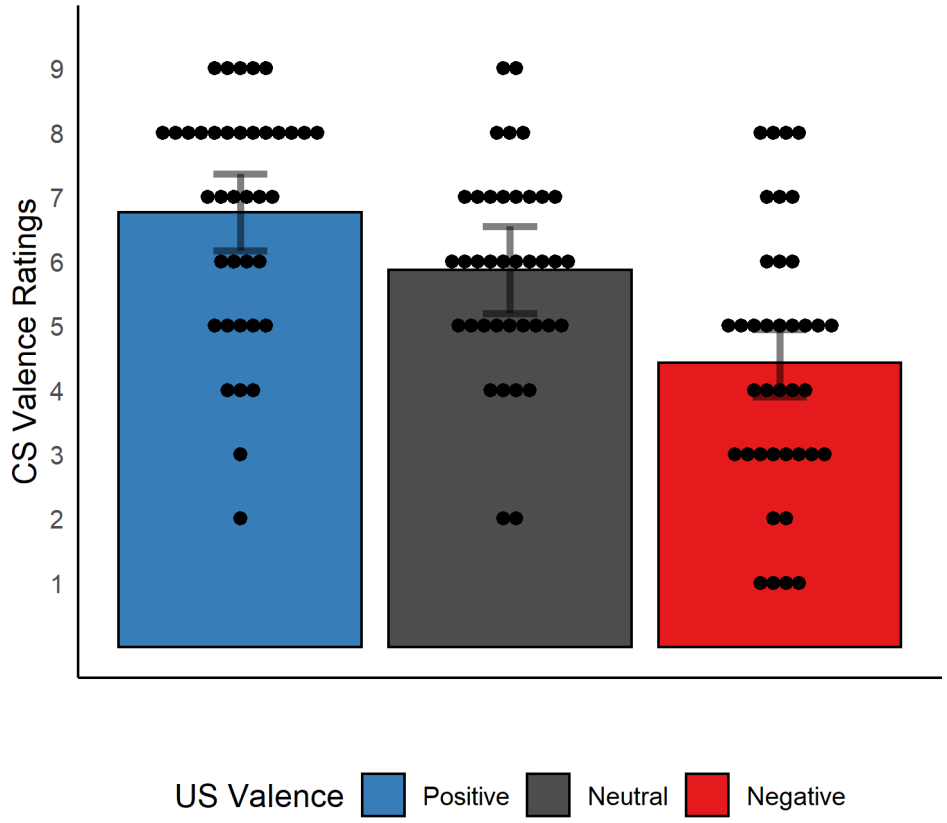

Figure 9. Mean explicit valence ratings with individual participant values plotted for backward CSs as a function of US valence (positive vs. neutral vs. negative). Error bars represent 95% confidence intervals of the mean.

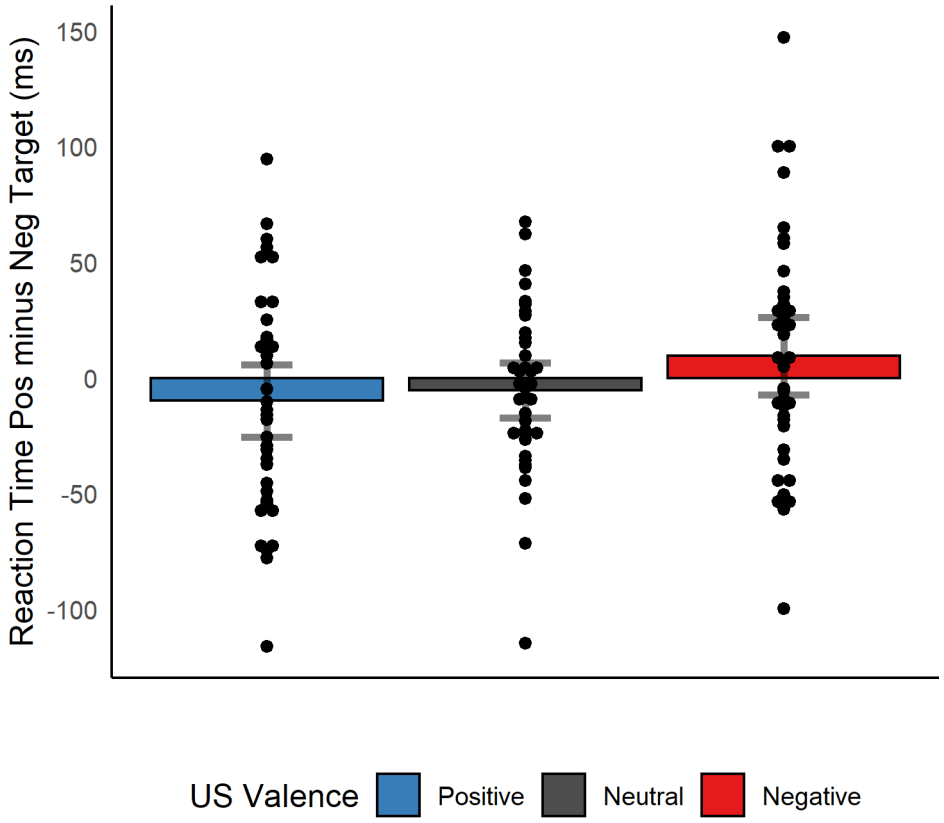

Figure 10. Difference scores (positive target words – negative target words) for reaction times with individual participant values plotted from the affective priming task for backward CSs as a function of US valence (positive vs. neutral vs. negative). Assimilation effects are represented by negative scores for positive CS primes and by positive scores for negative CS primes. Error bars show 95% confidence intervals of the mean.

Table 1: US intensities in the two groups (measured by a handheld digital sound level meter C-DSM1)

Group	Positive US – Melody	Negative US – Scream
Low Intensity	47dBA	72dBA
High Intensity	74dBA	88dBA

Note. dBA = Decibel A-Scale. US = Unconditional Stimulus

3.9 Supplementary Material – Experiment 1

Startle blink magnitude – Extinction

Figure S1 shows habituation occurring across blocks regardless of group or conditioning type. A main effect of block, F (2, 58) = 58.674, p < .001, ηp^2 = .669, was qualified by a marginal US valence x block interaction, F (2, 58) = 3.083, p = .053, ηp^2 = .096. Follow-up analyses revealed larger responses to CSs paired with positive USs at block 1, F (1, 59) = 4.176, p = .045, ηp^2 = .066, and no differences between CSs at blocks 2, F (1, 59) = 0.057, p = .812, ηp^2 = .001, and 3, F (1, 59) = 1.403, p = .241, ηp^2 = .023. As the interaction was marginal, we also followed up the main of effect of block, and found larger responses at block 1 when compared with block 2, t (59) = 10.65, p < .001, d = 1.36, and block 3, t (59) = 9.06, p < .001 d = 1.16.

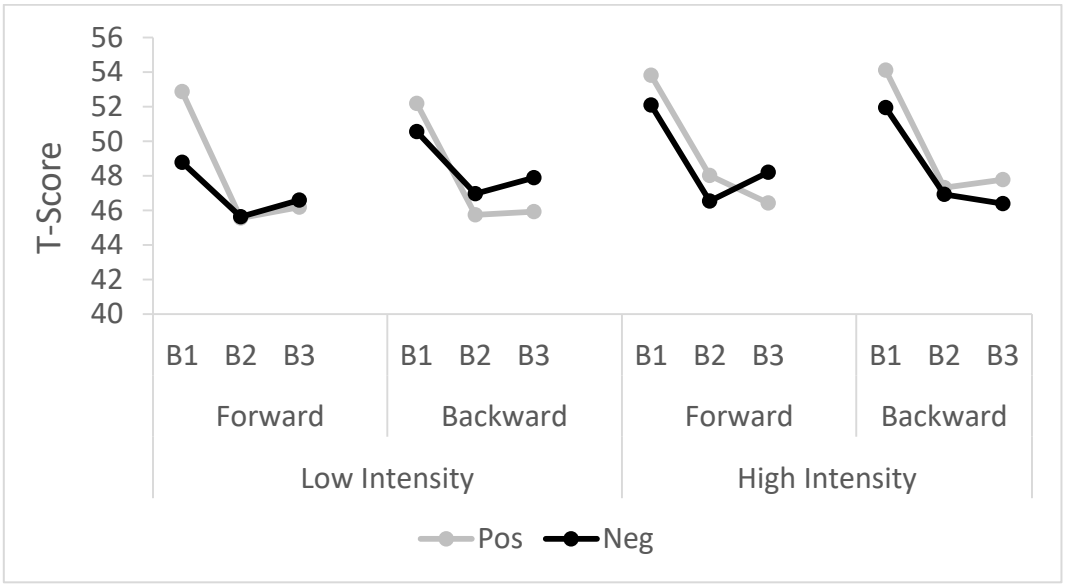

Figure S1. Startle blink magnitude (*T*-scores) by block (1, 2, and 3) during extinction, for CSs that were paired with positive and negative USs for forward and backward conditioning in the low intensity and high intensity groups.

Startle blink latency – Acquisition

Figure S2 below suggests faster blink onset following probes presented during CSs paired with negative USs than CSs paired with positive USs for forward conditioning, and faster following probes presented during CSs paired with positive USs than CSs paired with negative USs for backward conditioning, in both groups. Main effects of conditioning type, F (1, 56) = 190.209, p < .001, ηp^2 = .773, and

block, F (2, 55) = 21.857, p < .001, ηp^2 = .443, a group x conditioning type interaction, F (1, 56) = 29.188, p < .001, ηp^2 = .343, and a conditioning type x US valence interaction, F (1, 56) = 35.373, p < .001, ηp^2 = .387, were qualified by a conditioning type x US valence x block interaction, F (2, 55) = 4.446, p = .016, ηp^2 = .139. Follow-up analyses revealed an assimilation effect for forward conditioning, as blink onset was faster during CSs paired with negative USs than CSs paired with positive USs, on blocks 1, F (1, 56) = 13.14, p = .001, ηp^2 = .190, and 2, F (1, 56) = 6.559, p = .013, ηp^2 = .105, and a contrast effect for backward conditioning, as blink onset was faster during CSs paired with positive USs than CSs paired with negative USs, on blocks 1, F (1, 56) = 18.417, p < .001, ηp^2 = .247, and 3, F (1, 56) = 4.465, p = .039, ηp^2 = .074.

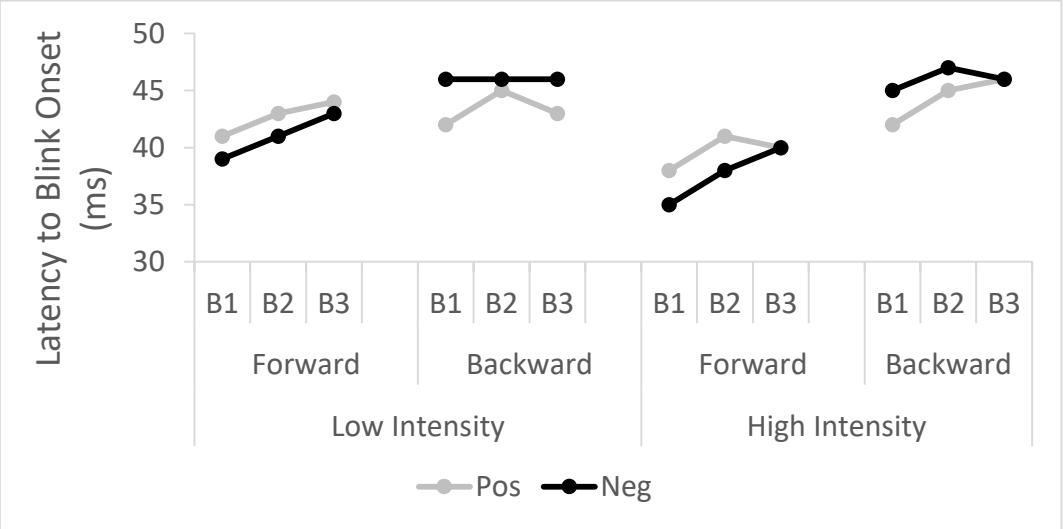

Figure S2. Time until blink onset following startle probe in milliseconds, during forward and backward CSs paired with positive and negative USs, in the low and high intensity groups across blocks 1, 2, and 3 during acquisition.

Startle blink latency – Extinction

As shown in figure S3, blink latency slows gradually across blocks in the low intensity group, and slows between blocks 1 and 2, then increases between blocks 2 and 3 in the high intensity group. A main effect of block, F (2, 48) = 14.327, p < .001, ηp^2 = .374, was qualified by a group x block interaction, F (2, 48) = 5.514, p = .007, ηp^2 = .187. Follow-up analyses revealed no differences between blocks in the low intensity group, F (2, 48) = 2.597, p = .085, ηp^2 = .098, and faster responses for block 1 and 3 than block 2, t (49) = 5.62, p < .001, d = 0.79, and t (49) = 4.15, p <

.001, d = 0.58, respectively, F (2, 48) = 18.836, p < .001, ηp^2 = .440, in the high intensity group.

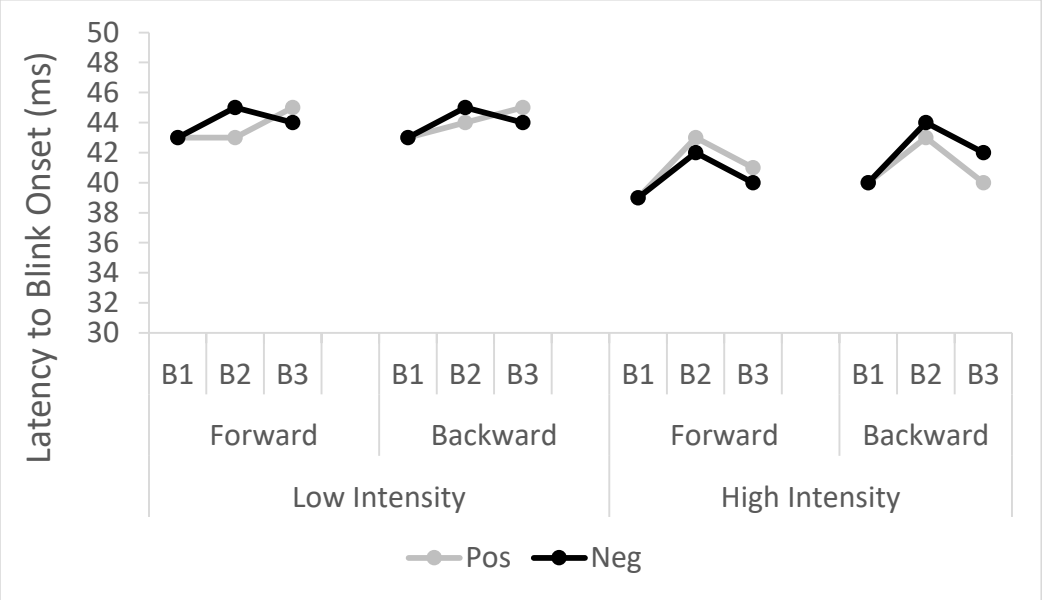

Figure S3. Time until blink onset following startle probe in milliseconds, during forward and backward CSs paired with positive and negative USs, in the low and high intensity groups across blocks 1, 2, and 3 during extinction.

Affective priming – Errors

Figures S4 below suggests assimilation effects for forward conditioning in both groups, and for backward conditioning in the low intensity group. In the high intensity group for backward conditioning, a contrast effect is suggested. The group x conditioning type interaction was significant, F (1, 59) = 9.257, p = .003, ηp^2 = .136. Follow-up analyses showed that priming scores in the low intensity group were significantly greater than 0 for backward conditioning, t (29) = 3.181, p = .003, d = 0.58, but not forward conditioning, t (29) = 1.511, p = .142, d = 0.28, and that forward and backward conditioning did not differ significantly from each other, t (28) = 0.876, p = .384, d = 0.16. In the high intensity group, follow-up analyses showed that forward conditioning was greater than 0, t (30) = 3.437, p = .002, d = 0.62, while backward conditioning was not, t (30) = 1.070, p = .293, d = 0.19, and that forward and backward conditioning differed significantly from each other, t (29) = 3.448, p = .001 d = 0.62.

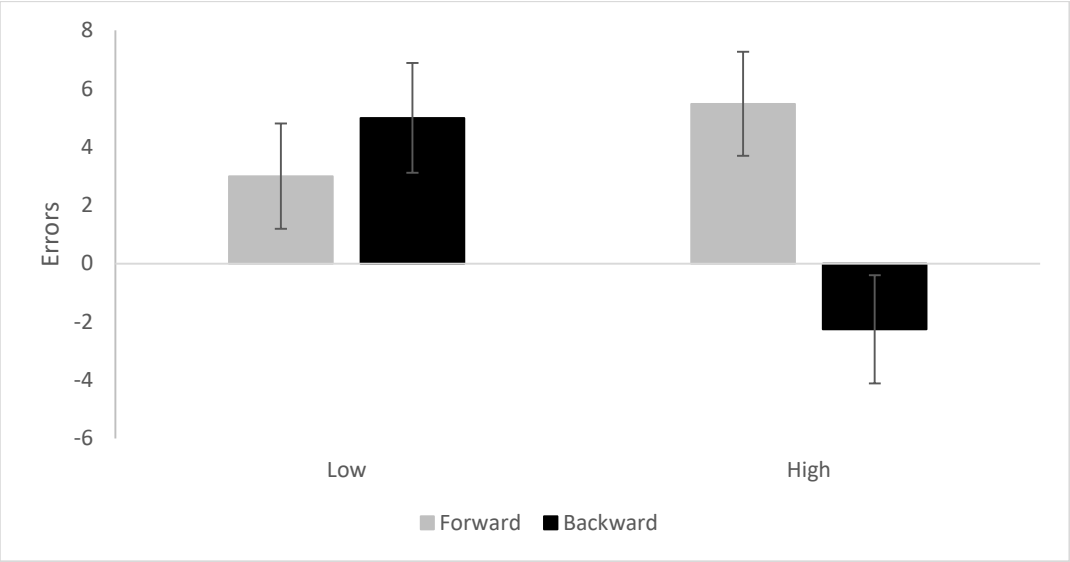

Figure S4. Priming scores for errors from the affective priming task for forward and backward conditioning in the low and high intensity groups. Error bars show 95% confidence intervals for the mean.

Skin conductance responding

Self-adhesive isotonic electrodes were attached to the thenar and hypothenar eminences of the non-preferred hand to record SCRs throughout the experiment. Responses were amplified at a gain of 5 μSiemens per volt by a Biopac MP150 system and recorded using AcqKnowledge 4.1.0 at a sampling rate of 1000 Hz. Respiration was measured by fitting a respiration belt around the participant's waist to control for SCR artefacts. SCR's were scored offline using AcqKnowledge 4.1.0. Responses were square root transformed and range correct by dividing each participant's response by their largest response, to reduce the skew of the data prior to analysis. Only first interval responses (1-4s) during forward CSs and US onset were analysed, as startle probes and CS/US overlap precluded any meaningful analysis of second interval responding or responding during backward CSs. Responses were then aggregated into five blocks, with each block containing the average of two consecutive trials. These data during forward CSs and US onset were then subjected to separate 2 (group: low intensity vs high intensity) x 2 (US valence: positive vs negative) x 5 (block: 1, 2, 3, 4, 5) mixed model ANOVA's. Four participants from the low intensity group and five participants from the higher intensity group were removed for being non-responders.

Skin conductance responding – First interval – Forward CS

Figure S5 shows habituation across blocks, and larger responses to CSs paired with negative USs regardless of group. This was confirmed by a main effect of US valence, F (1, 53) = 5.658 p = .021, ηp^2 = .096, showing larger responses to CSs paired with negative USs over CSs paired with positive USs, and a main effect of block, F (4, 50) = 23.861, p < .001, ηp^2 = .656, which shows larger responses at block 1 than blocks 2, t (54) = 9.84, p < .001, d = 1.33, 3, t (54) = 8.11, p < .001, d = 1.09, 4, t (54) = 6.94, p < .001, d = 0.94, and 5, t (54) = 6.93, p < .001 d = 0.93. While it may look like group interacts with valence, this was not the case, Group × Valence, F (1, 53) = 2.197 p = .144, ηp^2 = .040, and Group × Valence × Block, F (4, 50) = 0.381 p = .821, ηp^2 = .030.

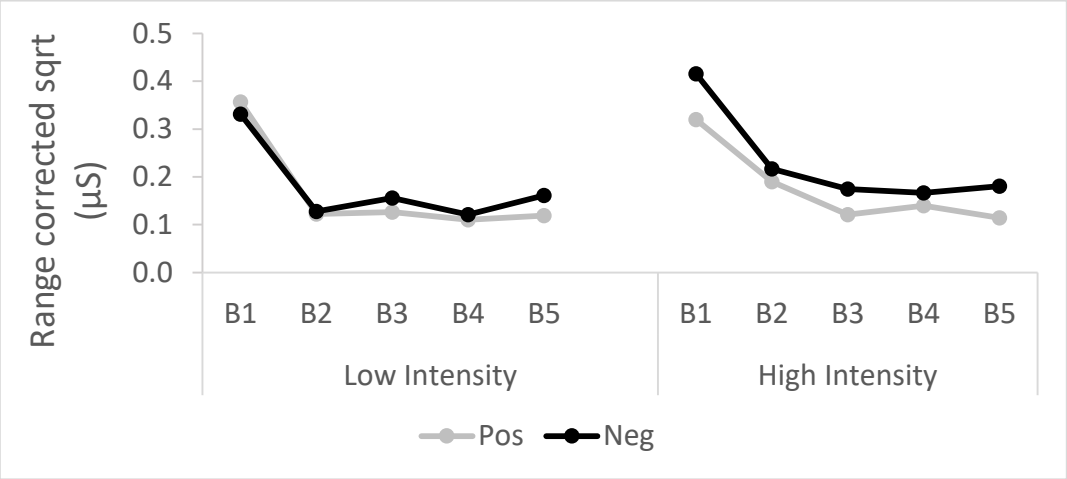

Figure S5. First interval skin conductance responses to CSs paired with positive and negative USs during acquisition, presented in blocks of 2 averaged responses per block (blocks 1, 2, 3, 4, and 5), for the low intensity and high intensity groups.

Skin conductance responding – First interval – US onset

Figure S6 shows habituation to the US, with larger responses to the negative US in the high intensity group. This was confirmed by a main effect of block, F (4, 50) = 14.267, p < .001, ηp^2 = .533, and a group x US valence interaction, F (1, 53) = 5.111, p = .028, ηp^2 = .088. Follow-up analyses revealed no differences between positive and negative USs in the low intensity group, F (1, 53) = 0.151 p = .669, ηp^2 = .003, and larger responses to the negative US than the positive US in the high intensity group, F (1, 53) = 7.764 p = .007, ηp^2 = .128.

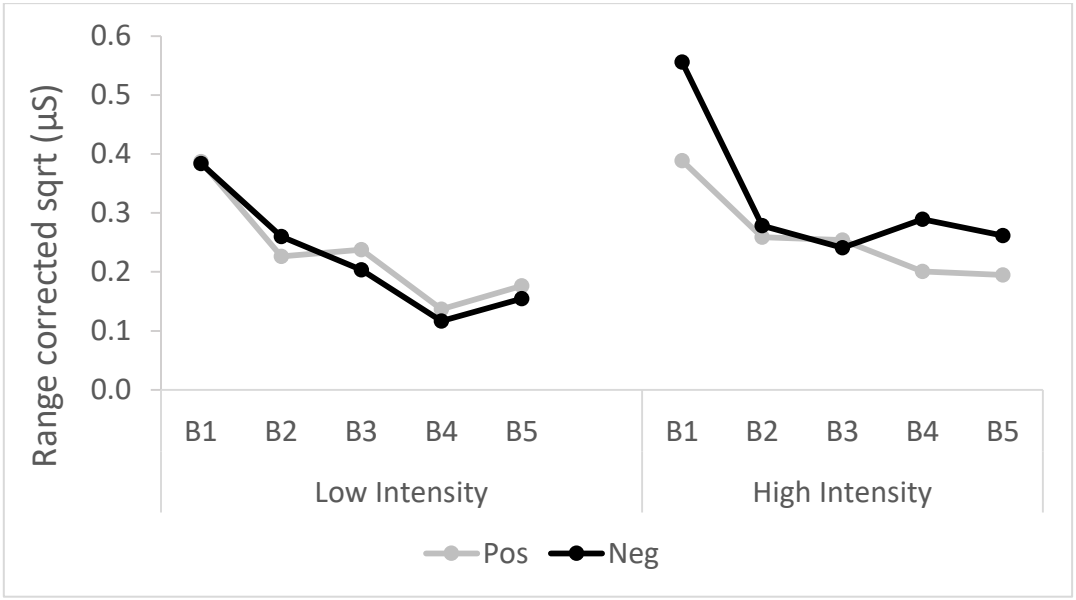

Figure S6. First interval skin conductance responses positive and negative USs during acquisition, presented in blocks of 2 averaged responses per block (blocks 1, 2, 3, 4, and 5), for the low intensity and high intensity groups.

3.10 Supplementary Material – Experiment 2

Startle blink magnitude – Extinction

Figure S7 shows a decrease in response from blocks 1 to 2. This is confirmed by the tests of within-subjects contrasts which showed a linear trend for block, F (1, 36) = 13.91, p = .001, ηp^2= .279.

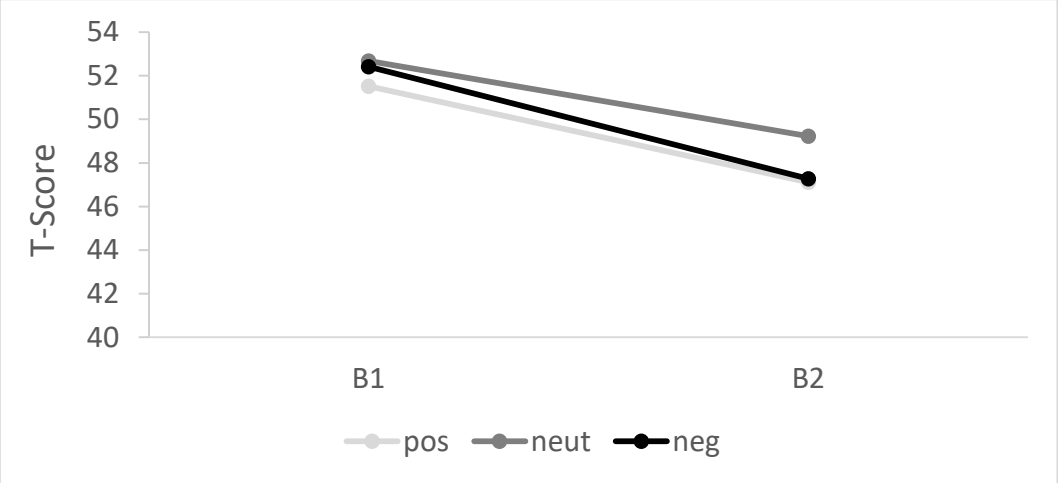

Figure S7. Startle blink magnitude (*T*-scores) by block (1 and 2) during extinction, for CSs that were paired with positive, neutral, and negative USs following backward conditioning.

Startle blink latency – Acquisition

Figure S8 shows faster blink onset during CSs paired with the positive US than CSs paired with the neutral US, and CSs paired with the negative US. This was confirmed by the tests of within-subject contrasts which showed a linear trend for US valence, F (1, 33) = 19.801, p < .001, ηp^2= .375

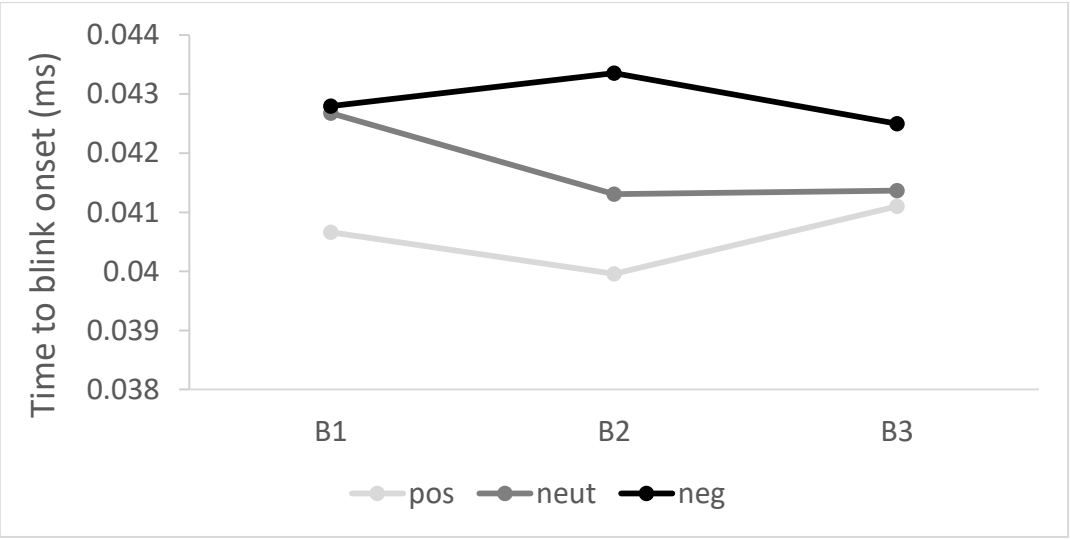

Figure S8. Time until blink onset following startle probe in milliseconds, during backward CSs paired with positive, neutral, and negative USs, across blocks 1, 2, and 3 during acquisition.

Startle blink latency – Extinction

Figure S9 suggests slower blinks to the CS paired with the negative US compared with CSs paired with neutral and positive USs at block 1, with faster responses to CSs paired with the negative US at block 2. However, no tests of within-subjects contrasts were significant, *F*s < 0.738, *ps* > .396, $\eta p^2 s$ < .021.

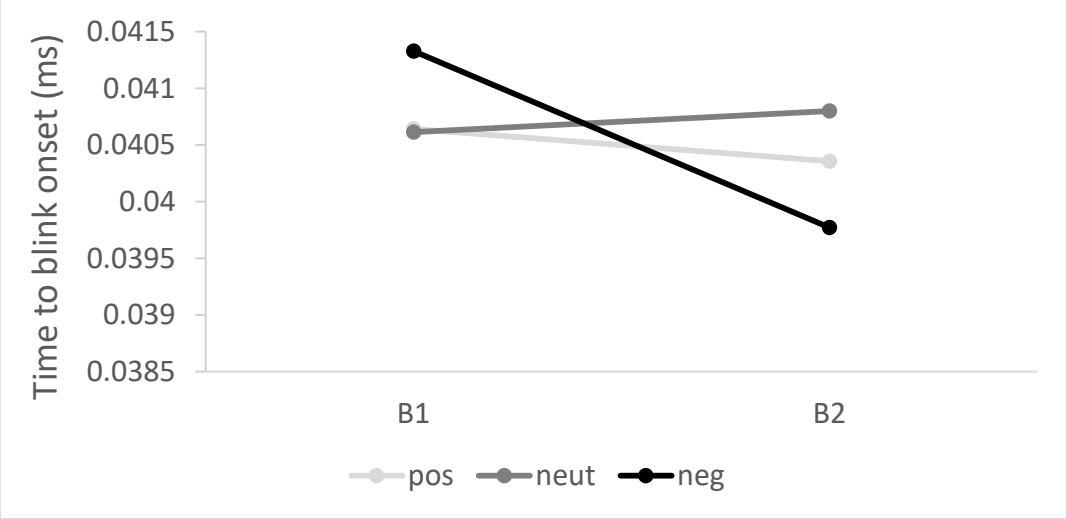

Figure S9. Time until blink onset following startle probe in milliseconds, during backward CSs paired with positive, neutral, and negative USs, across blocks 1 and 2 during extinction.

Affective priming – Errors

Figure S10 suggests a quadratic trend, however neither the quadratic, F (1, 37) = 1.488, p = .230, ηp^2= .039, or linear, F (1, 37) = 1.716, p = .198, ηp^2= .044, trend analyses were significant.

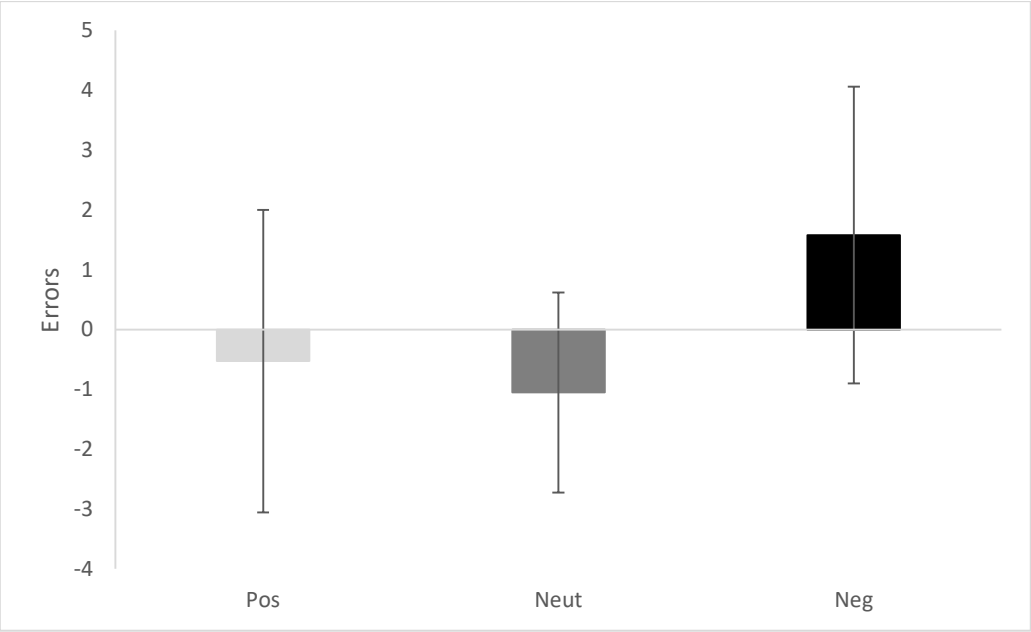

Figure S10. Difference scores (positive target words – negative target words) for errors from the affective priming task for CSs paired with positive, neutral, and negative USs. Error bars show 95% confidence intervals of the mean.

Skin Conductance Responding – US Onset – First Interval Response

Figure S11 shows skin conductance responses decreasing across blocks, with larger skin conductance responses to the negative US than the neutral and positive USs. This was confirmed by main effects of US valence, F (2, 35) = 15.827, p < .001, ηp^2 = .475, and block, F (3, 34) = 13.188, p < .001, ηp^2 = .538, and a significant US valence x block interaction, F (6, 31) = 4.86, p < .001, ηp^2 = .485. At block's 1 and 2, responses to USneg were greater than responses to USpos and USneut, F (2, 35) = 23.568, p < .001, ηp^2= .574, and F (2, 35) = 6.774, p = .003, ηp^2= .279. No differences in responses size were found between US valence at block 3, F (2, 35) = 1.564, p = .224, ηp^2= .082, and block 4, F (2, 35) = 0.947, p = .398, ηp^2= .051.

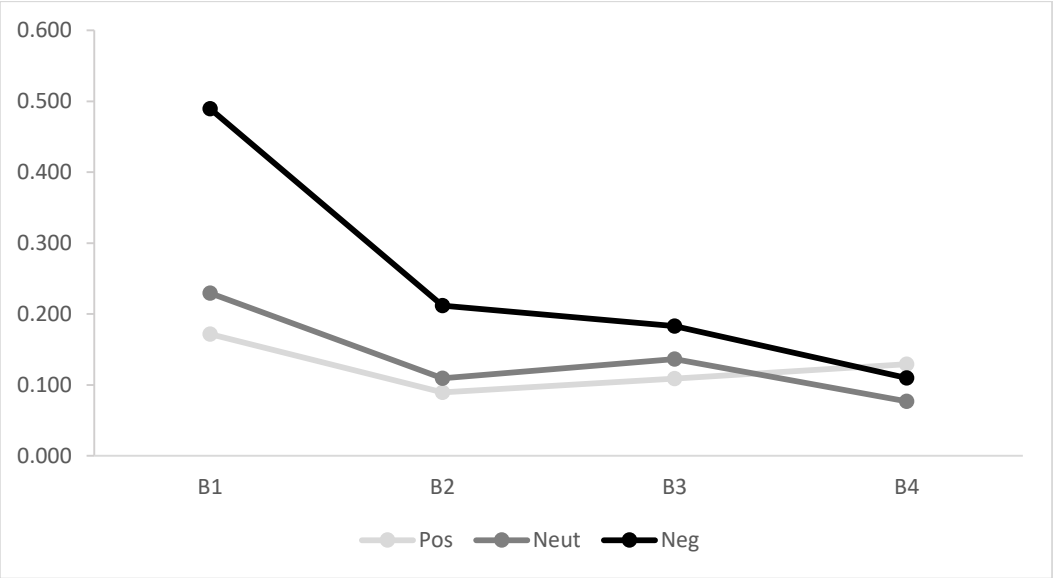

Figure S11. First interval skin conductance responses to positive, neutral, and negative USs during acquisition, presented in blocks of 2 averaged responses per block (blocks 1, 2, 3, and 4).

Chapter 4: Startle during backward evaluative conditioning is not modulated by instructions

Luke J. S. Green
Curtin University

Camilla C. Luck
Curtin University

Ottmar V. Lipp
Curtin University

Author Notes

This work was supported by an Australian Government Research Training Program Scholarship to Luke Green and grants DP180111869 and SR120300015 from the Australian Research Council to Ottmar Lipp.

All data and materials (excluding sounds from the IADS database) are available at https://osf.io/8xu25/.

Correspondence concerning this article should be sent to: Luke J S Green, School of Psychology, Curtin University, GPO Box U1987 Perth WA 6845, Australia. Email: luke.green2@postgrad.curtin.edu.au.

4.1 Abstract

Instructions highlighting that backward conditional stimuli (CS) stop unconditional stimuli (US) result in their acquiring valence opposite to that of the US on explicit measures of valence. We assessed whether such instructions would influence startle blink modulation in the same way. Two groups were presented with concurrent forward and backward evaluative conditioning (CS-US-CS) using cartoon aliens as CSs, and pleasant, neutral, and unpleasant sounds as USs. Startle magnitude was measured during conditioning and valence ratings were assessed after conditioning. Participants in the 'start-stop' instructions group (n = 41) were instructed to learn whether CSs started or stopped US presentations, while participants in the 'observe' instructions group (n = 41) were told to pay attention to the stimuli as they would be asked questions about them after the experiment. In the start-stop instructions group backward CSs paired with positive USs were rated as less pleasant than backward CSs paired with neutral and negative USs (contrast effect) whereas ratings of backward CSs did not differ in the observe instructions group. Startle magnitude was larger during backward CSs paired with positive USs in comparison to CSs paired with neutral or negative USs in both instruction groups. Startle blink modulation was unaffected by instructions, suggesting that startle indexes the emotional state at the time of probe presentation rather than CS valence based on propositional information about the function of the CS.

Keywords: Evaluative conditioning, startle modulation, backward conditioning, attitudes, propositional learning

Valence refers to how pleasant or unpleasant a given stimulus is and can be acquired and changed via fear and evaluative conditioning (Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010; Lipp, 2006). In these procedures, neutral stimuli are paired with valenced unconditional stimuli (USs) resulting in the valence of the now conditional stimulus (CS) changing. The direction in which valence changes (i.e. towards pleasant or unpleasant) can depend on the temporal relationship between CSs and USs (forward vs. backward conditioning), the propositional relationship between CSs and USs (i.e. are CSs and USs 'friends' or 'enemies' or does the CS 'start' or 'stop' the US), and the measures used to assess valence change (Andreatta, Mühlberger, Yarali, Gerber, & Pauli, 2010; Fielder & Unkelbach, 2011; Mallan, Lipp, & Libera, 2008; Moran & Bar-Anan, 2013). The influence of these factors on valence change has important implications for our understanding of many areas of psychology, such as human emotional learning, attitude formation, and prejudice (Corneille & Stahl, 2019; De Houwer, Thomas, & Baeyens, 2001; Olson & Fazio, 2006). Moreover, our likes and dislikes influence decision making from the trivial to the profound and it has been suggested that residual negative valence following successful anxiety disorder treatment can lead to relapse of fear in those with anxiety disorders (Dirikx, Hermans, Vansteenwegen & Baeyens, 2004; Galdi, Arcuri, & Gawronski, 2008; Gibson, 2008; LeBel & Campbell, 2009; Zbozinek, Hermans, Prenoveau, Liao & Craske, 2015). Understanding the processes that underlie the acquisition of valence is a necessary endeavour in contemporary psychology's quest to decrease psychopathology and increase acceptance of individuals from all walks of life.

The acquisition of valence can be measured by means of self-report, where participants are asked to rate how pleasant or unpleasant they find a given stimulus (i.e. explicit valence ratings). Explicit valence ratings following forward (CS-US) and backward (US-CS) conditioning without instructions about CS/US relations reveal that the CS acquires valence in the direction of the US (an assimilation effect; Green, Luck, Gawronski, & Lipp, 2019; Mallan et al., 2008; see Hofmann et al., 2010 for a review and meta-analysis). When instructions highlight that each set of CSs either starts or stops a pleasant or unpleasant US, backward CSs presented with positive USs are rated as less pleasant than backward CSs presented with negative USs (a contrast effect; Moran & Bar-Anan, 2013; Green et al., 2019). Moran and Bar-Anan (2013) demonstrated this in a concurrent forward and backward evaluative

conditioning paradigm (CS-US-CS) employing pleasant and unpleasant sound USs, and Green et al. (2019) confirmed that the instructions highlighting the roles of the CSs were required for backward CS contrast effects to emerge in a picture-picture paradigm. Contrast effects have also been demonstrated following forward conditioning when information that changes the relationship between the CS and US is presented. For example, a CS that 'loathes' a negative US is rated more positively than a CS that 'loves' a negative US, even though the US is the same (Förderer & Unkelbach, 2012; Unkelbach & Fiedler, 2016). Instructional manipulations are robust in eliciting contrast effects, however, as will be discussed below, contrast effects can occur in certain circumstances without an instructional manipulation (Andreatta, Mühlberger, Glotzbach-Schoon, & Pauli, 2013). Moreover, the effect of instructional manipulations on other measures of valence acquisition is currently unknown.

In addition to explicit valence ratings, valence acquisition can also be measured using the startle blink reflex. The startle reflex is a component of a defensive response displayed by humans and other animals that can be elicited by a loud noise probe (Blumenthal et al., 2005). The magnitude of the eye blink response that occurs as part of this reflex can be reliably measured using EMG and provides an index of emotional state as it is potentiated during aversive states and inhibited during appetitive states (Bradley, Cuthbert, & Lang, 1990; Vrana, Spence, & Lang, 1988; but see Lipp, Siddle, & Dall, 2003). This generally means that startles elicited during CSs presented *before* aversive USs (forward conditioning) will be larger than startles elicited during CSs presented with neutral or appetitive USs, akin to an assimilation effect (Andreatta et al., 2010; Andreatta et al., 2013, Green, Luck, & Lipp, 2020a; Luck & Lipp, 2017, but see also Mallan et al., 2008). However, when CSs are presented *after* aversive USs (backward conditioning), a different pattern emerges (Andreatta et al., 2010; Andreatta et al., 2013; Green et al., 2020a; Luck & Lipp, 2017). Andreatta et al. (2010) used the startle blink reflex to demonstrate in humans that backward conditioning with aversive electro-tactile stimuli results in 'relief learning' as the CS signals the end of the aversive US. This is akin to a contrast effect, as CSs presented *after* aversive USs acquire positive properties, shown by attenuated startle responses in comparison to unpaired CSs (CS-). Explicit valence ratings, on the other hand, revealed an assimilation effect, though when a

concurrent forward and backward conditioning paradigm (CS-US-CS) was employed by Andreatta et al. (2013), a backward CS contrast effect was found.

Green et al. (2020a) extended the work by Andreatta et al. (2013) employing both aversive and appetitive sound USs in a concurrent forward and backward conditioning experiment (CS-US-CS; Experiment 1), and aversive, neutral, and appetitive sound USs in a backward conditioning experiment (US-CS; Experiment 2). Startle responses were larger during CSs following appetitive USs than aversive USs in Experiment 1 and larger during CSs following appetitive USs than aversive and neutral USs in Experiment 2. These contrast effects suggest that 'disappointment learning' occurs from pairing a CS with the offset of an appetitive stimulus, much like 'relief learning' occurs when pairing a CS with the offset of an aversive stimulus. However, differing from Andreatta et al. (2010; 2013), an instructional manipulation adapted from Moran and Bar-Anan (2013) was employed which highlighted that the CSs started or stopped (or stopped only in Experiment 2) the USs. Moreover, Green et al. (2020a) used sound USs, which are arguably less intense than the electro-tactile USs used by Andreatta et al. (2013). The fact that, in absence of explicit instructions, Green et al. (2019) found no evidence of relief and disappointment learning on explicit ratings using picture USs while Andreatta et al. (2013) did when using an electro-tactile US, suggests that a certain threshold of US intensity may be required for relief and disappointment learning to occur. Thus, it is currently unknown whether the backward CS contrast effect reported by Green et al. (2020a) for explicit ratings and startle modulation is a true reflection of relief and disappointment learning, or the result of the instructional manipulation highlighting that the role of each set of CSs is to either start or stop the US.

The aim of the current study was to assess the influence of instructional manipulations highlighting that each set of CSs either starts or stops USs on the startle blink reflex. A concurrent forward and backward conditioning procedure employing aversive, appetitive, and neutral USs and six different sets of cartoon aliens as CSs was presented to a 'start-stop instructions' group and an 'observe instructions group'. In the 'start-stop instructions' group participants were told that each set of CSs either started or stopped the USs. In the 'observe instructions' group participants were told to pay attention to the stimuli as they would be asked questions about them at the end. Explicit valence ratings were measured after conditioning and startle blink magnitude was measured during conditioning. Affective priming was

also measured to ensure that procedural features remained constant with earlier experiments using the research protocol. However, as affective priming did not additional information pertinent to the proposed research question, affective priming results are reported in the supplement. We predicted backward CS contrast effects in the start-stop instructions group and backward CS assimilation effects in the observe instructions group and assimilation effects for forward conditioning in both groups on the startle blink reflex and explicit valence ratings.

4.2 Method

4.2.1 Participants

Eighty five undergraduate students from Curtin University participated in this experiment for course credit following ethical approval from the Curtin University Human Research Ethics Committee. Participants were unable to participant if they were pregnant or suffering from cardiac or seizure related disorders, but were not screened with regard to psychopathology. Three participants were excluded as they had participated in a previous study that employed a similar conditioning procedure. The final sample consisted of 82 students (55 female), *M* age = 20.83, *SD* = 3.73, with 41 participants per group. The sample size was based on previous research by Andreatta et al. (2013), Moran and Bar-Anan (2013), and Moran, Bar-Anan, and Nosek (2016), which yielded large effect sizes ranging between ηp^2 = .15 and .60 for their within-participant interactions of interest with samples ranging from 28 to 68 participants. Based on these studies, we predicted that 40 participants per group would provide sufficient power to detect the effects of interest. Two participants in the start-stop instructions group and 27 participants in the observe instructions group failed the recollective memory test (i.e. incorrectly stated the relationship between CSs and USs for one or more contingencies). This is most likely because the start-stop instructions emphasise that there is a relationship between CSs and USs, whereas the observe instructions do not. Analyses were run with and without these participants, however, results will be reported from the entire sample as the results from only those who could verbalise the contingencies did not deviate in a meaningful way.

4.2.2 Apparatus/Stimuli

Four families of four aliens from Moran and Bar-Anan (2013; available at https://osf.io/cqsnj/), and two additional families of four aliens created for this

experiment were used as CSs (all 960 x 720 pixels). The CS families differed in colour and head shape. The positive and negative USs were also taken from Moran and Bar-Anan (2013); the pleasant guitar melody from the start of 'The Shape of My Heart' by Sting (USpos; 47dBA), and an aversive human scream (USneg; 72dBA). The neutral US was the sound of a train passing a train station (sound #425 from IADS-2; 56dBA). A Biopac MP150 system was used to record Orbicularis Oculi electromyogram (EMG), skin conductance responding, and respiration, using AcqKnowledge Version 4.1 at a sampling rate of 1000Hz. Two 4mm Ag/AgCl cup electrodes filled with electrode gel and attached using double-sided adhesive electrode collars were used to measure Orbicularis Oculi EMG. Three sets of electrodes from different brands were used, although electrodes were never mixed across brands (Med-associates, Discount disposable, and Gereonics). One electrode was attached to the area of skin under the pupil of the left eye, and the second under the corner of the left eye. Impedance was checked to ensure conductivity, but as impedance changes over time the number was not recorded. To ensure that conductivity was as good as possible without abrading the skin, participants washed the areas where electrodes were to be attached with water and non-alcohol based soap. As it is difficult to achieve an impedance of less than 10 k/ohms without abrading the skin, our criterion for adequate contact was that upon initial measurement (about 30 seconds after electrode attachment) an impedance of less than 100 k/ohms was observed and that the impedance was steadily decreasing. In the case of a reading above 100 k/ohms or a stagnant reading, electrodes were removed and the process from cleaning the areas to checking electrode contact was repeated. This repetition only happened for one participant, and upon repeating the procedure electrode contact was confirmed using the above standard. We also presented three habituation startle probes and observed the EMG responses before continuing with the experiment to ensure adequate signal to noise ratio. The startle eliciting stimulus was a 50ms 105dBA white noise burst with near instantaneous rise time presented via a custom built noise-generator. A Biopac EMG100C amplifier was used to amplify the EMG signal at a gain of 5000. The high pass filter was set to 10 Hz, and the low pass filter to 500 Hz. Two self-adhesive isotonic Biopac EL507 electrodes were attached to the thenar and hypothenar eminences of the non-dominant hand to measure electrodermal responding. These responses were DC amplified at a gain of 5 µSiemens per Volt by a Biopac EMG100C amplifier. To

control for breathing and movement related artefacts in electrodermal responding, respiration was measured by a chest gauge (Biopac TSD-201). Recording skin conductance responses is standard practice in psychophysiological experiments in our laboratory. However, as we used a higher ratio of probed to non-probed trials than in previous studies measuring startle modulation and skin conductance responding concurrently, we did not analyse skin conductance responses due to contamination by the startle probes. Ten positive words (*pleasant, good, outstanding, beautiful, magnificent, marvellous, excellent, appealing, delightful, and nice*) and 10 negative words (*unpleasant, bad, horrible, miserable, hideous, dreadful, painful, repulsive, awful, and ugly*) taken from Hu, Gawronski, and Balas (2017a, 2017b) were used as target stimuli in the affective priming task (Calibri 11 point font), and the CSs were used as primes. Presentations of acoustic and visual stimuli during the main experiment and the explicit valence ratings task, the affective priming task, and the memory test, generation of event markers, and recording of speeded responses and ratings were controlled by DMDX (Forster & Forster, 2003). A 24inch BenQ LED monitor at a resolution of 1920 × 1080 with a refresh rate of 60 Hz was used to show visual stimuli. Auditory USs and startle probes were presented with Sennheiser HD-25-1 headphones.

4.2.3 Procedure

Participants signed up for the study using Curtin University's online School of Psychology Research Participation Scheme. Participants read the information sheet upon arrival to the lab and were played each US for 30 seconds before providing informed consent. After signing the consent form, participants washed and dried the area under their left eye and their hands, with soap and water. Next, the recording equipment was attached, three habituation startles were presented, and a three minute baseline recording of skin conductance responding was taken. Participants were then told about the experiment. The start-stop instructions group was told to learn which alien family was responsible for starting the human, musical, and metropolitan sounds, and which alien family was responsible for stopping these sounds. They were also told that they would be tested on this at the end (full instructions available at https://osf.io/8xu25/). The observe instructions group was told that families of aliens would be presented with human, musical, and metropolitan sounds, and that they needed to learn which creatures were presented with which sounds for a later memory test. The researcher started the script and the

relevant instructions were presented on the screen. Participants then completed the acquisition phase, which was based on the CS-US-CS procedure used by Moran and Bar-Anan (2013; see Figure 1A). In this procedure, eight positive US trials, eight negative US trials, and eight neutral US trials were presented. CSs were pictures of cartoon aliens presented for 8s each, with a 2s overlap between CS-US and US-CS. The CSs were six families of cartoon aliens differing in head shape and colour, with each family containing four aliens. Each CS from each family was presented twice and CSs were counterbalanced between-participants. USs were presented for 10, 15, 20, 25, or 30s, with a pleasant melody as the positive US, a human scream as the negative US, and the sound of a train passing by as the neutral US. Startles were presented at 4.5 or 5.5s after forward CS onset, and 6.5 and 7.5s after backward CS onset (4.5 or 5.5s after US offset; see Figure 1B). Startle probes were presented during each forward and backward CS on six out of eight trials, totalling 36 probes. These probes were assigned randomly within forward and backward CSs separately. The inter-trial intervals were 12, 14, or 16s, and startle probes were presented half-way through the interval on half of the trials, for a total of 12 startle probes. After acquisition, participants were told to rate how much they liked each family of creatures (CSs). Participants were shown each family of aliens (four aliens per family) separately and asked to provide a rating of how much they liked each family on a scale from 1 = don't like at all, to 9 = like a lot. The experimenter then explained the affective priming task, and after 24 practice trials, participants completed the main affective priming task. In this task, a 500ms fixation cross was presented, followed by a 200ms CS prime, and then the target word until the participant responded by pressing the right 'SHIFT' key if the target word was positive, and the left 'SHIFT' key if the target word was negative. The prime stimuli were the four alien creatures from the six families that were used as CSs. CS families were each presented with 10 good words and 10 bad words for a total of 120 trials. Two aliens from each family were presented with each good word and two aliens from each family were presented with each bad word. Both reaction time and errors in categorising target words were measured. Next, the experimenter told participants the memory test would follow. Participants in the start-stop instructions group were shown each family separately, and asked "What was the role of the creatures in this picture? 1. Started the human sound. 2. Stopped the human sound. 3. Started the musical sound. 4. Stopped the musical sound. 5. Started the metropolitan sound. 6.

Stopped the metropolitan sound". Participants in the observe instructions group were shown each family separately, and asked "What was the relationship between the creatures in this picture and the sounds? 1. Preceded the human sound. 2. Followed the human sound. 3. Preceded the musical sound. 4. Followed the musical sound. 5. Preceded the metropolitan sound. 6. Followed the metropolitan sound". The recording equipment was then disconnected, and participants were asked to fill out a post-experimental questionnaire including US valence and arousal ratings, and demographic questions. Age, gender, and ethnicity were recorded. Participants were asked to rate how pleasant/unpleasant the three US sounds and the startle probes were (7-point scale ranging from -3 = very unpleasant to 3 = very pleasant), how intense the three US sounds were (7-point scale from 0 = not at all to 6 = very intense), and how startling the startle probes were (7-point scale from 0 = not at all to 6 = very startling). Participants were also asked if they were colour-blind. Participants were debriefed and thanked for their time. The entire experiment took approximately 1 hour.

4.2.4 Scoring, response definition, and statistical analyses

Mixed-model ANOVAs ($\alpha = .05$, Pillai's trace reported) were used to analyse data in IBM SPSS Statistics 25. Pairwise comparisons were used to follow-up significant interactions and Bonferroni-Holm corrections were applied (Holm, 1979). To limit the length of the main paper, startle latency data are reported in the supplementary materials. As stated earlier, skin conductance response data were not analysed due to contamination by the startles probes presented during the majority of the CSs. Affective priming data are reported in the supplement as they did not provide additional information pertinent to the proposed research question.

Startle blink magnitude. Raw EMG signals were notch-filtered at 50 Hz, and high and low passed filtered at 30 and 500 Hz, before being rectified and smoothed by calculating a five point moving average. Startle blink magnitude was visually defined as the largest peak within a 120ms window from probe onset, so long as the response began 20-60ms after probe onset (Blumenthal et al., 2005). Trials where response onset could not be visually identified in this window were classified as non-response trials, scored as zero, and included in the analysis. Trials where responses could not be differentiated from excessive background EMG activity or where spontaneous blinks occurred between the startle probe onset and response window onset were scored as missing (6.37%). Trials were aggregated by

US valence for forward and backward conditioning separately and transformed into *T*-scores to control for individual differences and variation across individual trials. *T*-scores were then subjected to a 2 (Instructions: start-stop vs. observe; between-participants) × 2 (Conditioning Type: forward vs. backward; within-participant) × 3 (US Valence: positive vs. neutral vs. negative; within-participant) mixed model ANOVA. 17 participants (start-stop instructions = 8, observe instructions = 9) performed the experiment while landscaping was occurring outside of the building which may have been distracting due to the acoustic noise produced by the workers. This may have resulted in slightly more zero responses in the sample including these participants (7.7%) than the sample without these participants (6.2%), while missing responses were comparable (6.37% vs 6.07%, respectively). However, the likelihood that this would have any systematic effect on the results is low considering that a similar number of participants from each group were affected. Moreover, the analyses were run excluding these participants and the pattern of results remained the same. Thus, data from the entire sample have been reported.

Explicit valence ratings. Participants rated how much they liked each family of CSs, with more positive valence denoted by a higher rating. These data were subjected to a 2 (Instructions: start-stop vs. observe; between-participants) × 2 (Conditioning Type: forward vs. backward; within-participant) × 2 (US Valence: positive vs. neutral vs. negative; within-participant) mixed model ANOVA. One participant from the observe group was excluded from the ANOVA due to missing data.

4.3 Results

Manipulation checks. Groups did not differ in gender, $\chi 2$ (1, N = 82) = .497, p = .481, ethnicity, $\chi 2$ (7, N = 82) = 8.93, p = .257, or age, t (80) = 1.13, p = .264, d = 0.25. US and startle probe valence ratings were subjected to a 2 (Instruction: start-stop vs. observe; between-participants) × 4 (Valence: positive vs. neutral vs. negative vs. startle probe; within-participant) mixed model ANOVA. A main effect of Valence, F (3, 77) = 386.24, p < .001, ηp^2 = .94, was qualified by a Group × Valence interaction, F (3, 77) = 3.72, p = .015, ηp^2 = .13. Follow-up analyses revealed that in the start-stop and observe instruction groups the positive US was rated as more positive than the neutral US, t (40) = 10.49, p < .001, d = 1.64, and, t (39) = 7.99, p < .001, d = 1.26, the negative US, t (40) = 24.56, p < .001, d = 3.84, and, t (39) =

21.98, $p < .001$, $d = 3.48$, and the startle probe, t (40) = 20.08, $p < .001$, $d = 3.14$, and, t (39) = 20.33, $p < .001$, $d = 3.21$, and that the neutral US was rated as more positive than the negative US, t (40) = 12.43, $p < .001$, $d = 1.94$, and, t (39) = 12.59, $p < .001$, $d = 1.99$, and the startle probe, t (40) = 9.98, $p < .001$, $d = 1.56$, and, t (39) = 13.09, $p < .001$, $d = 2.07$. The Group × Valence interaction was driven by the fact that in the start-stop instructions group the negative US was rated as significantly more negative than the startle probe, t (40) = 4.34, $p < .001$, $d = 0.68$, while in the observe instructions condition this difference was not significant, t (39) = 0.65, $p = .518$, $d = 0.10$.

US and startle probe intensity ratings were also subjected to a 2 (Instructions: start-stop vs. observe; between-participants) × 4 (Valence: positive vs. neutral vs. negative vs. startle probe; within-participant) mixed model ANOVA. A main effect of valence, F (3, 78) = 33.68, $p < .001$, $\eta p^2 = .56$, and a marginal Group × Valence interaction, F (3, 78) = 2.43, $p = .071$, $\eta p^2 = .09$, were detected. The USpos was rated as less intense than the USneut, t (40) = 2.86, $p = .005$, $d = 0.45$, the USneg, t (40) = 8.11, $p < .001$, $d = 1.27$, and the startle probe, t (40) = 9.79, $p < .001$. $d = 1.53$, The USneut was also rated as less intense than the USneg, t (40) = 6.82, $p < .001$, $d = 1.06$, and the startle probe, t (40) = 8.53, $p < .001$, $d = 1.33$. The USneg and startle probe did not differ from each other, t (40) = 1.55, $p = .125$, $d = 0.24$. Startle magnitude during the inter-trial intervals did not differ between groups, t (80) = 0.84, $p = .405$, $d = 0.19$.

4.3.1 Explicit valence ratings

Figure 2 suggests an assimilation effect for forward conditioning in both groups and for backward conditioning in the observe instructions group, as CSs paired with positive USs appear more pleasant than CSs paired with neutral or negative USs. Moreover, contrast effects are suggested for backward conditioning in the start-stop instructions group, as CSs paired with negative USs appear more pleasant than CSs paired with neutral and positive USs. Main effects of conditioning type, F (1, 79) = 4.50, $p = .037$, $\eta p^2 = .05$, and US valence, F (1, 78) = 33.89, $p < .001$, $\eta p^2 = .47$, and a Conditioning Type × US Valence interaction, F (1, 78) = 40.26, $p < .001$, $\eta p^2 = .51$, were qualified by a Group × Conditioning Type × US Valence interaction, F (1, 78) = 4.32, $p = .017$, $\eta p^2 = .10$. Follow up analyses confirmed the assimilation effect for forward conditioning as CSs paired with positive USs were rated as more pleasant than CSs paired with neutral USs, ps <

.001, and CSs paired with neutral USs were rated as more pleasant than CSs paired with negative USs, *p*s < .001, for the start-stop instructions group, F (2, 78) = 36.90, $p < .001$, $\eta p^2 = .49$, and the observe instructions group, F (2, 78) = 27.22, $p < .001$, $\eta p^2 = .41$. A contrast effect was suggested for backward CSs in the start-stop instructions group, F (2, 78) = 6.21, $p = .003$, $\eta p^2 = .14$, as backward CSs presented after positive USs were rated as more unpleasant than CSs presented after neutral USs, $p = .005$, and negative USs, $p = .001$. CSs presented after neutral USs were rated as more unpleasant than CSs presented after negative USs, $p = .048$, however, when applying the Bonferroni-Holm correction the latter comparison was not significant. No backward conditioning was observed in the observe instructions group, *p*s > .120, F (2, 78) = 1.43, $p = .246$, $\eta p^2 = .04$.

4.3.2 Startle blink magnitude

Figure 3 suggests larger startle responses during CSs presented before negative USs, and to a lesser extent before positive USs, compared to CSs presented before neutral USs in both groups. Startle responses during CSs following positive USs appear larger when compared to responses during CSs following neutral and negative USs and responses during CSs following negative USs appear smaller when compared to responses during CSs following neutral USs in both groups. A main effect of conditioning type, F (1, 80) = 67.13, $p < .001$, $\eta p^2 = .46$, was qualified by a Conditioning Type × US Valence interaction, F (2, 79) = 15.22, $p < .001$, $\eta p^2 = .28$. The Conditioning Type × US Valence interaction indicated that for forward conditioning, blink magnitude was larger during CSs paired with negative USs than CSs paired with neutral USs, t (81) = 2.15, $p = .035$, $d = 0.24$, however this comparison was not significant when applying the Bonferroni-Holm correction. Blink magnitude during CSs paired with positive USs did not differ from blinks during CSs paired with neutral USs, t (80) = 0.27, $p = .786$, $d = 0.03$, or negative USs, t (80) = 1.73, $p = .088$, $d = 0.19$. For backward conditioning, blink magnitude was larger during CSs following positive USs when compared with CSs following negative USs, t (80) = 5.89, $p < .001$, $d = 0.65$. Blink magnitude was also larger during CSs following neutral USs when compared with CSs following negative USs, t (80) = 2.26, $p = .027$, $d = 0.25$, and larger during CSs following positive USs than CSs following neutral USs, t (80) = 2.32, $p = .023$, $d = 0.26$, however, when applying the Bonferroni-Holm correction the latter two comparison were not significant.

4.4 Discussion

The current experiment assessed whether instructions emphasising the CSs' role in controlling US presentations affect startle modulation in the same way as explicit valence ratings. A concurrent forward and backward conditioning procedure (CS-US-CS) with sound USs and pictures of cartoon aliens as CSs was presented to participants and startle blink magnitude was measured during forward and backward CSs. Explicit valence ratings were assessed after acquisition. In the start-stop group, participants were told that each set of CSs would either start or stop one of the sound USs and that they had to learn the role of each set of CSs for a later memory test. In the observe group, participants were told to pay attention to the CSs and USs as they would be asked questions about them at the end. In past research (Green et al., 2019), backward CS contrast effects were observed on explicit ratings after start-stop instructions whereas backward CS assimilation effects were observed after observe instructions. This finding was partially replicated in the current study in that for explicit valence ratings, backward CS contrast effects were observed in the start-stop instructions group and both groups showed assimilation effects for forward CSs. However, in the observe instructions group the backward CS assimilation effect was not significant. For blink startle, larger responses were observed during backward CSs paired with positive USs compared to negative USs, indicating backward CS contrast effects regardless of instructions. Startle responses during forward CSs were larger for CSs paired with the negative US compared to CSs paired with the neutral US before the Bonferroni-Holm correction, indicating a potential assimilation effect. However, responses during CSs paired with positive USs did not differ from the other two conditions. Thus, startle modulation is not affected by instructional manipulations in the same way as explicit valence ratings.

The current findings suggest that, in contrast to explicit valence ratings, startle magnitude during backward CSs is not moderated by instructional manipulations. Consistent with Green et al. (2020a), it seems that relief and disappointment learning at US offset are responsible for the observed backward CS contrast effects in startle. However, these results differ from those observed by Mallan et al. (2008). In this backward conditioning study, positive, neutral, and negative picture USs were presented before pictures of geometric shapes as CSs. Mallan et al. (2008) found startle potentiation during CSs backwardly paired with

both positive and negative USs in relation to neutral USs. This could suggest that both CSs had gained negative valence, a result not supported by the ratings of explicit CS valence which suggested an assimilation effect, or more likely, that startle was larger during stimuli that attracted attention, regardless of valence. When taken together, these findings suggest that either a certain threshold level of US intensity may be required for relief and disappointment learning to occur, or that the emergence of contrast or assimilation effects on startle modulation during backward conditioning varies with US modality. Future research should attempt to equate US intensity across modalities in order to investigate whether the modality of the US (sound vs picture vs shock etc.) or the intensity of the US is responsible for the discrepancy in results across these studies (Andreatta et al., 2013, 2017; Green et al., 2020a; Mallan et al., 2008).

One may argue that, rather than backward conditioning to the CSs, startle responding after US offset (and during the backward CSs) may reflect the emotion experienced due to the US ending (i.e. relief or disappointment). This seems unlikely as Andreatta et al. (2010; 2013) found evidence of relief learning during CSs presented without the US during extinction and during backward CSs presented after a 6s inter stimulus interval between US offset and CS onset. This suggests that even though it is possible that startle in our study reflects the emotion experienced after US offset, conditioning to the CS is likely to have occurred. A second explanation for the backward CS contrast effects observed for the startle blink reflex that does not rely on 'relief and disappointment' learning is that the startle blink reflex is sensitive to the difference in valence between the USs and the initially neutral backward CSs. In comparison to an aversive US, the neutral backward CS may seem relatively pleasant whereas in comparison to an appetitive US, a neutral backward CS may seem relatively unpleasant. These contrasts may result in startle inhibition during the backward CS following the aversive US and startle facilitation during the backward CS following the appetitive US. Both alternatives can be tested in studies that uncouple the presentations of USs and backward CSs by either presenting unpaired CSs during partially-reinforced acquisition, or by assessing startle modulation during an extinction phase that immediately follows acquisition.

Explicit valence ratings revealed the same pattern of backward CS contrast effects in the start-stop instructions group as previously reported in studies using similar instructions, while no such finding emerged in the observe instructions group

(Green et al., 2019; Moran & Bar-Anan, 2013; Moran et al., 2016). These findings again confirm that instructions emphasising CS agency are required for the emergence of backward CS contrast effects on explicit valence ratings (unless shock USs are employed as in Andreatta et al., 2013). The lack of backward *assimilative* conditioning in the observe instructions group was unexpected. The overlap between the US and the backward CS should have increased the size of the expected assimilation effect as simultaneous conditioning also occurs (see Green, Luck, & Lipp, 2020b; Mallan et al., 2008), though Green et al. (2020b) showed that presenting a concurrent forward CS (CS-US-CS) reduces the backward CS assimilation effect generally seen in a simple backward conditioning design (US-CS). However, this reduction was not so large that an assimilation effect was not observed. It is possible that US intensity or US modality may play a role here as Mallan et al. (2008) found backward CS assimilation effects with picture USs and Andreatta et al. (2013) found backward CS contrast effects with electro-tactile USs. The intensity of the sound USs employed here may fall between those of the picture and electro-tactile USs used by Mallan et al. (2008) and Andreatta et al. (2013), thus resulting in a null effect if evaluative conditioning shifts from assimilation to contrast with increasing US intensity. However, as mentioned previously, the effects of US intensity on backward conditioning require further examination.

It could be argued that the dissociation between startle magnitude and explicit valence ratings in the observe instruction group and the fact that explicit valence ratings are affected by instructions while startle blink magnitude is not, are due to startle being measured online during acquisition and explicit valence ratings being measured after completion of the acquisition phase. This hypothesis was tested by Luck and Lipp (2017) as a potential explanation for the dissociation between startle magnitude and explicit valence ratings observed by Andreatta et al. 2010, in a backward conditioning experiment with a 100ms gap between shock US offset and the onset of geometric shape CSs. Startle blink and explicit valence ratings were measured concurrently during acquisition using an online measure of explicit valence. In contrast to the hypothesis, a dissociation between startle magnitude and explicit valence ratings was observed, suggesting that time of measurement does not account for the dissociation between measures.

Dissociations between startle and explicit reports of stimulus valence have also been observed in comparisons between clinical groups. Hazlett et al. (2007), for

instance, found larger startle during unpleasant word stimuli in participants diagnosed with borderline personality disorder than in healthy controls, whereas controls rated the words as more unpleasant than those with borderline personality disorder. While startle and valence ratings dissociated in Hazlett et al. (2007), both measures revealed assimilation effects. The dissociation was driven by psychopathology leading to assimilation effects of different sizes, but not effects in opposite directions. This pattern of between-participant dissociation differs from the pattern of within-participant dissociation found in the current study and in previous backward conditioning research, as the within-participant dissociation shows an assimilation effect on one measure and a contrast effect on the other (Andreatta et al., 2010; Luck & Lipp, 2017). The fact that within-participant dissociations can occur resulting in opposing patterns of results (i.e. assimilation on one measure and contrast on the other) may suggest that startle and explicit valence ratings reflect different aspects of conditional responding.

According to Russell's (2003) framework, negative core affect can be experienced while still conceding that a pleasant stimulus is in fact pleasant, suggesting that emotional experience and the ability to label stimulus valence can be independent from each other. Support for this notion was shown by Weber, Shinkareva, Kim, Gao, and Wedell (2020), as explicit valence ratings of the CSs were influenced by the affective valence elicited by the USs (i.e. pleasant or unpleasant emotional state), independent of how much participants liked the USs (i.e. neutral, low, or high liking of the US – explicit stimulus valence). Applied to the current findings, this may suggest that startle modulation indicates the pure emotional response to the backward CS whereas explicit valence ratings reflect the entire learning episode integrating the emotional response with appraisals of the backward CS based on, for instance, its perceived agency due to the instructional manipulation or the information provided by specific procedural features (see difference between CS-US-CS and US-CS procedures; Andreatta et al., 2013; Green et al., 2020a). Support for the idea that startle and valence ratings may be measuring different things was shown by Patrick, Bradley and Lang (1993), as criminal psychopaths showed inhibited startle responding during unpleasant images suggesting pleasant emotion was experienced, yet the images were explicitly rated as aversive. While this dissociation appeared to be due to psychopathology, it showed that measures supposedly assessing the same thing can dissociate, suggesting that a

dissociation in participants who were not screened for psychopathology may result because different aspects of a conditional response are being measured. Additionally, Green et al. (2020) found backward CS contrast effects on startle and ratings in a CS-US-CS design and a dissociation between startle and ratings in a US-CS design. If startle and ratings are measuring the same aspect of a conditional response this dissociation should not have occurred. Future research should assess this explanation of the dissociation between startle modulation and explicit valence ratings observed here. This research could utilise conditioning procedures that involve qualifiers of the relationship between the CS and US as employed by Fiedler and Unkelbach (2011) or Förderer and Unkelbach (2012), or the provision of additional information about the CS as employed by Luck and Lipp (2018).

A limitation of the current study was that participants were not screened for psychopathology. As mentioned previously, differences between healthy controls and those with psychopathology can lead to differences in the size of assimilation effects. Fortunately, as we were interested in dissociations resulting in contrast effects, as opposed to differences in the extent of assimilation effects, the possibility that some participants displayed a degree of psychopathology cannot account for the outcomes of the current study. A further limitation was that 17 participants completed the experiment with potentially distracting acoustic noise coming from outside of the building due to landscaping works. Fortunately, it is unlikely that any systematic effects resulted as these participants were distributed equally across groups and the pattern of results did not change with and without these participants. While the analysis without these participants was likely underpowered, the chances that instructions had an effect that could not be observed are low based on the small effect size ($\eta p^2 = .001$). Moreover, the observation that contrary to explicit ratings the pattern of startle modulation shown here is not affected by differences in instructions is consistent with findings that the type of conditioning design used (concurrent forward and backward conditioning: CS-US-CS vs. backward conditioning only : US-CS) affects explicit valence ratings, but not startle modulation (Green et al., 2020a).

In conclusion, the present study found that startle modulation was unaffected by instructional manipulations as contrast effects were observed during backward CSs regardless of the instruction used. Moreover, this study confirmed that backward CS contrast effects emerged on explicit valence ratings only when instructions

highlighting the roles of the CSs were presented. These findings indicate that instructional manipulations do not affect startle modulation in the same way as explicit valence ratings in a within-participant forward and backward conditioning procedure (CS-US-CS). Taken together with past findings from research on backward conditioning, the current results raise interesting questions about the complexities of emotional processes and how they manifest – even in very simple emotion induction paradigms.

4.5 References

Andreatta, M., Mühlberger, A., Glotzbach-Schoon, E., & Pauli, P. (2013). Pain predictability reverses valence ratings of a relief-associated stimulus. *Frontiers in Systems Neuroscience, 7*(53), 1-12, doi:10.3389/fnsys.2013.00053

Andreatta, M., Mühlberger, A., Yarali, A., Gerber, B., & Pauli, P. (2010). A rift between implicit and explicit conditioned valence in human pain relief learning. *Proceedings of the Royal Society of London B: Biological Sciences, 277*, 2411-2416. doi: 10.1098/rspb.2010.0103

Blumenthal, T. D., Cuthbert, B. N., Filion, D. L., Hackley, S., Lipp, O. V., & Van Boxtel, A. (2005). Committee report: Guidelines for human startle eyeblink electromyographic studies. *Psychophysiology, 42*, 1-14. doi: 10.1111/j.1469-8986.2005.00271.x

Bradley, M. M., Cuthbert, B. N., & Lang, P. J. (1990). Startle reflex modification: Emotion of attention? *Psychophysiology, 27*, 513-522. doi:10.1111/j.1469-8986.1990.tb01966.x

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*, 161-189. doi: 10.1177/1088868318763261

Cuthbert, B. N., Bradley, M. M., & Lang, P. J. (1990). Probing picture perception: Activation and emotion. *Psychophysiology, 33*, 103-111. doi:10.1111/j.1469-8986.1996.tb02114.x

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*, 853- 869. doi:10.1037//0033-2909.127.6.853

Dirikx, T., Hermans, D., Vansteenwegen, D., Baeyens, F., & Eelen, P. (2004). Reinstatement of extinguished conditioned responses and negative stimulus valence as a pathway to return of fear in humans. *Learning & Memory, 11*(5), 549-554. doi: 10.1101/lm.78004

Fiedler, K., & Unkelbach, C. (2011). Evaluative conditioning depends on higher order encoding processes. *Cognition and Emotion, 25,* 639-656.

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

Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers, 35,* 116-124. doi:10.3758/BF03195503

Galdi, S., Arcuri, L., & Gawronski, B. (2008). Automatic mental associations predict future choices of undecided decision makers. *Science, 321,* 1100-1102. doi: 10.1126/science.1160769

Gibson, B. (2008). Can evaluative conditioning change attitudes toward mature brands? New evidence from the Implicit Association Test. *Journal of Consumer Research, 35,* 178-188. doi:10.1086/527341

Green, L. J. S., Luck, C., Gawronski, B., & Lipp, O. V. (2019). Contrast effects in backward evaluative conditioning: Exploring effects of affective relief/disappointment versus instructional information. *Emotion*. Advance online publication. doi:10.1037/emo0000701

Green, L. J. S., Luck, C., & Lipp, O. V. (2020a). How disappointing: Startle modulation reveals conditional stimuli presented after pleasant unconditional stimuli acquire negative valence. *Psychophysiology, 57*(8), 1-16. doi:10.1111/psyp.13563

Green, L. J. S., Luck, C., & Lipp, O. V. (2020b). Assimilation or contrast effects in backward evaluative conditioning: The role of US offset predictability. Submitted.

Hazlett, E. A., Speiser, L. J., Goodman, M., Roy, M., Carrizal, M., Wynn, J. K., . . . New, A. S. (2007). Exaggerated Affect-Modulated Startle During Unpleasant Stimuli in Borderline Personality Disorder. *Biological Psychiatry, 62*(3), 250-255. doi:10.1016/j.biopsych.2006.10.028

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. *Psychological Bulletin, 136,* 390-421. doi:10.1037/a0018916

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6,* 65-70.

Hu, X., Gawronski, B., & Balas, R. (2017a). 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,* 17-32.

Hu, X., Gawronski, B., & Balas, R. (2017b). Propositional versus dual-process accounts of evaluative conditioning: II. The effectiveness of counter-conditioning and counter-instructions in changing implicit and explicit evaluations. *Social Psychological and Personality Science, 8,* 858-866.

LeBel, E. P., & Campbell, L. (2009). Implicit partner affect, relationship satisfaction, and the prediction of romantic breakup. *Journal of Experimental Social Psychology, 45,* 1291-1294. doi: 10.1016/j.jesp.2009.07.003

Luck, C. C., & Lipp, O. V. (2017). Startle modulation and explicit valence evaluations dissociate during backward fear conditioning. *Psychophysiology, 54,* 673-683.

Luck, C. C., & Lipp, O. V. (2018). Verbal instructions targeting valence alter negative conditional stimulus evaluations (but do not affect reinstatement rates). *Cognition and Emotion,* 1-20. doi:10.1080/02699931.2017.1280449

Lipp, O. V. (2006). Human fear learning: Contemporary procedures and measurement. In M. G. Craske, D. Hermans & D. Vansteenwegen (Eds.), (2006). Fear and learning: From basic processes to clinical implications (pp. 37-52). Washington: APA Books.

Lipp, O. V., Siddle, D. A. T., & Dall, P. J. (2003). The effects of unconditional stimulus valence and conditioning paradigm on verbal, skeletal, and

autonomic indices of Pavlovian conditioning. *Learning and Motivation, 34,* 32-51. doi:10.1016/S0023-9690(02)00507-6

Mallan, K. M., Lipp, O. V., & Libera, M. (2008). Affect, attention, or anticipatory arousal? Human blink startle modulation in forward and backward affective conditioning. *International Journal of Psychophysiology, 69.* 9-17. doi:10.1016/j.ijpsycho.2008.02.005

Moran, T., and Bar-Anan, Y. (2013). The effect of object-valence relations on automatic evaluation. *Cognition and Emotion, 27.* 743-752. doi:10.1080/02699931.2012.732040

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2016). The assimilative effect of co-occurrence on evaluation above and beyond the effect of relational qualifiers. *Social Cognition, 34.* 435-461. doi: 101521soco2016345435

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

Patrick, C. J., Bradley, M. M., & Lang, P. J. (1993). Emotion in the criminal psychopath: Startle reflex modulation. *Journal of Abnormal Psychology, 102,* 82-92. doi:10.1037/0021-843X.102.1.82

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review, 110.* 145-172. doi:10.1037/0033-295X.110.1.145

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

Vrana, S. R., Spence, E. L., & Lang, P. J. (1988). The startle probe response: A new measure of emotion? *Journal of Abnormal Psychology, 97,* 487-491. doi:10.1037/0021-843X.97.4.487

Zbozinek, T. D., Hermans, D., Prenoveau, J. M., Liao, B., & Craske, M. G. (2015). Post-extinction conditional stimulus valence predicts reinstatement fear:

Relevance for long-term outcomes of exposure therapy. *Cognition and Emotion, 29*(4), 654-667. doi: 10.1080/02699931.2014.930421

4.6 Figures

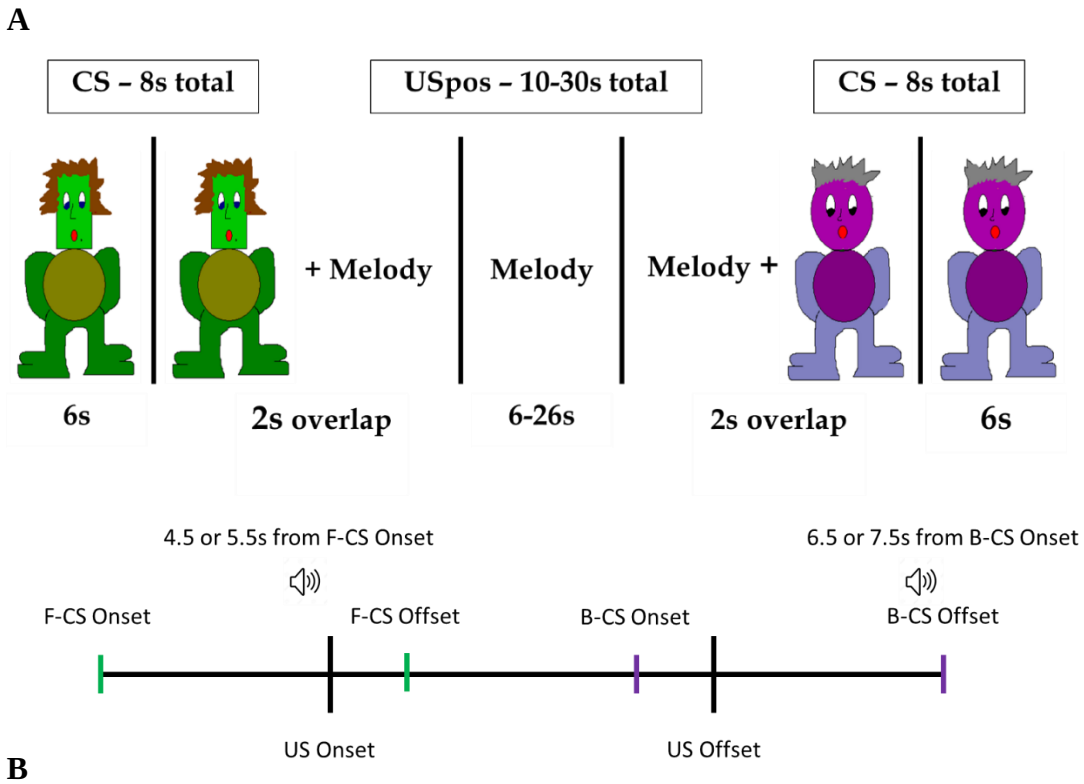

Figure 1. A) Example of a positive CS-US-CS trial. Forward and backward CSs were presented alone for 6 seconds and overlapping with the US for 2 seconds (8 seconds of total CS presentation). USs varied in duration for 10, 15, 20, 25, or 30 seconds. CS = Conditional stimulus, USpos = Positive unconditional stimulus. B) Example of startle probe timing relative to F-CS, US, and B-CS onset and offset. Startle probes were presented at 4.5 or 5.5 seconds after F-CS onset and 6.5 or 7.5 seconds after B-CS onset. F-CS = Forward conditional stimulus, US = Unconditional stimulus, B-CS = Backward conditional stimulus, speaker picture represents startle probe.

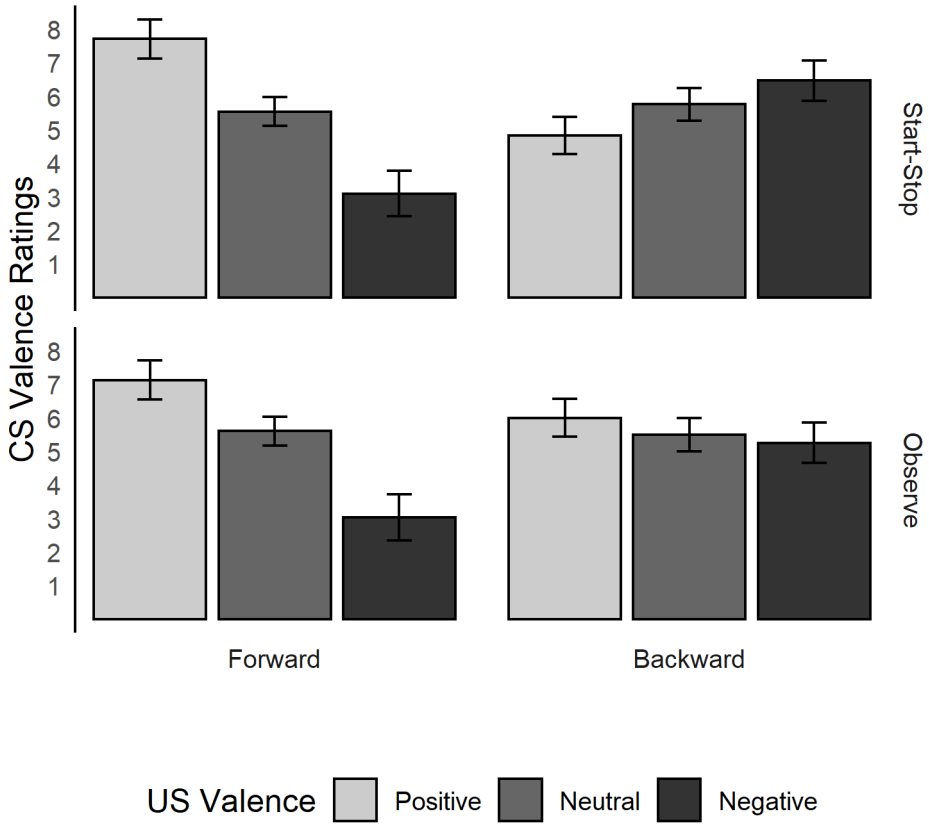

Figure 2. Mean explicit valence ratings for forward and backward CSs as a function of US valence (positive vs. negative) and instructions (start-stop vs. observe). Error bars represent 95% confidence intervals of the mean.

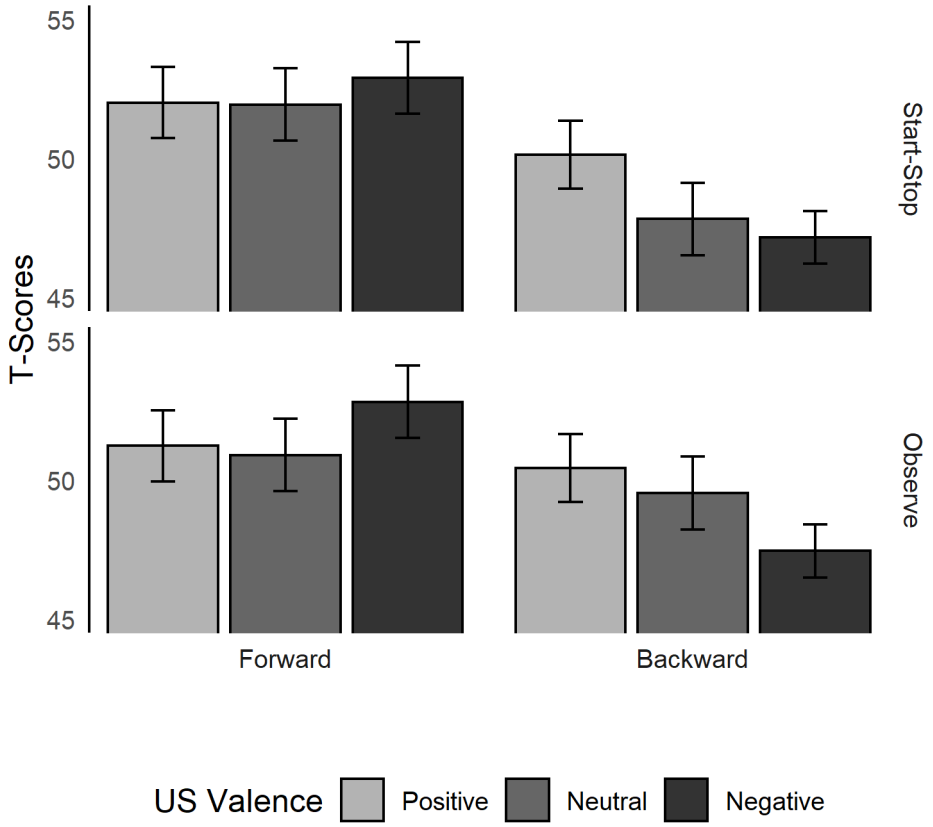

Figure 3. Startle blink magnitude (*T*-scores) for forward and backward CSs as a function of US valence (positive vs. neutral vs. negative) and instructions (start-stop vs. observe). Error bars represent 95% confidence intervals of the mean.

4.7 Supplementary Materials

4.7.1 Affective priming

Participants categorised positive or negative target words following the presentation of all CSs. Incorrect categorisation of the target word was scored as an error. Responses deemed to be outside the window of a response suggestive of task adherence, such as responses faster than 300ms and slower than 1000ms, were also scored as errors (Koppehele-Gossel, Hoffmann, Banse, & Gawronski, 2020). Participants were removed from the analyses if they made more than 25% errors, resulting in the removal of one participant from the observe instructions group (final $n = 81$, observe instructions group $n = 40$). Mean reaction times and percentage of errors were calculated for each CS family and target word pair. These means were then used to calculate difference scores for reaction times within each CS prime (CS prime/positive target word – CS prime/negative target word), such that assimilation effects are represented by negative scores for positive CS primes and by positive scores for negative CS primes. These scores were subjected to a 2 (Instructions: start-stop vs. observe; between-participants) × 2 (Conditioning Type: forward vs. backward; within-participant) × 3 (US Valence: positive vs. neutral vs. negative) mixed model ANOVA.

Reaction Times. Figure S1 suggests a linear trend for US valence for forward conditioning and a quadratic trend for backward conditioning regardless of group. A main effect of US valence, F (2, 78) = 3.73, $p = .028$, $\eta p^2 = .09$, was qualified by a Conditioning Type × US valence interaction, F (2, 78) = 4.25, $p = .018$, $\eta p^2 = .10$. Follow-up analyses showed that difference scores for forward CSs paired with positive USs were significantly smaller than difference scores for forward CSs paired with neutral USs, t (79) = 3.31, $p = .001$, $d = 0.37$, and negative USs, t (79) = 2.96, $p = .004$, $d = 0.33$. No other comparisons for forward or backward conditioning were significant, ts < 1.70, ps > .094. One-sample t-tests showed that assimilation effects occurred for CSs forwardly paired with positive USs, t (80) = 3.49, $p = .001$, $d = 0.39$, and for CSs backwardly paired with positive USs, t (80) = 2.29, $p = .024$, $d = 0.25$, as the difference scores were significantly smaller than zero. CSs backwardly paired with negative USs showed a contrast effect, as the difference score was also significantly smaller than zero, t (80) = 3.00, $p = .004$, $d = 0.33$.

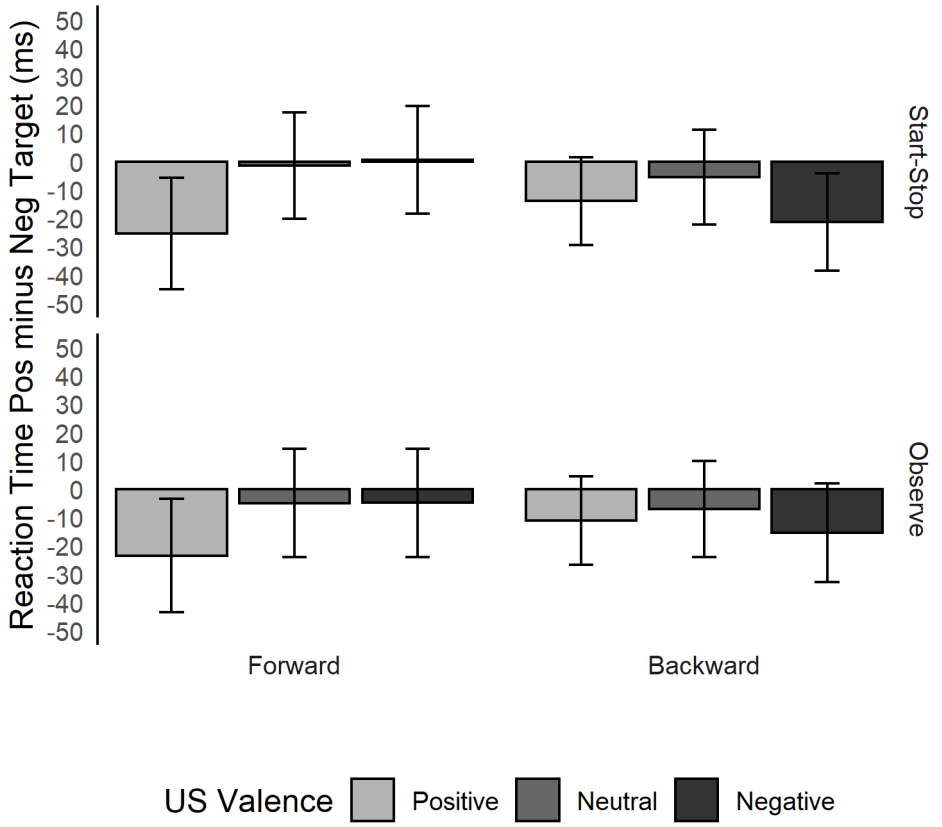

Figure S1. Difference scores (positive target words – negative target words) for reaction times from the affective priming task for forward and backward CSs as a function of US valence (positive vs. neutral vs. negative) and instructions (start-stop vs. observe). Assimilation effects are represented by negative scores for positive CS primes and by positive scores for negative CS primes. Error bars show 95% confidence intervals of the mean.

4.7.2 Startle blink latency

Figure S2 below suggests faster blink onset following probes presented during CSs paired with negative USs than CSs paired with positive and neutral USs for forward conditioning and faster blinks following probes presented during CSs paired with positive USs than CSs paired with negative and neutral USs for backward conditioning, in both groups. The final sample subjected to the ANOVA comprised 39 in the start-stop group and 40 in the observe group due to missing responses causing the ANOVA to remove these participants. A main effect of conditioning type, F (1, 77) = 71.59, $p < .001$, $\eta p^2 = .482$, was qualified by a Group × Conditioning Type interaction, F (1, 77) = 4.42, $p = .039$, $\eta p^2 = .054$, and a Conditioning Type x US Valence interaction, F (2, 76) = 7.74, $p = .001$, $\eta p^2 = .169$.

The interaction between group and conditioning type was driven by faster blink onset in the observe instructions group than the start-stop instructions group for backward conditioning, F (1, 77) = 5.56, p = .021, ηp^2 = .067, while groups did not differ for forward conditioning, F (1, 77) = 2.15, p = .146, ηp^2 = .027. Follow-up analyses for the interaction between conditioning type and US valence revealed that blink onset was faster during CSs presented after positive USs than CSs presented after negative USs, t (78) = 2.24, p = .003, d = 0.25. Blink onset was also faster during CSs presented after positive USs than CSs presented after neutral USs, t (78) = 2.40, p = .033, d = 0.27, however, when applying the Bonferroni-Holm correction this comparison was not significant. No other pairwise comparisons reached significance, ts < 1.67, ps > .099.

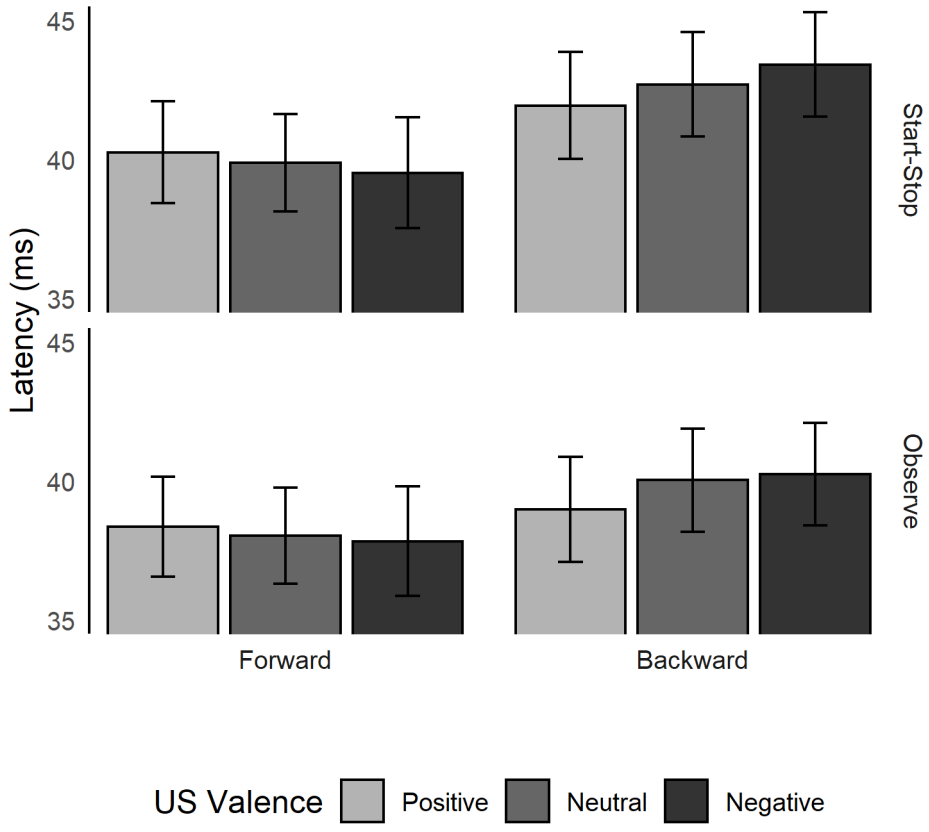

Figure S2. Startle blink latency (ms) for forward and backward CSs as a function of US valence (positive vs. neutral vs. negative) and instructions (start-stop vs. observe). Error bars represent 95% confidence intervals of the mean.

Chapter 5: Assimilation or contrast effects in backward evaluative conditioning: The role of US offset predictability

Luke J. S. Green

Curtin University

Camilla C. Luck

Curtin University

Ottmar V. Lipp

Curtin University

Author Notes

This work was supported by an Australian Government Research Training Program Scholarship to Luke Green and grant DP180111869 from the Australian Research Council to Ottmar Lipp.

Correspondence concerning this article should be sent to: Luke J S Green, School of Psychology, Curtin University, GPO Box U1987 Perth WA 6845, Australia. Email: luke.green2@postgrad.curtin.edu.au.

Data and materials are available at https://osf.io/4mtph/.

5.1 Abstract

Backward evaluative conditioning has been shown to result in assimilative effects where backward conditional stimuli (CS) acquire the valence of the unconditional stimulus (US) or in contrast effects where backward CSs acquire valence opposite to the US. The current experiments were designed to assess whether US offset predictability, manipulated by varying US duration (fixed vs. variable) and US-CS overlap (overlap vs. no overlap) determines the nature of backward evaluative conditioning. Experiment 1 employed backward conditioning only ($US_{Pleasant}$-CS_1; $US_{Unpleasant}$-CS_2) whereas Experiment 2 employed concurrent forward and backward conditioning (CS_3-$US_{Pleasant}$-CS_1; CS_4-$US_{Unpleasant}$-CS_2). Backward CS assimilation effects emerged in Experiment 1 and did not differ as a function of US duration or US-CS overlap. In Experiment 2, US-CS overlap led to stronger backward assimilation effects than when CSs and USs did not overlap. These findings provide no support for the notion that the nature of backward CS evaluations varies as a function of US offset predictability and suggest that past findings of backward CS contrast effects were driven by the instructional manipulations used in those studies.

Keywords: Associative learning, backward conditioning, evaluative conditioning, evaluative learning, forward conditioning

How likes and dislikes are acquired and changed is an important line of enquiry in psychology because they influence decision making in all facets of life – from choosing romantic partners, to deciding which career path to follow or which political party to support (e.g., Galdi, Arcuri, & Gawronski, 2008; Gibson, 2008; LeBel & Campbell, 2009). Moreover, our likes and dislikes have been implicated as an important element in the formation of implicit and explicit attitudes and prejudice (Corneille & Stahl, 2019; Olson & Fazio, 2006). Further understanding the basic processes underlying the acquisition of likes and dislikes may assist in developing more successful methods for changing attitudes and combating prejudice, as well as assisting in understanding decision making and consumer behaviour.

Evaluative conditioning (EC) is one of the basic processes through which likes and dislikes can be acquired and changed (De Houwer, 2007). EC occurs when a neutral stimulus (CS) is paired with a valenced unconditional stimulus (US) and the valence of the CS changes (De Houwer, Thomas, & Baeyens, 2001). Examples of EC are especially prominent in advertising, such as when a popular celebrity endorses a new product resulting in favourable evaluations and the potential purchase of this once neutral product. A common method for investigating EC in the laboratory is the picture-picture paradigm. This paradigm involves presenting pictures of neutral (CSs) and valenced stimuli (USs) in temporal and/or spatial proximity and measuring changes in stimulus valence (Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010; Levey & Martin, 1975). Changes in stimulus valence as a result of pairing these pictures can be indexed through both explicit and implicit measures. Self-report explicit measures such as valence ratings involve asking participants how pleasant or unpleasant they find a certain stimulus. Implicit measures, such as the implicit association test (IAT; Greenwald, McGhee, & Schwartz, 1998) or affective priming (Fazio, Jackson, Dunton, & Williams, 1995), are reaction time based measures that require participants to categorise stimuli by pressing keys on a keyboard that correspond to specific categories. These measures exploit the fact that we are faster to evaluate stimuli that are congruent in valence than stimuli that are incongruent in valence (De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009; Gawronski & De Houwer, 2014).

Valence change resulting from EC can manifest in two different formats, assimilation or contrast. Assimilation effects occur when the CS acquires the valence of the US (Mallan, Lipp, & Libera, 2008; Staats & Staats, 1957). Contrast effects

occur when the CS acquires valence that is opposite to that of the US and seem to emerge in the picture-picture paradigm only when additional information is provided about the relationship between the CS and US (Fiedler & Unkelbach, 2011; Förderer & Unkelbach, 2012; Moran & Bar-Anan, 2013; Moran, Bar-Anan, & Nosek, 2016). For example, Moran et al. (2016) demonstrated contrast effects in a within-participant forward and backward conditioning experiment where they presented pleasant and unpleasant picture USs that appeared to be controlled by four different families of alien creatures (CSs). Participants were told that each set of CSs had a different role to play; that one would start the unpleasant stimulus, one would stop the unpleasant stimulus, one would start the pleasant stimulus, and one would stop the pleasant stimulus. Moran et al. (2016) reported assimilation effects for forward CSs, i.e. CSs that started pleasant USs became pleasant and CSs that started unpleasant USs became unpleasant. However, for backward CSs, contrast effects emerged, i.e. the CSs that stopped the unpleasant US became pleasant and CSs that stopped the pleasant US became unpleasant.

The backward conditioning results reported by Moran et al. (2016) are inconsistent with those from studies that did not incorporate instructional manipulations and assessed forward and backward conditioning between-participants (CS-US vs. US-CS; see Kim, Sweldens, & Hutter 2016; Mallan et al., 2008) instead of within-participants (CS-US-CS). For example, Mallan et al. (2008) demonstrated assimilation effects in a between-participants picture-picture paradigm presenting geometric shapes (CSs) either before (forward conditioning), during (simultaneous conditioning), or after (backward conditioning) pleasant and unpleasant pictures (USs). CSs paired with pleasant USs were evaluated as more pleasant than CSs paired with unpleasant USs on explicit and implicit measures with no difference across the conditioning groups (see also Hofmann et al. 2010 for a meta-analysis).

Green, Luck, Gawronski, and Lipp (2019) assessed the impact of instructions on backward evaluative conditioning comparing those employed by Moran et al. (2016) and those used by Mallan et al. (2008) in a within-participant CS-US-CS picture-picture paradigm. The Moran et al. (2016) instructions informed participants about the role of each CS in starting or stopping the US and had yielded backward CS contrast effects. The Mallan et al. (2008) instructions asked participants to pay attention to the stimuli as they will be asked questions about them later and had yielded backward CS assimilation effects. Green et al. (2019) confirmed that the

instructions determined the nature of backward evaluative conditioning as contrast effects were found using the Moran et al. (2016) instructions and assimilation effects using the Mallan et al. (2008) instructions. While the results of Green et al. (2019) suggest that instructions are crucial in determining the nature of backward evaluative conditioning in a within-participant CS-US-CS paradigm, below we will discuss two alternative accounts that may explain backward CS contrast effects without reliance on instructional manipulations.

Moran et al. (2016) presented picture USs for durations of 3, 4, and 5 seconds. It is possible that presenting USs that vary in duration renders the offset of the US unpredictable, thus increasing its salience. Increasing US salience may then lead to an opponent process occurring whereby relief or disappointment are triggered at the offset of unpleasant or pleasant USs, respectively (Andreatta, Mühlberger, Yarali, Gerber, & Pauli, 2010; Green, Luck, & Lipp, 2020; Luck & Lipp, 2017). These opponent emotional reactions may then be associated with the backward CSs resulting in backward CS contrast effects. Evidence for such opponent emotional processes was provided by Andreatta, Mühlberger, Glotzbach-Schoon, and Pauli (2013) in a within-participant forward and backward conditioning experiment (CS-US-CS) employing geometric shapes as CSs and an aversive electro-tactile stimulus as the US. After conditioning, participants rated the backward CS, which was presented 6s after the US offset, as more pleasant than the forward CS and a neutral CS that was presented alone. This suggests that relief experienced after the offset of the aversive US had been conditioned to the backward CS. It should be noted that the interval between US offset and backward CS onset does not seem to affect relief learning. Luck and Lipp (2017) showed the same relief learning effect with a 100ms gap between US offset and backward CS onset as did Andreatta et al. (2010) with an 6s gap between US offset and backward CS onset. If US duration variability increases US salience sufficiently such that US offset elicits an opponent emotional reaction, then this may support the backward CS contrast effects observed by Moran et al. (2016) in absence of an additional instructional manipulation.

In addition to varying the duration of the USs, overlapping the US and CS may influence backward CS learning by rendering the backward CS a powerful signal of US offset. The backward CS that signals the offset of an aversive stimulus may lead to that CS acquiring positive properties as it signals the end of an aversive event. The same would occur for the backward CS that signals the offset of an

appetitive stimulus with that backward CS acquiring negative properties. Moran and Bar-Anan (2013) overlapped visual CSs and sound USs that lasted 10, 15, 20, 25, or 30 seconds in a CS-US-CS paradigm similar that used by Moran et al. (2016). The end of the forward CS overlapped with the start of the US for 2000ms, and the end of the US overlapped with the start of the backward CS for 2000ms. If overlapping CSs and USs does render the backward CS a predictor of US offset, then the backward CS contrast effects observed by Moran and Bar-Anan (2013) may be due to the scheduling of CS and US even in absence of an explicit instructional manipulation that highlights this. It should be noted, however, that overlapping CSs and USs may also result in simultaneous conditioning which has been shown to result in assimilation effects (Mallan et al., 2008). Thus, determining the potential effects of US/CS overlap on backward evaluative conditioning seems required.

We have detailed two separate and plausible accounts for the backward CS contrast effects as observed by Moran et al. (2016) and Moran and Bar-Anan (2013). Moreover, based on an underpowered pilot study conducted in our lab (see https://osf.io/4mtph/) we have preliminary evidence suggesting that varying US duration and US/CS overlap may interact to yield backward CS contrast effects. To confirm this observation in an adequately powered study, the aim of Experiment 1 was to determine whether US offset predictability, manipulated by varying US duration and overlapping US-CS presentations such that the backward CSs signal the offset of the US, would be sufficient to result in backward CS contrast effects. If this is the case, backward CS contrast effects like those shown by Moran et al. (2016) and Moran and Bar-Anan (2013) should be observed in absence of an explicit instructional manipulation highlighting the role of the CSs.

5.2 Experiment 1

A 2 × 2 factorial design was employed to assess the effects of US duration variability (fixed vs. variable) and US-CS overlap (overlap vs. no overlap) in a simple backward conditioning procedure. Two identical training phases were performed successively and explicit valence ratings were measured before and after each phase for a total of three measurements. We used two post-acquisition rating measurements to confirm that evaluative conditioning effects increase with more trials. Affective priming was measured before the first and after the second training

phase for a total of two measurements as presenting 80 CSs alone could be considered an extinction phase. Based on our pilot work, we predicted that in an adequately powered study backward CS assimilation effects on explicit valence ratings would occur for the fixed duration-no overlap, fixed duration-overlap, and variable duration-no overlap groups, while backward CS contrast effects were predicted in the variable duration-overlap group. We expected the predicted effects to be larger after the second training phase (post-training 2) than the first training phase (post-training 1). As implicit measures have only revealed backward CS contrast effects when instructional manipulations highlighting the relation between CSs and USs are presented on every trial, we predicted assimilation effects on affective priming for all groups (Hu, Gawronski, & Balas, 2017a; see also Bading, Stahl, & Rothermund, 2020).

5.2.1 Method

Participants. Following ethical approval from the Curtin University Human Research Ethics Board, two hundred and sixty participants were recruited through Amazon Mechanical Turk using TurkPrime (M-Turk; Litman, Robinson, & Abberbock, 2017; mean age = 36.03, *SD* = 11.02). An *a*-priori power calculation in G*Power to detect a medium sized interaction (η_p^2 = .07) at .80 power and α = .05 based on an ANOVA *F* test suggested 188 participants (within-between interaction; 4 groups [Fixed-no overlap; Fixed-overlap; Variable-overlap; Variable-no overlap]; 3 measurement points [Pre-training; Post-training 1; Post-training 2]; Effect size specification: as in Cohen [1988]; nonsphericity correction: 1. Note: US duration and US/CS overlap have been collapsed into one factor for the power analysis as G*Power only allows for two factors to be entered). US valence was not included in this power analysis as we planned to create difference scores based on this factor and use them as the dependent variable in our analyses (CS paired with positive US-CS paired with negative US). In order to ensure adequate power to enable follow-up *t*-tests if our highest order interaction was significant we also performed an a-priori power calculation in G*Power to detect a medium sized effect (*d* = 0.5) at .80 power for a two-tailed independent *t*-test. The power analysis suggested 64 participants for each of four groups. The fixed-no overlap group consisted of 67 participants; the fixed-overlap group included 65 participants; the variable-no overlap group consisted of 63 participants; the variable-overlap group comprised 65 participants. Groups did not differ on gender, $\chi^2(3) = 0.78$, $p = .855$, or ethnicity, $\chi^2(15) = 11.45$, $p = .720$.

Age was submitted to a 2 (US Duration: fixed vs. variable; between-participants) × 2 (US-CS overlap: no overlap vs. overlap; between-participants) between-groups ANOVA and an interaction was found, $F(1, 252) = 5.06$, $p = .025$, $\eta_p^2 = .020$, $BF_{incl} = 1.44$. Pairwise comparisons showed that the mean age was higher in the fixed duration-no overlap group ($M = 39.06$, $SD = 13.01$) compared to the variable duration-no overlap group ($M = 34.18$, $SD = 8.89$), $t(254) = 2.53$, $p = .012$, $d = 0.32$, $BF_{10} = 0.45$, and the fixed duration-overlap group ($M = 34.73$, $SD = 9.94$), $t(254) = 2.25$, $p = .025$, $d = 0.28$, $BF_{10} = 0.57$.

Apparatus/stimuli. Four cartoon aliens created by Moran and Bar-Anan (2013) were used as CSs (materials from Moran and Bar-Anan, 2013, available at https://osf.io/cqsnj/). Each alien differed in colour and head shape. Four positive pictures depicting pleasant sceneries and baby animals and four negative pictures depicting aggressive humans and animals from the International Affective Picture System (IAPS; Lang, Bradley, & Cuthbert, 1999) were used as USs (1050, 1300, 1440, 1710, 5833, 6313, 6560, and 8190). Inquisit 4 Web by Millisecond Software ™ (2016) was used to run the experiment and to record responses in all tasks.

Explicit valence ratings. Participants were shown two of the CSs individually and asked to "Please rate on a scale of 1 to 9, where 1 = unpleasant and 9 = pleasant". CSs were selected randomly across participants.

Affective priming task. Each CS was presented twice with 10 positive target words and 10 negative target words for a total of 80 trials. Trials were randomised across participants. Following a 500ms fixation cross, the CS prime was presented for 200ms followed by the target word until the participant provided their response. Participants were instructed to press the *I* key if the target word was positive and the *E* key if the target word was negative. A red cross appeared on the screen if an error was made. Target words were taken from Hu et al. (2017a; Hu, Gawronski, & Balas, 2017b). The positive words were *pleasant, good, outstanding, beautiful, magnificent, marvellous, excellent, appealing, delightful,* and *nice.* The negative words were *unpleasant, bad, horrible, miserable, hideous, dreadful, painful, repulsive, awful,* and *ugly.* Twenty practice trials were presented before the first affective priming task only (pre-training) using the same stimuli as in the main task.

Recollective Memory Test. The current study also included measures of recollective memory for exploratory purposes. Details of the test and the results are available in the supplementary materials at https://osf.io/4mtph/.

Demographics questionnaire. Participants provided their age, gender, and ethnicity. They were also asked to provide information about the environment in which they completed the task and if they had any comments.

Procedure. Participants selected the HIT (human intelligence task) on M-Turk and read the description of the study. Participants were presented with an information sheet outlining the experiment, informed that they could withdraw at any time by pressing 'ctrl + q', and prompted to press 'continue' if they consented to participate. If participants pressed 'continue', informed consent was implied. The first explicit valence ratings and affective priming task were presented followed by the first training phase. The first training phase comprised 12 positive and 12 negative trials randomly presented with inter-trial intervals of 4, 5, and 6 seconds. Each trial consisted of a pleasant or unpleasant US, followed by a backward CS (the CS following the pleasant US is labelled CSp and the CS following the unpleasant US is labelled CSu). This US-CS paradigm was adapted from Green, Luck, and Lipp (2019; Experiment 2) which was originally based on the CS-US-CS paradigm from Moran and Bar-Anan (2013). We used one CS from each of the two alien families used in those studies and four positive and four negative pictures as USs. CSs were fully counter-balanced across participants resulting in 12 training sequences. USs were flashed for 1s followed by a blank screen for 200ms resulting in 5 x 1s flashes for a total of 5s presentation time in the fixed groups. USs were presented for either 3, 5, or 7s of total presentation time in the variable groups. In the overlap groups the 1st second of CS presentation coincided with the last second of US presentation resulting in an overlapping image of the CS and US as in Pleyers, Corneille, Luminet, and Yzerbyt (2007). CSs were presented for 5s in total during each trial in all groups (see Figure 1 for depiction of trial sequences).

Participants received the following instructions:

In this task you will be presented with a series of pictures. Please pay attention to which pictures follow each other as you will be tested on this at the end of the experiment.

After the first training phase, the second round of explicit valence ratings was presented. Participants were then instructed that the experiment would continue as before and shown the same instructions as before the first training phase. The second training phase was identical to the first with only trial order differing due to randomisation. After the second training phase, participants performed a third round

of explicit valence ratings and the second affective priming task. This was followed by the memory test and demographics questionnaire. Participants then received a completion code to receive their payment and were thanked for their participation. The experiment took approximately 25 minutes on average to complete, and participants were compensated US-$4.50.

Statistical analyses. IBM SPSS Statistics 25 was used for frequentist data analysis. We also report Bayesian analyses to supplement the frequentist analyses. JASP 0.10.0.0 was used for Bayesian analyses of the full models and the Bayes Factor package in R was used to conduct follow-up *t*-tests. EC scores for explicit valence ratings were calculated by subtracting valence ratings of CSs paired with unpleasant USs from those of CSs paired with pleasant USs (CSp – CSu) separately for pre-training, post-training 1, and post-training 2. Assimilation effects were reflected by positive EC scores and contrast effects by negative EC scores. In the affective priming task, trials were considered as error trials when the target word was incorrectly categorised or when reaction times were shorter than 300ms or longer than 1000ms, as these were deemed to be outside the window of a valid response (see Koppehele-Gossel, Hoffmann, Banse, & Gawronski, 2020). Participants were removed from the analyses if the percentage of error trials was larger than 25% (n = 83). The high number of exclusions suggests that some participants may not have been native English speakers or were providing dummy responses instead of completing the task properly. Of the remaining 177 participants, 8.78% of trials were excluded as errors. Priming scores were calculated as the difference in response times between incongruent and congruent trials: ([CSs paired with pleasant USs/unpleasant target words + CSs paired with unpleasant USs/pleasant target words] – [CSs paired with pleasant USs/pleasant target words + CSs paired with unpleasant USs/unpleasant target words]). Priming scores were calculated separately for pre-training and post-training. Positive priming scores suggest an assimilation effect, while negative scores suggest a contrast effect. EC scores were subjected to both Frequentist and Bayesian 2 (US Duration: fixed vs. variable; between-participants) × 2 (US-CS Overlap: no overlap vs. overlap; between-participants) × 3 (Time: pre-training vs post-training 1 vs post-training 2; within-participant) mixed ANOVAs, and priming scores were subjected to both Frequentist and Bayesian 2 (US Duration: fixed vs. variable; between-participants) × 2 (US/CS Overlap: no overlap vs. overlap; between-participants) × 2 (Time: pre-training vs post-training;

within-participant) mixed ANOVAs. Significant interactions were followed-up with pairwise comparisons and one sample *t*-tests where appropriate. Pillai's trace values of the multivariate solution are reported for main effects and interactions (α criteria = .05). BF_{10} values from the Bayesian model comparison are reported for main effects and follow-up *t*-tests, and $BF_{inclusion}$ (BF_{incl}) values from the effects analysis (across matched models) are reported for interactions using default priors (*t*-tests: Cauchy prior scale: 0.707; ANOVAs: r scale fixed effects = 0.5, r scale random effects = 1, r scale covariates = 0.354, auto sample). The $BF_{inclusion}$ (across matched models) compares models containing the effect of interest with equivalent models stripped of that effect. This leaves a model containing only the interaction effect without lower order effects contributing to the model (known as the Baws approach; see Mathôt, 2017, for a discussion).

5.2.2 Results

Explicit valence ratings. Mean EC scores depicted in Figure 2 suggest assimilation effects at post-training 1 and 2 in all groups. The ANOVA revealed a significant main effect of time, $F(2, 249) = 144.29$, $p < .001$, $\eta_p^2 = .537$, $BF_{10} = 3.71 \times 10^{56}$. The three-way interaction between US variability, US/CS overlap, and time was not significant, $F(2, 249) = 1.42$, $p = .244$, $\eta_p^2 = .011$, $BF_{incl} = 0.31$. Follow-up analyses revealed a significantly larger assimilation effect at post-training 2 in comparison to post-training 1, $t(253) = 6.81$, $p < .001$, $d = 0.43$, $BF_{10} = 2.59 \times 10^{8}$, and pre-training, $t(253) = 16.96$, $p < .001$, $d = 1.06$, $BF_{10} = 7.60 \times 10^{7}$, and a significantly larger assimilation effect at post-training 1 compared to pre-training, $t(253) = 11.92$, $p < .001$, $d = 0.75$, $BF_{10} = 9.73 \times 10^{4}$. One sample *t*-tests indicate that EC scores were significantly larger than 0 at post-training 1, $t(259) = 12.65$, $p < .001$, $d = 0.78$, $BF_{10} = 4.70 \times 10^{25}$, and post-training 2, $t(259) = 18.24$, $p < .001$, $d = 1.13$, $BF_{10} = 8.52 \times 10^{44}$, but not pre-training, $t(253) = 0.20$, $p < .001$, $d = 0.01$, $BF_{10} = 1.40$.

Affective priming. Mean priming scores from the two affective priming tasks completed at pre-training and post-training 2, depicted in Figure 3, suggest an assimilation effect at post-training 2 in the overlap groups only. However, the ANOVA revealed only an assimilation effect at post-training 2 demonstrated by a significant main effect of time, $F(1, 173) = 10.89$, $p < .001$, $\eta_p^2 = .059$, $BF_{10} = 25.88$. One sample *t*-tests indicate that priming scores were significantly larger than 0 at post-training 2, $t(176) = 3.28$, $p = .001$, $d = 0.25$, $BF_{10} = 14.43$, but not pre-training,

$t(176) = 0.82$, $p = .414$, $d = 0.06$, $BF_{10} = 0.12$. The three-way interaction between US duration, US-CS overlap, and time was not significant, $F(1, 173) = 0.33$, $p = .856$, $\eta_p^2 < .001$, $BF_{incl} = 0.09$.

5.2.3 Discussion

The aim of Experiment 1 was to determine whether contrast effects could be observed on explicit valence ratings and in affective priming for backward CSs that signal the offset of USs of varying durations. To achieve this, we manipulated US offset predictability by varying the duration of the USs and the overlap between the USs and the backward CSs. Explicit valence ratings and affective priming both revealed an assimilation effect at post-training 2 (and post-training 1 for ratings) indicating that varying US duration and US-CS overlap did not result in backward CS contrast effects in a backward conditioning only procedure either individually, or when interacting with each other.

Explicit valence ratings show that CSs following pleasant USs become more pleasant and CSs following unpleasant USs become less pleasant. This assimilation effect is shown to be stronger after the second training phase (post-training 2) than after the first (post-training 1), suggesting that more training trials lead to stronger EC effects. The hypotheses that the backward CS contrast effects observed by Moran et al. (2016) and Moran and Bar-Anan (2013) were driven by the association of an emotional experience (relief or disappointment) with a previously neutral stimulus (backward CS) or that US-CS overlap would enable the backward CSs to signal the offset of the USs resulting in backward CS contrast effects were not supported.

5.3 Experiment 2

The results from Experiment 1 suggest that US offset predictability did not determine the nature of the backward CS effects observed by Moran et al. (2016) and Moran and Bar-Anan (2013). However, in addition to US duration and US-CS overlap, Moran and Bar-Anan (2013) also included a forward CS in a CS-US-CS design which enabled prediction of the onset of the US. Andreatta et al. (2013) presented different participants with a forward or backward fear conditioning procedure in one experiment (CS-US vs. US-CS) and with a within participant forward and backward fear conditioning design in a second (CS-US-CS). They found that when backward fear conditioning was presented alone participants rated the

backward CS presented after an unpredictable aversive electro-tactile US as unpleasant (assimilation effect). However, in the within participant CS-US-CS design, participants rated the backward CS presented after a predictable aversive electro-tactile US as pleasant (contrast effect). These findings suggest that presenting a second CS before a US-CS pairing that provides predictive information about the onset of an aversive US may lead to backward CS contrast effect as observed by Moran and Bar-Anan (2013). Therefore, it is possible that USs must be predictable in order for US offset predictability to influence backward CS evaluations.

To investigate the possibility that backward CS evaluations would differ as a result of US offset predictability when US onset was predictable, we replicated Experiment 1 with the addition of a concurrent forward CS (CS-US-CS). We hypothesised that backward CS contrast effects would occur in the variable duration overlap group and assimilation effects would occur in the three other groups on explicit valence ratings. As in Experiment 1, we predicted assimilation effects for all groups on affective priming. In order to determine whether US predictability influences backward CS evaluations and whether this interacts with US offset predictability and US/CS overlap, we also planned to analyse backward CS evaluations between the two experiments.

5.3.1 Method

Participants. As in Experiment 1, participants were recruited through M-Turk according to a-priori power calculations in G*Power which indicated that 256 participants were required to detect a medium sized interaction ($\eta_p^2 = .07$) at .80 power and $\alpha = .05$ on an ANOVA *F* test (within-between interaction; 4 groups [Fixed-no overlap; Fixed-overlap; Variable-overlap; Variable-no overlap]; 6 measurement points [Pre-training forward; Pre-training backward; Post-training 1 forward; Post-training backward; Post-training 2 forward; Post-training 2 backward]; Effect size specification: as in Cohen [1988]; nonsphericity correction: 1. Note: US duration and US/CS overlap have been collapsed into one factor and conditioning type and time have been collapsed into one factor for the power analysis as G*Power only allows for two factors to be entered). As per Experiment 1, 256 participants were also required to detect significant results on follow-up *t*-tests. The final sample comprised 273 participants (mean age = 37.39, *SD* = 11.09). The fixed duration-no overlap group consisted of 64 participants, the fixed duration-overlap group included 67 participants, the variable duration no-overlap group comprised 66 participants,

and the variable duration-overlap group consisted of 76 participants. Groups did not differ on gender, $\chi^2(3) = 3.85$, $p = .279$, or ethnicity, $\chi^2(15) = 18.41$, $p = .242$. Age was submitted to a 2 (US Duration: fixed vs. variable; between-participants) × 2 (US-CS overlap: no overlap vs. overlap; between-participants) between-groups ANOVA and a two-way interaction was found, $F(1, 262) = 4.90$, $p = .028$, $\eta_p^2 = .018$, $BF_{incl} = 1.76$. Pairwise comparisons showed that the mean age was marginally higher in the variable duration overlap group ($M = 39.04$, $SD = 11.77$) compared to the variable duration no overlap group ($M = 35.47$, $SD = 10.75$), $t(264) = 1.92$, $p = .056$, $d = 0.24$, $BF_{10} = 0.18$. Mean age did not differ between the fixed duration overlap ($M = 36.13$, $SD = 10.20$) and fixed duration no overlap ($M = 38.55$, $SD = 11.27$) groups, $t(264) = 1.23$, $p = .219$, $d = 0.15$, $BF_{10} = 0.26$.

Explicit valence ratings. The measurement of explicit valence ratings was identical to Experiment 1 with the addition of the two forward CSs.

Affective priming task. Each of the four CSs was presented twice with each positive and negative target word for a total of 160 trials. All other details were the same as in Experiment 1.

Recollective memory test. The recollective memory test was identical to Experiment 1 with the addition of the two forward CSs. Results are reported in the supplementary materials at https://osf.io/4mtph/.

Demographics questionnaire. The demographics questionnaire was identical to Experiment 1.

Apparatus/stimuli. The apparatus/stimuli were the same as Experiment 1 with the exception that every participant saw four CSs. CS presentation was fully counterbalanced across participants resulting in 24 trial sequences.

Procedure. The procedure was identical to Experiment 1 with the addition of forward CS presentations on each trial. In the overlap groups the US was presented during the last second of forward CS presentation so that the forward CS and the US overlapped for 1 second. The first second of backward CS presentation overlapped with the final second of US presentation resulting in an overlapping image of the CS and US. CSs were presented for 5s in total during each trial in all groups. As in Experiment 1, fixed duration groups received 5s of total US presentation and variable duration groups received US presentations that varied between 3, 5, and 7s of total US presentation time (see Figure 4 for depiction of trial sequences). The experiment took 35 minutes on average to complete, and participants were compensated US $6.

Statistical analyses. As in Experiment 1, EC scores for explicit valence ratings and priming scores were calculated separately for forward and backward conditioning at pre-training, post-training 1, and post-training 2. As in Experiment 1, participants with more than 25% errors and outliers in the priming task were excluded (n = 66). Of the remaining 207 participants 8.96% of trials were excluded as errors. EC and priming scores were analysed as in Experiment 1 with the addition of the factor Conditioning Type (forward vs backward; within-participants). In order to assess whether US predictability influences the impact of US duration and US-CS overlap, we subjected backward CS evaluations from Experiments 1 and 2 to both Frequentist and Bayesian 2 (Experiment: backward conditioning only vs. forward and backward conditioning; between-participants) × 2 (US duration: fixed vs. variable; between-participants) × 2 (US-CS Overlap: no overlap vs. overlap; between-participants) × 3 (Time: pre-training vs post-training 1 vs post-training 2; within-participants) mixed ANOVAs. All other details were the same as in Experiment 1.

5.3.2 Results

Explicit valence ratings. Mean EC scores depicted in Figure 5 suggest that assimilation effects at post-training 1 and 2 occurred for forward conditioning in all groups and in the overlap groups only for backward conditioning. Main effects of conditioning type, $F(1, 269) = 97.98$, $p < .001$, $\eta_p^2 = .267$, $BF_{10} = 3.65 \times 10^{17}$, and time, $F(2, 268) = 223.65$, $p < .001$, $\eta_p^2 = .54$, $BF_{10} = 2.90 \times 10^{83}$, and two-way interactions between conditioning type and time, $F(2, 268) = 60.71$, $p < .001$, $\eta_p^2 = .312$, $BF_{incl} = 1.27 \times 10^{13}$, US-CS overlap and conditioning type, $F(1, 269) = 9.12$, $p = .003$, $\eta_p^2 = .033$, $BF_{incl} = 56.02$, and US-CS overlap and time, $F(2, 268) = 8.55$, $p < .001$, $\eta_p^2 = .060$, $BF_{incl} = 15.67 \times 10^3$, were qualified by a three-way interaction between US-CS overlap, conditioning type, and time, $F(2, 268) = 6.78$, $p = .001$, $\eta_p^2 = .048$, $BF_{incl} = 2.17$. The two-way interaction between US duration and conditioning type was also significant, $F(1, 269) = 4.76$, $p = .030$, $\eta_p^2 = .017$, $BF_{incl} = 2.51$. The four-way interaction between US duration, US-CS overlap, conditioning type, and time was not significant, $F(2, 268) = 1.75$, $p = .176$, $\eta_p^2 = .013$, $BF_{incl} = 0.12$. Pairwise comparisons for the three-way interaction between US-CS overlap, conditioning type, and time revealed larger EC scores in the overlap groups than the no overlap groups at post-training 2 for forward conditioning, $t(271) = 2.06$, $p = .041$, $d = 0.25$, $BF_{10} = 0.84$, and larger EC scores in the overlap groups than the no

overlap groups for backward conditioning at post-training 1, $t(271) = 5.54$, $p < .001$, $d = 0.67$, $BF_{10} = 1.91 \times 10^5$, and post-training 2, $t(271) = 6.35$, $p < .001$, $d = 0.77$, $BF_{10} = 1.30 \times 10^7$. EC scores were also larger in the overlap groups than the no overlap groups at post-training 1 for forward conditioning, however, this comparison was marginal, $t(271) = 1.84$, $p = .068$, $d = 0.22$, $BF_{10} = 0.58$. Table 1 depicts test statistics from one sample *t*-tests showing that EC scores were greater than zero for forward and backward conditioning in both the no overlap and overlap groups at post-training 1 and post-training 2, but not pre-training.

Pairwise comparisons for the two-way interaction between US duration and conditioning type revealed assimilation effects for both conditioning types in both US duration groups. Significant differences were found between the assimilation effects for forward and backward conditioning in both the fixed and variable duration groups. Difference scores were calculated between forward and backward conditioning in each group and were subjected to an independent samples *t*-test to determine what was driving the interaction. The difference between forward and backward conditioning assimilation effects was larger in the fixed duration group than the variable duration group, $t(271) = 2.27$, $p = .024$, $d = 0.28$, $BF_{10} = 1.51$.

Affective priming. Mean priming scores depicted in Figure 6 suggest an assimilation effect at post-training 2 in the fixed groups for forward conditioning. A main effect of conditioning type, $F(1, 203) = 4.37$, $p = .038$, $\eta_p^2 = .021$, $BF_{10} = 1.89$, was qualified by a Conditioning Type × Time interaction, $F(1, 203) = 5.91$, $p = .016$, $\eta_p^2 = .028$, $BF_{incl} = 0.28$. The four-way interaction between US variability, US/CS overlap, conditioning type, and time was not significant, $F(1, 203) = 0.44$, $p = .510$, $\eta_p^2 = .002$, $BF_{incl} = 0.28$. Follow-up analyses of the significant two-way interaction between conditioning type and time revealed larger priming scores at post-training 2 in comparison to pre-training for forward conditioning, $t(206) = 2.97$, $p = .003$, $d = 0.21$, $BF_{10} = 5.57$, but not backward conditioning, $t(206) = 0.22$, $p = .824$, $d = 0.02$, $BF_{10} = 0.08$. One sample *t*-tests showed that priming scores were greater than zero for forward conditioning at post-training 2 only, $t(206) = 4.98$, $p < .001$, $d = 0.35$, $BF_{10} = 8.19 \times 10^3$, all other $t < 1.18$, $p > .238$, $d < .08$, $BF_{10} < 0.15$.

Explicit valence ratings – cross experiment comparison. Figure 7 shows mean EC scores for backward CSs only across experiments as a function of US duration, US-CS overlap, and time. The figure suggests that backward CS assimilation effects at post-training 1 and 2 were found in the overlap groups in both

experiments and only in Experiment 1 but not Experiment 2 in the no overlap groups. A main effect of time, $F(2, 518) = 185.09$, $p < .001$, $\eta_p^2 = .417$, $BF_{10} = 1.90 \times 10^{24}$, and two-way interactions between experiment and time, $F(2, 518) = 23.97$, $p < .001$, $\eta_p^2 = .085$, $BF_{incl} = 3.23 \times 10^5$, and US-CS overlap and time, $F(2, 518) = 9.09$, $p < .001$, $\eta_p^2 = .034$, $BF_{incl} = 1.44$, were qualified by a significant three-way interaction between experiment, US-CS overlap, and time, $F(2, 518) = 5.10$, $p = .006$, $\eta_p^2 = .019$, $BF_{incl} = 0.03$. The four-way interaction between experiment, US duration, US-CS overlap, and time was not significant, $F(1, 518) = 1.53$ $p = .218$, $\eta_p^2 = .006$, $BF_{incl} = 0.05$.

Follow-up analyses of the three-way interaction should be interpreted cautiously due to the Bayes factor opposing the frequentist results. EC scores in the no overlap condition were larger in Experiment 1 than Experiment 2 at post-training 1, $F(1, 519) = 23.18$ $p < .001$, $\eta_p^2 = .043$, $BF_{10} = 3.43 \times 10^4$, and post-training 2, $F(1, 519) = 54.70$ $p < .001$, $\eta_p^2 = .095$, $BF_{10} = 4.58 \times 10^{11}$, but not pre-training, $F(1, 519) = 0.01$ $p = .919$, $\eta_p^2 < .001$, $BF_{10} = 0.34$. In the overlap condition, EC scores were larger in Experiment 1 than in Experiment 2 at post-training 2 only, $F(1, 519) = 9.54$ $p = .002$, $\eta_p^2 = .018$, $BF_{10} = 7.05$. No differences were found at post-training 1, $F(1, 519) = 1.91$ $p = .168$, $\eta_p^2 = .004$, $BF_{10} = 0.31$, or pre-training, $F(1, 519) = 0.47$, $p = .492$, $\eta_p^2 = .001$, $BF_{10} = 0.86$.

Affective priming – cross experiment comparison. Mean priming scores depicted in Figure 8 suggest assimilation effects at post-training 2 in the overlap groups in Experiment 1. A main effect of time, $F(1, 376) = 5.30$, $p = .022$, $\eta_p^2 = .014$, $BF_{10} = 1.02$, was qualified by a two-way interaction between Experiment and time, $F(1, 376) = 6.77$, $p = .010$, $\eta_p^2 = .018$, $BF_{incl} = 2.29$. The four-way interaction between Experiment, US duration, US-CS overlap, and time was not significant, $F(1, 376) = 0.95$, $p = .329$, $\eta_p^2 = .003$, $BF_{incl} = 0.81$. Pairwise comparison showed that there was no difference between pre-training and post-training 2 in Experiment 2, $t(383) = 0.22$, $p = .826$, $d = 0.01$, $BF_{10} = 0.08$, and that priming scores at post-training 2 were significantly larger than at pre-training in Experiment 1, $t(383) = 3.34$, $p = .001$, $d = 0.17$, $BF_{10} = 15.58$. One sample *t*-tests revealed that priming scores in Experiment 2 did not differ from zero at pre-training, $t(206) = 0.94$, $p = .349$, $d = 0.07$, $BF_{10} = 0.12$, or post-training 2, $t(206) = 0.91$, $p = .365$, $d = 0.06$, $BF_{10} = 0.12$, and that in Experiment 1 priming scores differed significantly from zero at

post-training 2, $t(176) = 3.28$, $p = .001$, $d = 0.25$, $BF_{10} = 14.43$, but not pre-training, $t(176) = 0.82$, $p = .414$, $d = 0.06$, $BF_{10} = 0.12$.

5.4 General Discussion

The aim of the current investigation was to assess whether manipulating US offset predictability by varying US duration and US/CS overlap would elicit backward CS contrast effects on explicit valence ratings. This was assessed in a backward conditioning only (US-CS; Experiment 1) and in a concurrent forward and backward conditioning design (CS-US-CS: Experiment 2). We included a forward CS in Experiment 2 to render US onset predictable as Andreatta et al. (2013) found backward CS contrast effects only when the US was predicted by a forward CS (CS-US-CS). Explicit valence ratings revealed only backward conditioning assimilation effects, suggesting that low US offset predictability does not lead to backward CS contrast effects regardless of whether the US is predicted or not. Moreover, these assimilation effects were larger in the overlap groups than the no overlap groups, providing strong evidence against our hypothesis that past findings of contrast effects in evaluative backward conditioning occurred due to a combination of US onset and offset predictability. The current investigation was not designed to provide support for or against current EC theories, however, our aim was to determine the effects of procedural features that may affect backward evaluative conditioning and thus have implications for theorising about the mechanisms underlying EC.

In Experiment 2, affective priming revealed assimilation effects for forward conditioning only, regardless of group, while explicit valence ratings showed that CSs paired with pleasant USs became more pleasant and CSs paired with unpleasant USs became less pleasant, for both forward and backward conditioning. These explicit assimilation effects were found in the no overlap and overlap groups, regardless of US duration. For forward conditioning, these effects were significantly larger in the overlap groups than the no overlap groups at post-training 2, and for backward conditioning, these effects were significantly larger in the overlap groups than the no overlap groups at post-training 1 and post-training 2. The finding that backward CS assimilation effects were larger in the overlap groups regardless of US duration contradicts our hypothesis that rendering a US predictable while presenting USs of variable duration that overlap with backward CSs would lead to backward CS

contrast effects. Rather than highlighting information about the US offset that elicits an opposing emotional reaction, it appears that overlapping the USs and CSs results in simultaneous conditioning that increases the size of the assimilation effect. This suggests that the backward CS contrast effects reported in past literature are in fact driven by the instructional manipulation. Moreover, these instructions appear to have an even larger effect than originally assumed, as according to our findings they are reversing the effect of simultaneous conditioning due to US/CS overlap.

Results from the current studies differ from those reported in the backward conditioning only (US-CS) pilot study which we were attempting to confirm. In the pilot study, explicit valence ratings revealed a significant three-way interaction between US duration, US/CS overlap, and time, and only a marginal backward CS contrast effect in the variable overlap condition when comparing the mean to zero. Affective priming data revealed no significant main effects or interactions. When taken together with the findings from the current report and given that the pilot study was underpowered with little evidence for backward CS contrast effects, it appears that the findings from the pilot study were unreliable. Thus, we were unable to confirm our predictions with a larger sample as originally intended.

Findings from the current studies are also in line with those of Green et al. (2019), who found that instructions highlighting the role of the CSs were required to elicit backward CS contrast effects on two conditioning paradigms that differed on CS and US presentation parameters (such as duration of stimulus presentations and number and type of CSs and USs). In other words, neither paradigm was conducive to backward CS contrast effects without the instructional manipulation. With the exception of a fear conditioning study employing an intense electro-tactile US (Andreatta et al., 2013), no support for the emergence of backward CS contrast effects without an instructional manipulation has been found. This suggests that backward conditioning leads to assimilation effects, unless either instructional manipulations that change the relationship between CSs and USs, or intense electro-tactile stimuli are included in a concurrent forward and backward conditioning design.

Comparing backward conditioning between Experiments 1 and 2 showed that EC scores were larger in Experiment 1 than Experiment 2 at post-training 1 and 2 for the no overlap groups and at post-training 2 only for the overlap groups. This pattern was not moderated by US duration. The large backward CS assimilation effects

observed in the overlap groups in Experiment 1 are likely the result of simultaneous presentation of USs and backward CSs, while the almost non-existent assimilation effects in the no overlap groups when the US was predictable (Experiment 2) suggest that the presence of a forward CS appears to inhibit backward CS learning. The conclusion that the presence of a forward CS inhibits backward CS learning was also supported by affective priming scores, as assimilative backward conditioning was evident in Experiment 1, but not in Experiment 2.

A plausible alternative explanation for the difference in EC scores between Experiments 1 and 2 is that adding a forward CS resulted in some participants showing contrast effects and others assimilation effects. However, when removing participants whose EC score suggested a contrast effect (-1 or less) in both experiments the pattern remained the same, i.e. EC scores were larger in Experiment 1 than Experiment 2. Thus, some participants showing contrast effects and others showing assimilation effects due to the addition of the forward CS was not responsible for the difference in EC scores between experiments. These findings suggest that having a forward CS results in smaller backward conditioning overall when compared with backward conditioning only, meaning that paradigms with a forward CS would render backward CSs more conducive to contrast effects because any other variable (i.e. instructions) capable of eliciting contrast effects would have to compete with smaller assimilation effects.

The inhibition of backward conditioning caused by the presence of a forward CS in Experiment 2 may be explained by temporal overshadowing. Overshadowing occurs when two neutral stimuli are presented in a compound followed by a US, resulting in the more salient of the two neutral stimuli acquiring greater conditional responding than the less salient neutral stimulus (Mackintosh, 1975; Pavlov, 1927). If the principle of overshadowing can be extended to stimuli that are presented sequentially (temporal overshadowing), then the forward CS could be considered the most salient CS in a CS-US-CS arrangement as it predicts the onset of the US. This would mean that the forward CS would acquire the majority of the associative strength supported by the US, leaving only a limited amount available for the backward CS, as demonstrated by the attenuated assimilation effects in the no overlap groups in Experiment 2. There is little research to date investigating overshadowing in EC and the research that has been performed reveals mixed results, however, there is evidence to suggest that overshadowing can occur in EC (i.e.,

Kattner & Green, 2015; Purkis & Lipp, 2010; Walther, Ebert, & Meinerling, 2011; but see also, Dwyer, Jarratt, & Dick, 2007). Thus, future research should investigate whether overshadowing can emerge across time and whether temporal overshadowing is capable of explaining the current data.

The results of the between experiment analyses for the affective priming data need to be interpreted with caution as the priming task used in Experiment 2 may not provide a true indication of backward CS learning. Both forward CS primes and backward CS primes were presented in a single affective priming task, and the presence of the forward CS primes, which were evaluated as more pleasant or unpleasant than the backward CS primes, may have limited the effects of backward CS primes. If so, the lack of backward CS priming in Experiment 2 may not indicate weaker backward evaluative conditioning as suggested above, but an anchoring effect in comparison to forward CS primes. Future research should assess the implicit valence of forward and backward CSs in separate tasks to avoid this potential confound (i.e. Moran & Bar-Anan, 2013; but see also Bading et al., 2020). Anchoring could be argued to have affected explicit valence ratings as well, as forward CSs were also presented together with backward CSs in a randomized sequence during the assessments. However, this appears unlikely as the between experiment differences were not consistent across the experimental conditions.

In conclusion, the current study revealed assimilation effects for backward conditioning in both experiments. This assimilation effect was larger in Experiment 2 when CSs and USs overlapped (simultaneous conditioning) than when they did not overlap. Additionally, backward conditioning was smaller in Experiment 2 when concurrent forward conditioning was presented (CS-US-CS) than in Experiment 1 when only backward conditioning (US-CS) was presented. These findings suggest that presenting concurrent forward conditioning has an inhibitory effect on backward EC. Moreover, this inhibitory effect was not as large when CSs and USs overlapped, suggesting that the US-CS relationship can be influenced by other aspects of the acquisition phase which results in different expressions of backward CS valence.

5.5 References

Andreatta, M., Mühlberger, A., Glotzbach-Schoon, E., & Pauli, P. (2013). Pain predictability reverses valence ratings of a relief-associated stimulus. Frontiers in Systems Neuroscience, 7(53), 1-12. doi: 10.3389/fnsys.2013.00053

Andreatta, M., Mühlberger, A., Yarali, A., Gerber, B., & Pauli, P. (2010). A rift between implicit and explicit conditioned valence in human pain relief learning. Proceedings of the Royal Society of London B: Biological Sciences, 277, 2411-2416. doi: 10.1098/rspb.2010.0103

Bading, K., Stahl, C., & Rothermund, K. (2020). Why a standard IAT effect cannot provide evidence for association formation: The role of similarity construction. Cognition and Emotion, 34,128-143. doi: 10.1080/02699931.2019.1604322

Cohen, J. (1988). Statistical power analysis for the behavioural sciences. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

De Houwer, J. (2007). A conceptual and theoretical analysis of evaluative conditioning. The Spanish Journal of Psychology, 10, 230-241. doi: 10.1017/S1138741600006491

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. Psychological Bulletin, 135, 347-368. doi: 10.1037/a0014211

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, 853-869. doi: 10.1037//0033-2909.127.6.853

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

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, 1013-1027. doi: 10.1146/annurev.psych.54.101601.145225

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

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

Galdi, S., Arcuri, L., & Gawronski, B. (2008). Automatic mental associations predict future choices of undecided decision makers. Science, 321, 1100-1102. doi: 10.1126/science.1160769

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.

Gibson, B. (2008). Can evaluative conditioning change attitudes toward mature brands? New evidence from the Implicit Association Test. Journal of Consumer Research, 35, 178-188. doi: 10.1086/527341

Green, L. J. S., Luck, C. C., Gawronski, B., & Lipp, O. V. (2019) Contrast effects in backward evaluative conditioning: Exploring effects of affective relief/disappointment versus instructional information. Emotion. Advance online publication. doi: 10.1037/emo0000701

Green, L. J. S., Luck, C. C., & Lipp, O. V. (2020). How disappointing: Startle modulation reveals conditional stimuli presented after pleasant unconditional stimuli acquire negative valence. Psychophysiology. Advance online publication. doi: 10.1111/psyp.13563

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. Journal of Personality and Social Psychology, 74, 1464-1480. doi: 10.1037/0022-3514.74.6.1464

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. Psychological Bulletin, 136, 390-421. doi: 10.1037/a0018916

Hu, X., Gawronski, B., & Balas, R. (2017a). 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,* 17-32.

Hu, X., Gawronski, B., & Balas, R. (2017b). Propositional versus dual-process accounts of evaluative conditioning: II. The effectiveness of counter-conditioning and counter-instructions in changing implicit and explicit evaluations. *Social Psychological and Personality Science, 8,* 858-866.

Inquisit 4 [Computer software]. (2016). Retrieved from https://www.millisecond.com.

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

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

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. doi: 10.1016/j.jesp.2019.103905

Lang, P.J., Bradley, M.M., & Cuthbert, B.N. (2008). International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8. University of Florida, Gainesville, FL.

LeBel, E. P., & Campbell, L. (2009). Implicit partner affect, relationship satisfaction, and the prediction of romantic breakup. Journal of Experimental Social Psychology, 45, 1291-1294. doi: 10.1016/j.jesp.2009.07.003

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

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. Behavior Research Methods, 49, 433-442. doi: 10.3758/s13428-016-0727-z

Luck, C. C., & Lipp, O. V. (2017). Startle modulation and explicit valence evaluations dissociate during backward fear conditioning. Psychophysiology, 54, 673-683. doi: 10.1111/psyp.12834

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. Psychological Review, 82, 276–298.

Mallan, K. M., Lipp, O. V., & Libera, M. (2008). Affect, attention, or anticipatory arousal? Human blink startle modulation in forward and backward affective conditioning. International Journal of Psychophysiology, 69. 9-17. doi: 10.1016/j.ijpsycho.2008.02.005

Mathôt, S. (2017) Bayes like a Baws: Interpreting Bayesian repeated measures in JASP. Retrieved from https://www.cogsci.nl/blog/interpreting-bayesian-repeated-measures-in-jasp

Moran, T., & Bar-Anan, Y. (2013). The effect of object-valence relations on automatic evaluation. Cognition and Emotion, 27, 743-752. doi: 10.1080/02699931.2012.732040Moran, T., Bar-Anan, Y., & Nosek, B. A. (2016). The assimilative effect of co-occurrence on evaluation above and

beyond the effect of relational qualifiers. Social Cognition, 34. 435-461. doi: 101521soco2016345435

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

Pavlov, I. P. (1927). Conditioned reflexes. Oxford: Oxford University Press.

Pleyers, G., Corneille, Luminet, O., & Yzerbyt, V. (2007). Aware and (dis)liking: Item-based analyses reveal that valence acquisition via evaluative conditioning emerges only when there is contingency awareness. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33, 130-144. doi: 10.1037/0278-7393.33.1.130

Purkis, H. M., & Lipp, O. V. (2010). Stimulus competition in pre/post and online ratings in an evaluative learning design. Learning and Motivation, 41, 84-94. doi: 10.1016/j.lmot.2009.12.001

Staats, C. K., & Staats, A. W. (1957). Meaning established by classical conditioning. Journal of Experimental Psychology, 54, 74–80. doi: 10.1037/h0047716

Walther, E., Ebert, I., & Meinerling, K. (2011). Does cue competition reduce conditioned liking of brands and products? Psychology and Marketing, 28, 520–538. doi:10.1002/mar.20399.

5.6 Figures and Tables

Table 1: *One-sample t-test statistics from the US-CS overlap, conditioning type, and time interaction.*

	df	*t*	*p*	*d*	BF_{10}
N_F_Pre	129	1.33	.185	0.12	0.23
N_F_Post 1	129	9.38	< .001	0.82	1.97×10^{13}
N_F_Post 2	129	12.75	< .001	1.03	2.95×10^{21}
N_B_Pre	129	0.77	.443	0.07	0.13
N_B_Post 1	129	3.10	.002	0.27	9.02
N_B_Post 2	129	3.27	< .001	0.29	14.87
O_F_Pre	142	< .01	> .999	< 0.01	0.09
O_F_Post 1	142	13.31	< .001	1.11	3.73×10^{23}
O_F_Post 2	142	16.91	< .001	1.41	3.50×10^{32}
O_B_Pre	142	.52	.603	0.04	0.11
O_B_Post 1	142	8.80	< .001	0.74	1.47×10^{12}
O_B_Post 2	142	10.13	< .001	0.85	2.94×10^{15}

Note. N = No Overlap Group. O = Overlap Group. F = Forward Conditioning. B = Backward Conditioning. Pre = Pre-Training. Post 1 = Post-Training 1. Post 2 = Post-Training 2.

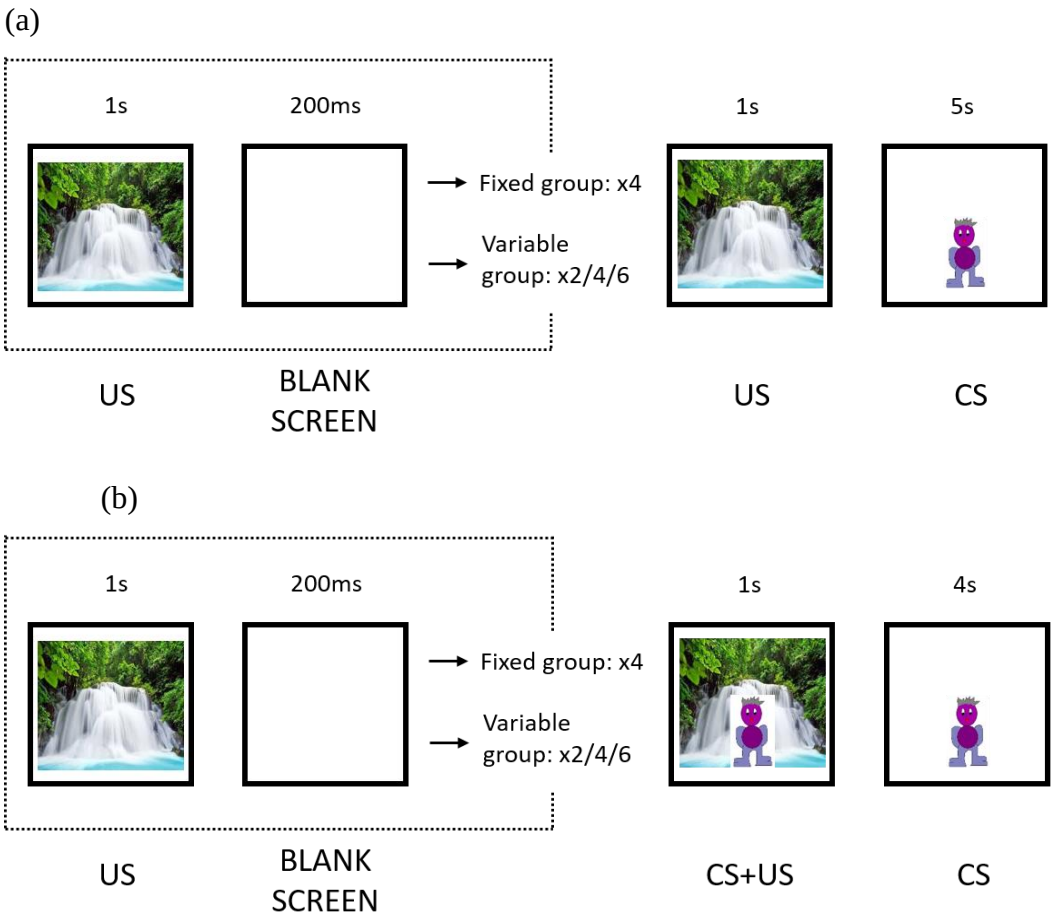

Figure 1. Trial sequence for no overlap (a) and overlap (b) groups in Experiment 1. In the fixed no overlap and overlap groups USs were flashed 5 times for 1s each by presenting a blank screen for 200ms after the first 4 USs. In the fixed no overlap group backward CS onset coincided with the offset of the fifth 1s US presentation. In the fixed overlap group the fifth 1s US presentation was presented with the CS overlaying the US. Backward CS onset coincided with CS+US offset and backward CSs were presented for 4s alone resulting in 5s of total backward CS presentation. In the variable no overlap (a) and variable overlap (b) groups USs were flashed 3, 5, or 7 times for 1s each by presenting a blank screen for 200ms after the first 2, 4, or 6 USs. In the variable no overlap group backward CS onset coincided with the offset of the third, fifth, or seventh 1s US presentation and backward CSs were presented for 5s each. In the variable overlap group the third, fifth, and seventh 1s US presentations were presented with the CS overlaying the US. Backward CS onset coincided with CS+US offset and backward CSs were presented for 4s alone resulting in 5s of total backward CS presentation.

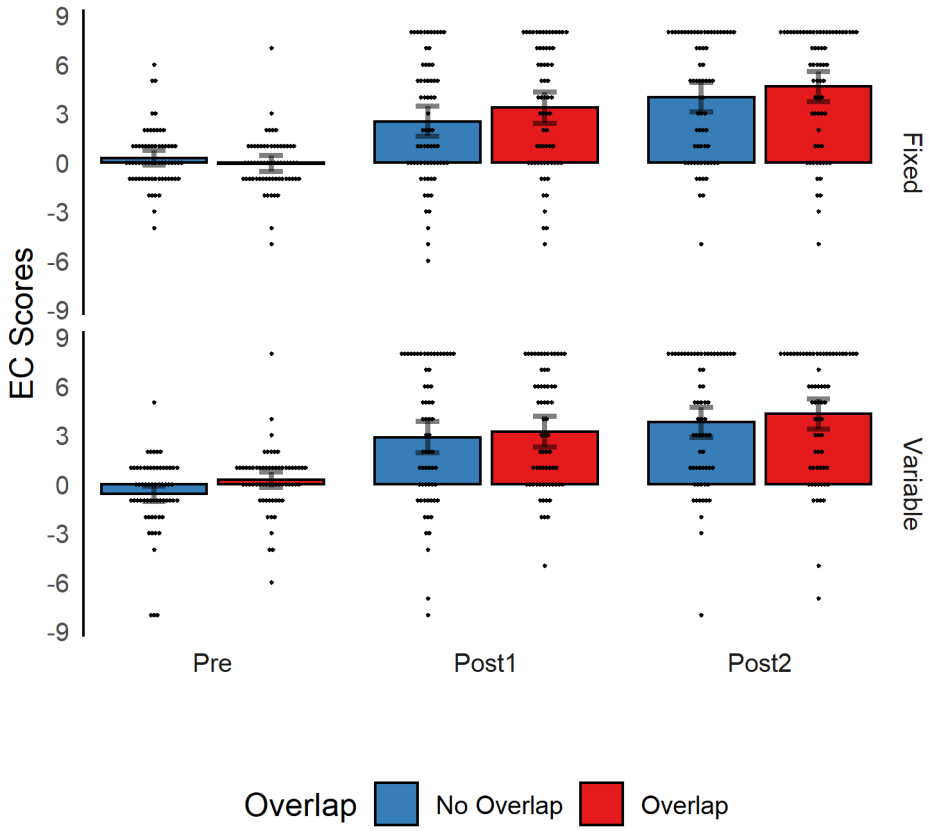

Figure 2. EC scores for backward conditioning measured at pre-training (Pre), post-training 1 (Post1), and post-training 2 (Post2), as a function of US-CS overlap ('No overlap' and 'Overlap') and US duration ('Fixed' and 'Variable') in Experiment 1. Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.

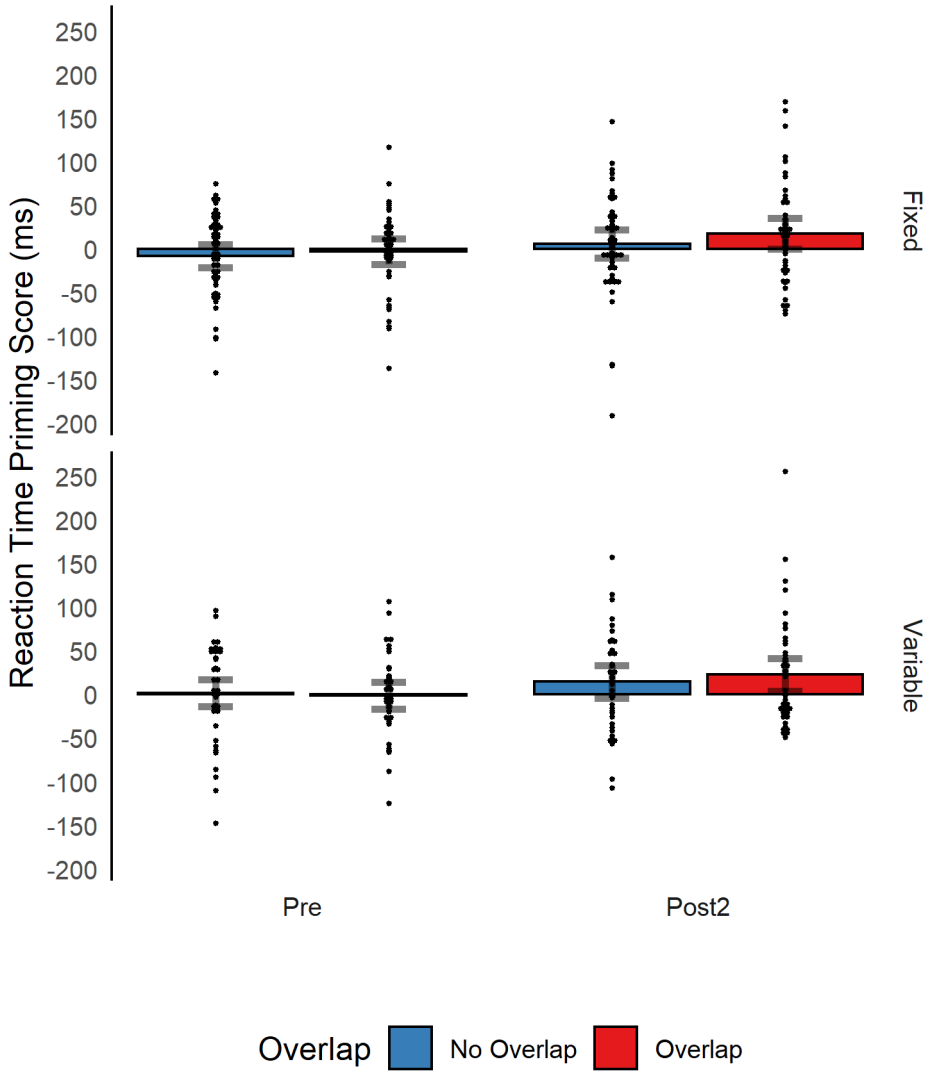

Figure 3. Priming scores for backward conditioning measured at pre-training (Pre) and post-training 2 (Post2) as a function of US-CS overlap ('No overlap' and 'Overlap') and US duration ('Fixed' and 'Variable') in Experiment 1. Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.

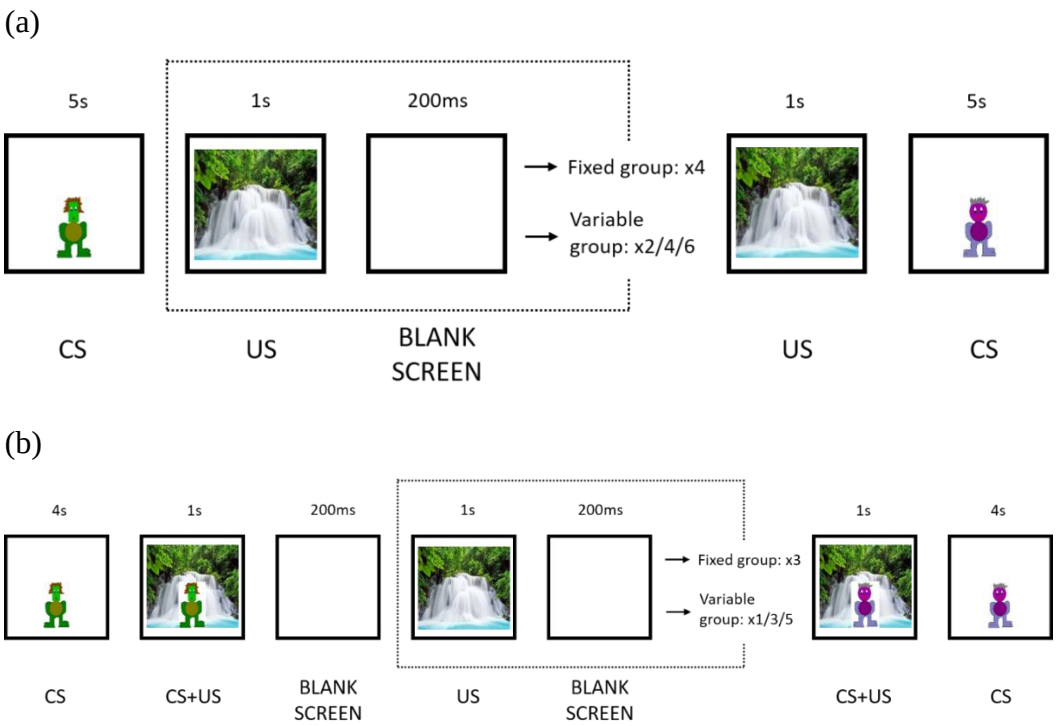

Figure 4. Trial sequence for no overlap (a) and overlap (b) groups in Experiment 2. In the fixed no overlap group (a) forward CSs were presented for 5s each and forward CS offset coincided with US onset. USs were flashed 5 times for 1s each by presenting a blank screen for 200ms after the first 4 USs. Backward CS onset coincided with the offset of the fifth 1s US presentation and backward CSs were presented for 5s. In the variable no overlap group (a) forward CSs were presented for 5s each and forward CS offset coincided with US onset. USs were flashed 3, 5, or 7 times for 1s each by presenting a blank screen for 200ms after the first 2, 4, or 6 USs. Backward CS onset coincided with the offset of the third, fifth, or seventh 1s US presentation and backward CSs were presented for 5s.

In the fixed overlap group (b) forward CSs were presented for 4s alone and overlaying the US for 1s resulting in 5s of total forward CS presentation. A 200ms blank screen was presented and USs were flashed 4 more times for 1s each by presenting a blank screen for 200ms after the following 3 USs. The fifth 1s US presentation was presented with the CS overlaying the US. In the variable overlap group (b) forward CSs were presented for 4s alone and overlaying the US for 1s resulting in 5s of total forward CS presentation. A 200ms blank screen was presented and USs were flashed 2, 4, or 6 more times for 1s each by presenting a blank screen for 200ms after the following 1, 3, or 5 USs. The third, fifth, and seventh 1s US presentations were presented with the CS overlaying the US. In both overlap groups

backward CS onset coincided with CS+US offset and backward CSs were presented for 4s alone resulting in 5s of total backward CS presentation.

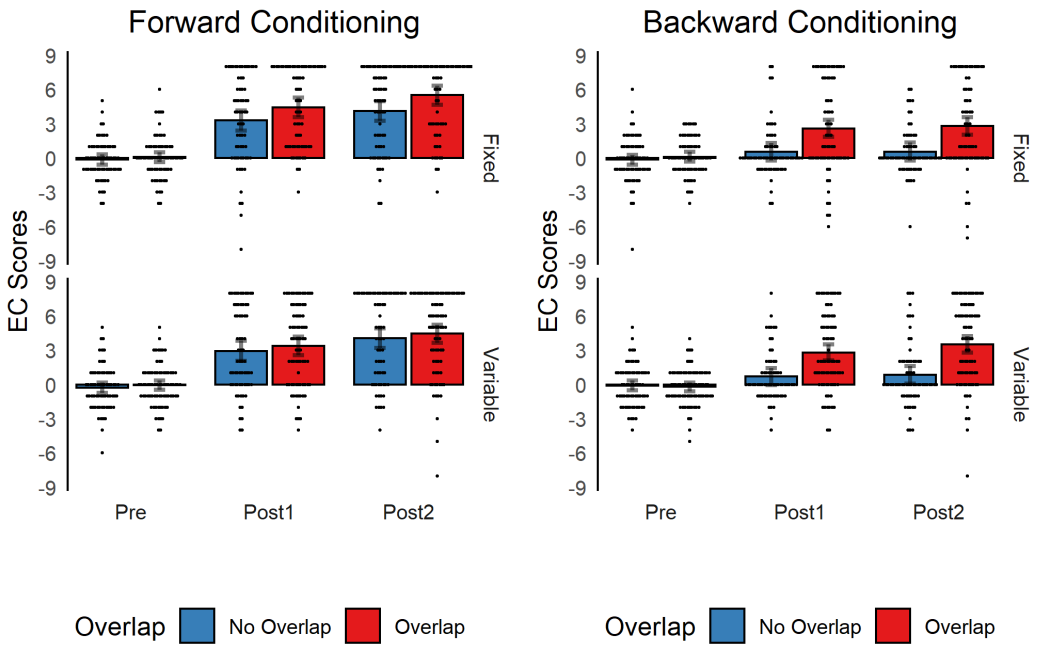

Figure 5. EC scores for forward and backward conditioning measured at pre-training (Pre), post-training 1 (Post1), and post-training 2 (Post2), as a function of US-CS overlap ('No overlap' and 'Overlap') and US duration ('Fixed' and 'Variable') in Experiment 2. Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.

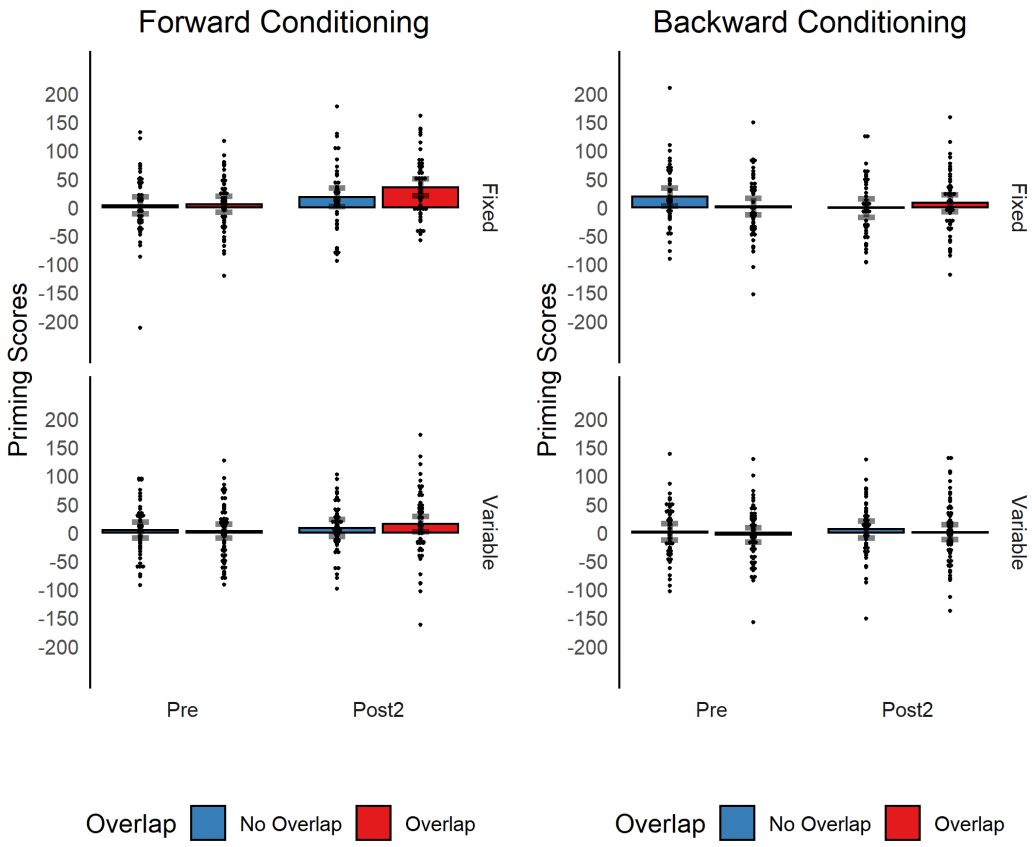

Figure 6. Priming scores for forward and backward CSs measured at pre-training (Pre) and post-training 2 (Post2) as a function of US-CS overlap ('No overlap' and 'Overlap') and US duration ('Fixed' and 'Variable') in Experiment 2. Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.

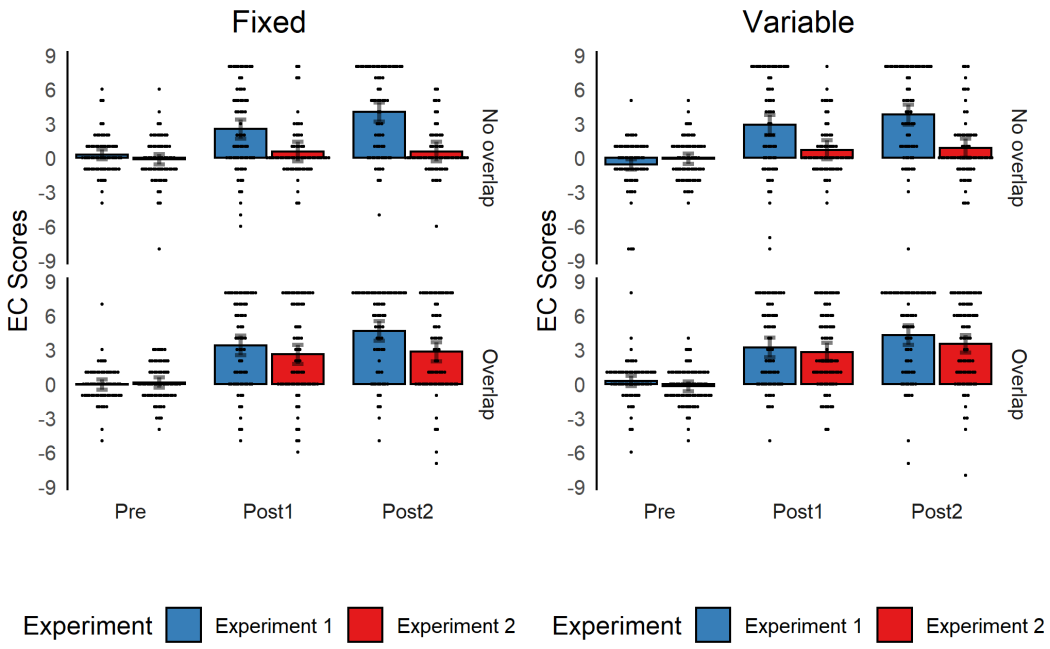

Figure 7. EC scores for backward CSs measured at pre-training (Pre), post-training 1 (Post1), and post-training 2 (Post2), as a function of Experiment (Experiment 1 and Experiment 2), US-CS overlap ('No overlap' and 'Overlap'), and US duration ('Fixed' and 'Variable'). Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.

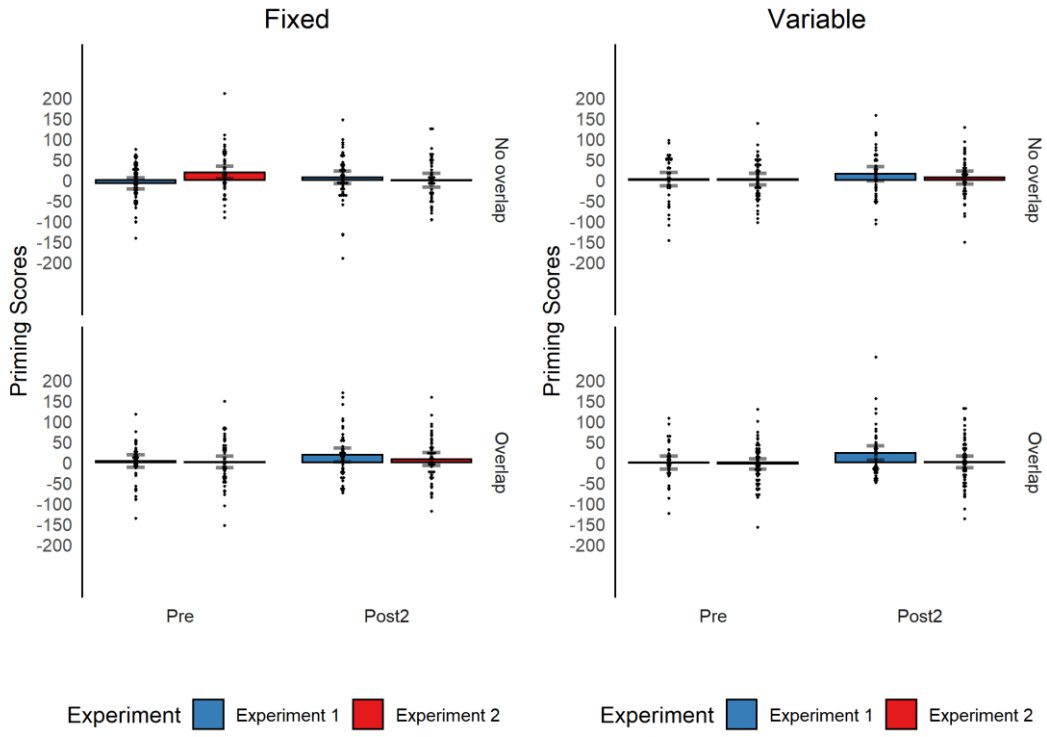

Figure 8. Priming scores for backward CSs at pre-training (Pre) and post-training 2 (Post2) as a function of Experiment (Experiment 1 and Experiment 2), US-CS overlap ('No overlap' and 'Overlap'), and US duration ('Fixed' and 'Variable'). Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.

5.7 Supplementary Material: Pilot Study Data

A 2 × 2 factorial design was employed to assess the effects of unconditional stimulus (US) duration variability (fixed vs. variable) and US-conditional stimulus (CS) overlap (overlap vs. no overlap) in a simple backward conditioning procedure. Explicit valence ratings and affective priming were measured before and after the training phase. Based on the paradigm used by Moran and Bar-Anan (2013) which presented variable USs with US/CS overlap, we predicted that backward CS contrast effects on explicit valence ratings would occur after training in the variable duration-overlap condition only. Backward CS assimilation effects on explicit valence ratings were predicted for the fixed duration-no overlap, fixed duration-overlap, and variable duration-no overlap groups. Implicit measures have only revealed backward CS contrast effects when instructional manipulations highlighting the relation between CSs and USs are presented on every trial, and therefore we predicted assimilation effects on affective priming for all groups (Hu, Gawronski, & Balas, 2017a; see also Bading, Stahl, & Rothermund, 2019).

5.7.1 Method

Design and Participants. A 2 (US Duration: fixed vs. variable; between-participants) × 2 (US-CS overlap: no overlap vs. overlap; between-participants) × 2 (US Valence: positive vs. negative; within-participant) × 2 (Time: pre-training vs post-training; within-participant) mixed design was used to determine the effect of US duration variability and US-CS overlap on CS valence after a backward conditioning procedure. Following ethical approval from the Curtin University Human Research Ethics Board, 96 students (mean age = 23.24, *SD* = 7.34; 25 male) *were* recruited from the School of Psychology at Curtin University in exchange for course credit resulting in 24 participants per group.

Apparatus/stimuli. Four cartoon aliens created by Moran and Bar-Anan (2013) were used as CSs (materials from Moran and Bar-Anan, 2013, available at https://osf.io/cqsnj/). Each alien differed in colour and head shape. Four positive pictures depicting pleasant sceneries and baby animals and four negative pictures depicting aggressive humans and animals from the International Affective Picture System (IAPS; Lang, Bradley, & Cuthbert, 1999) were used as USs (1050, 1300, 1440, 1710, 5833, 6313, 6560, and 8190). DMDX (Forster & Forster 2003) was used to run the experiment and to record responses in all tasks.

Explicit valence ratings. Participants were shown the two backward CSs and the two CSs that were not presented during training and asked to rate how pleasant they found the stimuli on a 9-point scale ranging from 1 (*unpleasant*) to 9 (*pleasant*).

Affective priming task. Each CS was presented twice with 10 positive target words and 10 negative target words for a total of 80 trials. This included the two backward CSs and two CSs that were not presented during training to serve as neutral CSs. Following a 500ms fixation cross, the CS prime was presented for 200ms followed by the target word until the participant provided their response. Participants were instructed to press the *I* key if the target word was positive and the *E* key if the target word was negative. Target words were taken from Hu et al. (2017a, 2017b). The positive words were *pleasant, good, outstanding, beautiful, magnificent, marvellous, excellent, appealing, delightful,* and *nice*. The negative words were *unpleasant, bad, horrible, miserable, hideous, dreadful, painful, repulsive, awful,* and *ugly.*

Recollective Memory Test. The current study also included measures of recollective memory for exploratory purposes, however, these data were not analysed.

Demographics questionnaire. Participants provided their age, gender, and ethnicity.

Procedure. Participants read the information sheet and provided informed consent by signing the consent form. Explicit valence ratings and affective priming were presented followed by the training phase. The training phase comprised 12 positive and 12 negative trials pseudo-randomly presented with inter-trial intervals of 4, 6, and 8 seconds. Each trial consisted of a pleasant or unpleasant US, followed by a backward CS. We used one CS from each of the four alien families from Moran and Bar-Anan (2013) and four positive and four negative pictures as USs. The four CSs were counter-balanced across participants using a Latin-square resulting in four training sequences. USs were flashed for 1s followed by a blank screen for 200ms resulting in 5 x 1s flashes for a total of 5s presentation time in the fixed groups. USs were presented for either 3, 5, or 7s of total presentation time in the variable groups. In the overlap groups the 1st second of CS presentation was presented within the final US presentation resulting in an overlapping image of the CS and US as in Pleyers, Corneille, Luminet, and Yzerbyt (2007). CSs were presented for 5s in total during each trial in all groups.

Participants received the following instructions:

In this task you will be presented with a series of pictures. Please pay attention to which pictures follow each other as you will be tested on this at the end of the experiment.

After the training phase was completed participants performed a final round of explicit valence ratings and affective priming. This was followed by the memory test and demographics questionnaire. The experiment took approximately 25 minutes on average to complete.

Statistical analyses. IBM SPSS Statistics 25 was used for data analysis. EC scores for explicit valence ratings were calculated by subtracting valence ratings of CSs paired with unpleasant USs (CSu) from those of CSs paired with pleasant USs (CSp; CSp – CSu). The CSs that were not presented were not included in the analyses. EC scores were calculated separately for pre and post-training. Assimilation effects were reflected by positive EC scores and contrast effects by negative EC scores. In the affective priming task, trials were considered as error trials when the target word was incorrectly categorised and when trials were considered outliers. Outliers were classed as trials on which reaction times were shorter than 200ms or longer than 1000ms, or three times the standard deviation of the participants' average response, as they were deemed to be outside the window of a valid response. Participants were removed from the analyses if the percentage of error trials was larger than 25% (n = 1). Priming scores were calculated as the difference in response times between incongruent and congruent trials: (CSs paired with pleasant USs/unpleasant target words + CSs paired with unpleasant USs/pleasant target words) – (CSs paired with pleasant USs/pleasant target words + CSs paired with unpleasant USs/unpleasant target words). Neutral CSs were not included in the analyses. Priming scores were calculated separately for pre and post-training. Positive priming scores suggest an assimilation effect, while negative scores suggest a contrast effect. EC scores were subjected to 2 (US Duration: fixed vs. variable; between-participants) × 2 (US-CS Overlap: no overlap vs. overlap; between-participants) × 2 (Time: pre-training vs post-training; within-participant) mixed ANOVAs, and priming scores were subjected to 2 (US Duration: fixed vs. variable; between-participants) × 2 (US/CS Overlap: no overlap vs. overlap; between-participants) × 2 (Time: pre-training vs post-training; within-participant) mixed ANOVAs. Significant interactions were followed-up with pairwise

comparisons and one sample *t*-tests where appropriate. Pillai's trace values of the multivariate solution are reported for main effects and interactions (α criteria = .05).

5.7.2 Results

Explicit valence ratings. Mean EC scores are depicted in Figure 1. The ANOVA revealed a significant three-way interaction between US variability, US/CS overlap, and time, F (1, 82) = 6.27, p = .014, η_p^2 = .071. Pairwise comparisons revealed larger responses in the fixed overlap group than the variable overlap group at post-training, t (84) = 3.14, p = .002, d = 0.68, but not pre-training, t (84) = 0.56, p = .577, d = 0.12, and larger responses in the fixed overlap group at post-training when compared to pre-training, t (85) = 2.89, p = .005, d = 0.31. One sample *t*-tests did not differ significantly from zero, all ts < 1.92, ps > .067, ds < 0.40. No differences were found between pre and post-training in each of the no overlap groups or between these groups at pre or post-training, all ts < 1.20, ps > .233, ds < 0.24.

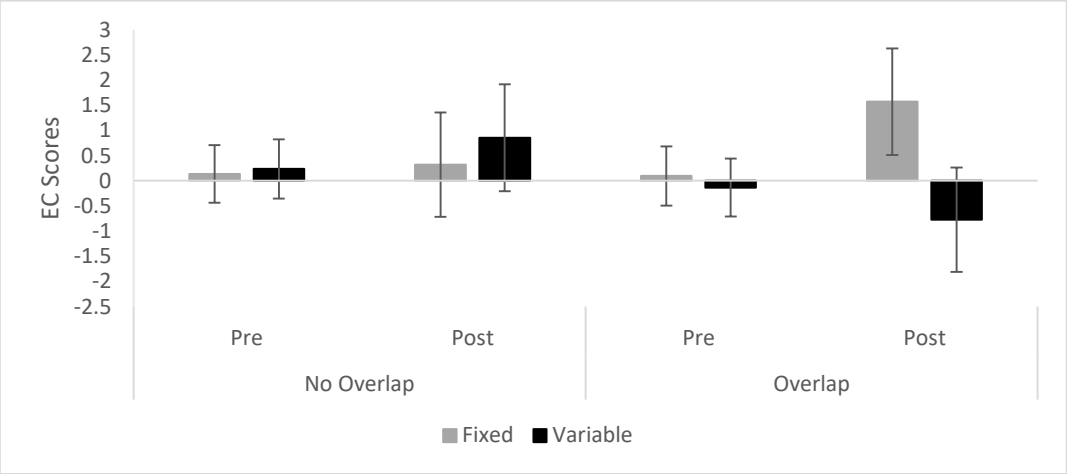

Figure 1. EC scores for backward conditioning measured at pre-training (pre) and post-training (post), as a function of US-CS overlap ('no overlap' and 'overlap') and US duration ('fixed' and 'variable') in Experiment 1. Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.

Affective priming. Mean priming scores from the two affective priming tasks completed at pre-training and post-training are depicted in Figure 2. No main effects or interactions attained significance, all Fs < 1.41, ps > 0.238, η_p^2s < .02.

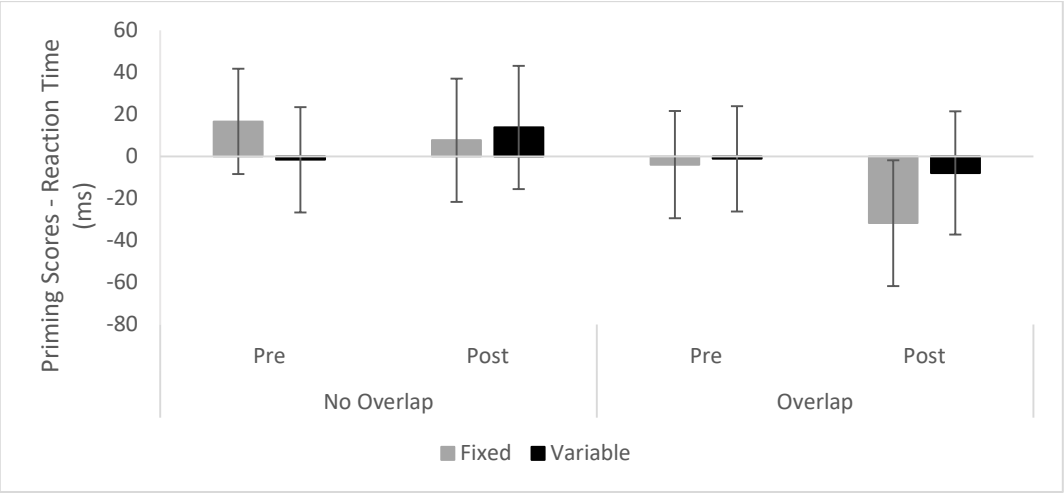

Figure 2. Priming scores for backward conditioning measured at pre-training (pre) and post-training (post) as a function of US-CS overlap ('no overlap' and 'overlap') and US duration ('fixed' and 'variable') in Experiment 1. Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.

5.7.3 Summary

Explicit valence ratings revealed some support for backward CS contrast effects emerging in the variable overlap group only, shown by a significant difference between the fixed and variable overlap groups at post-training. However, EC scores in the variable overlap group at post-training were not significantly different from zero or from pre-training. Based on the fact that this study was underpowered, it seems plausible that the backward CS EC scores may have become significantly less than zero with more participants. If so, the backward CS contrast effects observed by Moran and Bar-Anan (2013) may have been possible without the need for an instructional manipulation.

5.7.4 References

Bading, K., Stahl, C., & Rothermund, K. (2019). Why a standard IAT effect cannot provide evidence for association formation: The role of similarity construction. *Cognition and Emotion.* doi: 10.1080/02699931.2019.1604322

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

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,* 161-189. doi: 10.1177/1088868318763261

Hu, X., Gawronski, B., & Balas, R. (2017a). 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,* 17-32. doi: 10.1177/0146167216673351

Hu, X., Gawronski, B., & Balas, R. (2017b). Propositional versus dual-process accounts of evaluative conditioning: II. The effectiveness of counter-conditioning and counter-instructions in changing implicit and explicit evaluations. *Social Psychological and Personality Science, 8,* 858-866.

Lang, P.J., Bradley, M.M., & Cuthbert, B.N. (2008). International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8. University of Florida, Gainesville, FL.

Moran, T., & Bar-Anan, Y. (2013). The effect of object-valence relations on automatic evaluation. *Cognition and Emotion, 27,* 743-752. doi: 10.1080/02699931.2012.732040

Pleyers, G., Corneille, Luminet, O., & Yzerbyt, V. (2007). Aware and (dis)liking: Item-based analyses reveal that valence acquisition via evaluative conditioning emerges only when there is contingency awareness. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33, 130-144. doi: 10.1037/0278-7393.33.1.130

5.8 Supplementary Material: Recollective Memory Test

Participants were shown the two CSs in Experiment 1 and four CSs in Experiment 2 and asked:

"Circle the appropriate answer below. Was this picture presented: Together with pleasant pictures, together with unpleasant pictures, together with pleasant and unpleasant pictures, I did not see this picture, I could not tell?"

Using the sum of correct responses on the memory test, accuracy scores on the test could range from zero to two in Experiment 1 and from zero to four in Experiment 2. Participants were also presented with each US and asked to indicate which CS came after each US in both experiments and before each US in Experiment 2. This procedure resulted in an accuracy score ranging from 0 to 8 in Experiment 1 and 0 to 16 in Experiment 2. Participants were classified as remembering the CS-US contingencies if they scored 100% on both memory tests.

5.8.1 Experiment 1

One participant from the fixed-overlap and one participant from the variable-no overlap groups failed to complete the recollective memory test. Out of the participants who completed the test, the following number of participants failed in each group: fixed-no overlap = 17, fixed-overlap = 14, variable-no overlap = 22, and variable-overlap = 18.

5.8.2 Experiment 2

One participant from the fixed-overlap and one participant from the variable-overlap groups failed to complete the recollective memory. One participant from the fixed-no overlap and one participant from the fixed-overlap groups received only the first section of the recollective memory test. The participant from the fixed-no overlap group did not pass the memory test, however, the participant from the fixed-overlap group scored 100% on aspects of the test they completed. For our purposes, we will consider this to be a pass. Out of the participants who completed the test, the following number of participants failed in each group: fixed-no overlap = 57, fixed-overlap = 30, variable-no overlap = 56, and variable-overlap = 35. We believe these numbers are high because of the number of relations participants needed to recall (16). However, based on the fact that we achieved conditioning in all groups, we do not believe this measure to be assessing contingency awareness.

Chapter 6: Discussion

Evaluation of people, stimuli, and events on the dimension of valence, i.e. how much we like or dislike something, has been shown to influence a wide range of behaviours including decision making, attitude formation, and susceptibility to relapse after successful anxiety treatment (Galdi et al., 2008; Gibson, 2008; LeBel & Campbell, 2009; Matsuda et al., 2020; Sheeran et al., 2016; Zbozinek, et al., 2015). Understanding how these evaluations are acquired and manipulated is an important question that can be investigated in laboratory studies by using evaluative conditioning procedures. In these procedures, neutral CSs are presented with valenced USs and changes in CS evaluations are measured. These changes can be measured by asking participants to rate how much they like the CSs, i.e. explicit valence ratings, or through implicit measures of CS valence, such as reaction time based measures like the affective priming task, or physiological measures, such as the startle blink reflex. Previous research assessing evaluative conditioning has revealed that presenting CSs after USs in backward conditioning procedures (US-CS or CS-US-CS) yielded either assimilation effects, where the backward CS acquired the valence of the US, or contrast effects, where the backward CS acquired valence opposite to that of the US. The current thesis examined under which conditions backward CS contrast effects would emerge in both backward conditioning only (US-CS) and concurrent forward and backward conditioning designs (CS-US-CS). As previous research suggested that instructional manipulations, affective relief, US onset predictability, US intensity, and US offset predictability may all exert some influence on backward CS contrast effects, the aim of this thesis was to determine under which conditions, and to what extent, each of these factors impact backward evaluative conditioning. Eight experiments were performed across four studies to address these aims.

6.1 Summary of Results

Instructions that highlight CS-US relations in picture-picture paradigms employing CS-US-CS designs have previously been offered to explain backward CS contrast effects. Chapter 2 examined whether specific procedural features of these designs may have been conducive to eliciting affective relief at US offset, thus resulting in backward CS contrast effects without the need for the instructions that highlighted CS-US relations. The first pilot study found no backward CS contrast

effects with relational or non-relational instructions. The second pilot study replicated earlier backward CS contrast effects using the exact paradigm and relational instructions employed by Moran et al. (2016). The main experiment used a 2 × 2 design to compare the Moran et al. (2016) paradigm and relational instructions with the Mallan et al. (2008) paradigm and non-relational instructions. Backward CS contrast effects were found on explicit valence ratings in both paradigms if relational instructions were employed, while assimilation effects were found in both paradigms if non-relational instructions were employed. Forward CS ratings revealed assimilation effects in all groups, with a larger effect in groups that received the relational instructions. Affective priming suggested assimilation effects for both forward and backward CSs in the relational instruction groups and, marginally, in the non-relational instruction groups. Contrary to predictions, backward CS contrast effects only emerged when relational instructions were presented. Procedural features of the different paradigms, such as using single or multiple CSs and USs, or using different presentation times for CSs and USs, had no influence on backward conditioning. Moreover, backward CS contrast effects due to affective relief were not elicited in the non-relational instruction groups. These experiments suggest that instructional manipulations are required for backward CS contrast effects to occur in picture-picture paradigms.

Chapter 3, Experiment 1 examined whether US intensity could explain differing patterns of backward conditioning results on explicit and implicit measures of valence in a concurrent forward and backward conditioning design (CS-US-CS). Picture CSs and sound USs were used and explicit valence ratings, startle modulation, and affective priming were recorded. Instructional manipulations highlighting CS-US relations were also presented prior to acquisition. Forward CSs were rated in line with US valence, whereas backward CS contrast effects were observed with high intensity USs only. US intensity had no impact on startle modulation as assimilation effects were found for forward CSs and contrast effects for backward CSs, regardless of group. Assimilation effects were observed on affective priming for forward conditioning only, with no influence from US intensity.

Chapter 3, Experiment 2 assessed whether the offset of a pleasant US would elicit an opponent emotional reaction that mirrored that of affective relief at the offset of an aversive US, i.e. disappointment. This was achieved by presenting a backward conditioning only (US-CS) design with the addition of a neutral US-CS

pairing in order to observe the individual effects of pleasant and unpleasant US offset on backward CS contrast effects. Instructions similar to Experiment 1 were presented. Confirming prior predictions, startle modulation revealed backward CS contrast effects with larger startle responding during backward CSs following pleasant USs in comparison to neutral and unpleasant USs. However, contrary to predictions, assimilation effects were observed on explicit valence ratings despite the instructional manipulation highlighting US-CS relations. No affective priming was observed.

These experiments showed that manipulations of US intensity could not explain differing patterns of backward conditioning results on explicit and implicit measures of valence. Startle modulation results suggested that opponent emotional responses occurred at the offset of both pleasant and unpleasant USs, regardless of US intensity or US onset predictability (CS-US-CS vs US-CS). Finally, it appears that the efficacy of instructional manipulations in eliciting backward CS contrast effects on self-report measures can be reduced by lowering US intensity and can be completely eliminated by making US onset unpredictable (US-CS instead of CS-US-CS).

Chapter 4 examined whether previous startle modulation results when using sound USs could be explained by relational instructions that highlighted the role of the CSs in starting or stopping the USs. It also examined whether relational instructions were required to observe backward CS contrast effects on explicit valence ratings when sound USs were used, as Andreatta et al. (2013) found backward CS contrast effects in a CS-US-CS design with no instructional manipulation when using an electro-tactile US. A concurrent forward and backward conditioning (CS-US-CS) design with pleasant, neutral, and unpleasant USs was presented to participants with instructions highlighting the role of the CSs in starting or stopping the USs (start-stop instructions), or instructions telling participants to pay attention to the stimuli (observe instructions). Startle modulation was unaffected by the instructions as both groups showed backward CS contrast effects, and forward conditioning was not significant. For explicit valence ratings, backward CS contrast effects were observed in the start-stop instructions group only, and no backward conditioning was found in the observe instructions group. Both groups showed assimilation effects for forward conditioning. As demonstrated in previous picture-picture paradigms, the instructions presented before the conditioning task dictate the

pattern of explicit valence ratings. Startle modulation during backward CSs, on the other hand, appears to be driven by the relief and disappointment experienced at US offset, with instructional manipulations having no influence on startle responding.

Chapter 5 examined whether increasing US offset predictability alone would elicit backward CS contrast effects without presenting instructions highlighting CS-US/US-CS relations, in a picture-picture paradigm. US offset predictability was manipulated by varying US duration and by overlapping backward CS onset with US presentations (US/CS overlap). A backward conditioning only design (US-CS) was used in Experiment 1 and explicit valence ratings and affective priming showed backward CS assimilation effects regardless of US duration or US/CS overlap. Experiment 2 used a concurrent forward and backward conditioning design (CS-US-CS) and assimilation effects were found on explicit valence ratings for forward and backward conditioning in all groups, though backward CS assimilation effects were significantly larger when CSs overlapped with USs. Assimilation effects on affective priming were observed for forward conditioning only. The influence of US onset predictability was also examined by comparing backward conditioning between experiments (Experiment 1: US-CS vs Experiment 2: CS-US-CS). Explicit valence ratings showed larger assimilation effects in Experiment 1 (US-CS) compared to Experiment 2 (CS-US-CS). Both experiments showed that increasing US offset predictability does not elicit backward CS contrast effects. In fact, increasing US offset predictability by overlapping CSs and USs led to larger assimilation effects. Finally, presenting concurrent forward conditioning (CS-US-CS) to make the USs predictable resulted in smaller backward CS assimilation effects than presenting backward conditioning alone (US-CS).

The following paragraphs will address each of the factors that were assessed as potential moderators in the acquisition of backward CS contrast effects, based on explicit valence ratings and startle modulation results. While affective priming was also measured and reported, issues with low reliability (both within-thesis and in previous studies), high error rates, and inconsistent findings across studies makes interpretation difficult (Gawronski & De Houwer, 2014). Affective priming showed assimilation effects, no priming effects, or nonsensical results such as large differences at pre-test, which do not add to the current discussion due to issues with implicit measures in general. Recent studies have shown that implicit measures can be designed to show assimilation or contrast effects depending on whether

instructions are presented on every trial and how the measures are internally structured (Bading et al., 2019; Hu, Gawronski, & Balas, 2017). Moreover, the validity of implicit measures, i.e. what implicit measures are actually measuring, has been drawn into question, which also questions their usefulness with regard to their implications for EC theory (Brownstein et al., 2019; Corneille & Hutter, 2020; Corneille & Stahl, 2019; Herring et al., 2013). Thus, for these reasons and the fact that the implications that stem from the affective priming results provide little insight into backward CS valence acquisition, the affective priming results will not be discussed further.

6.2 CS and US Presentation

Results from Chapters 2-5 suggest that paradigmatic differences such as using single or multiple CSs, single or multiple USs, and varying presentation parameters of CSs and USs, had no impact on backward evaluative conditioning. The fact that backward evaluative conditioning did not differ when presenting one CS per US valence compared with presenting multiple CSs belonging to the same group per US valence, suggests that neutral stimuli sharing similar features (i.e. colour and head shape) that are individually paired with valenced USs may become functionally equivalent to a single CS. No differences in EC between single or multiple USs may also reflect that multiple USs of the same valence become functionally equivalent, as USs of the same valence serve the same purpose regarding valence change. However, as according to Hofmann et al. (2010) there is little evidence to suggest that EC depends on the number of CS-US pairings, it is also plausible that each individual CS acquired the same amount of EC as when multiple CSs were used, and that each individual US had the same potential to support EC as when multiple USs were used. Previous research also concluded that using different CS and US timings had no impact on EC, which was corroborated by the current thesis as varying CS/US timings and even flashing USs to highlight differences in US presentation durations had no influence on EC. The current thesis extends previous research by showing that the factors mentioned above, which have been previously shown to have no influence on forward evaluative conditioning, also have no influence on backward evaluative conditioning.

6.3 Relational Instructions

Results from the CS-US-CS picture-picture paradigm from Chapter 2 confirmed previous research by showing that backward CS contrast effects occurred

only when relational instructions were presented, regardless of paradigmatic features believed to increase or decrease the likelihood of backward CS contrast effects (Moran & Bar-Anan, 2013; Moran et al., 2016). This conclusion was supported by Chapter 5, as backward CS assimilation effects were found in both US-CS and CS-US-CS picture-picture paradigms when manipulating procedural factors hypothesised to lead to backward CS contrast effects in the absence of relational instructions. Results from Chapter 4 again confirmed that backward CS contrast effects emerged for explicit valence ratings only when relational instructions were presented, this time using picture CSs and sound USs in a CS-US-CS paradigm. In this study, startle modulation was unaffected by relational instructions as backward CS contrast effects were found regardless of instructions. The same pattern of results was found in Experiment 1 from Chapter 3 in a similar CS-US-CS picture-sound paradigm using relational instructions, as backward CS contrast effects emerged on explicit valence ratings and startle modulation. The only study that did not find backward CS contrast effects on explicit valence ratings when presenting relational instructions was the picture-sound backward conditioning only (US-CS) experiment from Chapter 3 (Experiment 2). In this study, a significant assimilation effect emerged on explicit valence ratings, while startle modulation revealed backward CS contrast effects.

Earlier work using relational instructions found contrast effects when using forward conditioning procedures (CS-US; Förderer & Unkelbach, 2012), and concurrent forward and backward conditioning procedures (CS-US-CS; Moran & Bar-Anan, 2013; Moran et al., 2016). However, no studies have attempted to demonstrate the effects of relational instructions in backward conditioning only designs (US-CS). The current thesis provided the first evidence showing that relational instructions did not lead to backward CS contrast effects on explicit valence ratings in a backward conditioning only design (US-CS), suggesting a potential boundary condition for the effects of relational instructions. These findings mostly support a propositional framework in which EC is determined by the relationship believed to exist between stimuli. The assimilation effect in the backward conditioning only design (US-CS) could be considered as evidence of an associative process, or it could be that the instructions were ineffective because there was no forward CS to contrast the role of the backward CS against. Hence, the meaning of the instructions changed based on the elements presented during

acquisition, which in turn changed the perception of the relationships between stimuli.

Startle modulation appears unaffected by relational instructions under all circumstances employed in this thesis. Startle modulation results could also be considered as support for an associative process as the opponent emotional reaction at US offset was elicited reliably regardless of instructional manipulations. However, a more plausible explanation could be that startle modulation reflects the emotion experienced at US offset (relief and disappointment) and explicit valence ratings assesses the relationship between USs and backward CSs based on the entire acquisition period, including the relational instructions. While it appears that relational instructions explicitly change the meaning of the backward CS in concurrent forward and backward conditioning designs (CS-US-CS), but not backward conditioning only designs (US-CS), future research should present relational instructions in backward conditioning only designs (US-CS) using picture-picture, picture-sound, and picture-shock paradigms, and compare these to CS-US-CS designs, to confirm the present results. Additionally, the use of multinomial modelling approaches may also provide further insight into the EC effects that occur within these paradigms (Heycke & Gawronski, 2020; Kukken, Hütter, & Holland, 2020).

6.4 Affective Relief

Affective relief occurs at the offset of an unpleasant or painful event, such as a shock or sound US (Andreatta et al. 2010; 2013; 2017). Stimuli that are presented during this period, such as backward CSs, can become conditioned to elicit relief (relief learning; Gerber et al., 2014). Relief learning is generally shown by startle inhibition during backward CSs that follow the aversive US. However, in a picture-picture backward conditioning only study (US-CS) with both pleasant and unpleasant USs, Mallan et al. (2008) found startle facilitation during both backward CSs, suggesting that arousal to both valenced stimuli was responsible for startle modulation. Chapter 4 and Experiment 1 from Chapter 3 investigated whether backward CS contrast effects indicative of affective relief similar to Andreatta et al. (2010; 2013) would emerge with sound USs and found that the pattern of startle responding supported an explanation based on affective relief. Experiment 2 from Chapter 3 confirmed that an opponent emotional response occurs at the offset of an unpleasant US (relief) and at the offset of a pleasant US (disappointment). Thus, the

backward CS contrast effects from Chapter 4 and Experiment 1 Chapter 3 were driven by affective responses to US offset that mirrored each other (relief and disappointment), as opposed to startle facilitation during backward CSs paired with the unpleasant US only driving backward CS contrast effects. These findings suggest that the pattern of startle modulation observed by Mallan et al. (2008) may have been due to lower levels of US intensity and/or US modality effects that were not capable of eliciting relief and disappointment at US offset.

Chapter 2 assessed whether affective relief could explain previous backward CS contrast effects in CS-US-CS picture-picture paradigms without the need for relational instructions (Moran & Bar-Anan, 2013; Moran et al., 2016). However, backward CS contrast effects were observed only when relational instructions were presented, suggesting that affective relief had no impact on backward CS contrast effects in past studies employing picture-picture paradigms. Chapter 5 also supported this conclusion in another picture-picture paradigm with both CS-US-CS and US-CS designs, as no backward CS contrast effects were observed. These findings are in line with other backward conditioning only (US-CS) studies that showed assimilation effects when no relational instructions were present (Hofmann et al., 2010; Kim, Sweldens, & Hütter, 2016; Mallan et al., 2008). However, as Chapters 3 and 4 and Andreatta et al. (2010) found assimilation effects on explicit valence ratings without relational instructions while startle modulation was suggestive of affective relief, it is possible that Chapters 2 and 5 may have shown affective relief on startle if it was measured. While plausible, this is unlikely because affective relief has not been shown on startle responding when using picture USs in backward conditioning (Mallan et al., 2008). Future research should investigate whether picture USs that are higher in arousal and further apart on valence ratings are capable of eliciting startle modulation patterns indicative of relief and disappointment learning.

The reliable finding of startle facilitation during backward CSs presented after pleasant USs in Chapters 3 and 4 suggests that the present paradigm using pleasant sound stimuli may provide an avenue to further investigate appetitive conditioning. While forward conditioning results were inconclusive, the reliable elicitation of negative emotion at the offset of the pleasant USs suggests that pleasant sound USs did in fact elicit pleasant emotion. Thus, simplifying the current paradigm may provide a fruitful approach for learning more about appetitive conditioning. Future research should compare CSs paired with shorter versions of the pleasant and

neutral sound USs in a forward conditioning only design (CS-US). If reliable differences between CSs emerge on both self-report and physiological measures of CS valence, then a reliable paradigm for assessing appetitive conditioning may have been found.

6.5 US/CS Overlap and US Onset Predictability

In Chapter 5, the influence of overlap between USs and backward CSs (US/CS overlap) on backward CS contrast effects was investigated. The hypothesis was that a backward CS predicting the offset of an aversive US may acquire positive properties from signalling safety from the aversive US, and that a backward CS predicting the offset of an appetitive US may acquire negative properties from signalling the end of a pleasant US. Instead, results showed that overlapping backward CSs with USs resulted in larger assimilation effects than if they did not overlap, potentially due to an additive effect of having a brief moment of simultaneous conditioning in addition to backward conditioning. The assimilation effects were also larger in backward conditioning only (US-CS) than concurrent forward and backward conditioning designs (CS-US-CS), implicating US onset predictability as a moderating factor in backward CS valence acquisition. Previous studies that used relational instructions that also overlapped backward CSs and USs to further highlight that backward CSs controlled USs may have created a counter-productive situation for finding backward CS contrast effects because overlapping backward CSs and USs increases assimilation effects, rather than increasing the likelihood of the instructions eliciting backward CS contrast effects (Moran & Bar-Anan, 2013; Moran et al., 2016). Alternatively, the relational instructions may have changed the interpretation of the US/CS procedure to one where the backward CS did in fact signal US offset, thus facilitating backward CS contrast effects. While both alternatives are plausible, Chapter 2 showed no differences between paradigms that differed on US/CS overlap, suggesting that in CS-US-CS paradigms the influence of US/CS overlap when relational instructions are present is negligible. In contrast, results from Experiment 2, Chapter 3, suggest that US/CS overlap in US-CS only paradigms has a large effect when relational instructions are present, as explicit valence ratings revealed assimilation effects, despite the fact that the relational instructions should have elicited backward CS contrast effects. US/CS overlap may have resulted in the US and backward CS appearing as related objects that share common features, as opposed to the backward CS signalling US offset. The end

result thus being a large assimilation effect that the relational instructions were not strong enough to overpower.

Putting aside US/CS overlap, whether the US can be predicted by a concurrent forward CS (CS-US-CS), or not (US-CS), appears to influence the nature of backward evaluative conditioning. Chapter 5 showed larger backward conditioning in the US-CS experiment than the CS-US-CS experiment. Moreover, Chapter 2, Chapter 4, and Experiment 1 from Chapter 3, all of which employed relational instructions in CS-US-CS designs, showed backward CS contrast effects, while Experiment 2 from Chapter 3 that employed relational instructions in a US-CS design, found backward CS assimilation effects. One plausible explanation for this is that the smaller assimilation effect that comes from the CS-US-CS design is easily overpowered by the relational instructions, whereas the stronger assimilation effect from the US-CS design is not. Another explanation that may work alone or in conjunction with the first, is that the meaning of each stimulus and their relation to each other changes depending on stimulus arrangement. In the US-CS design, the relational instructions may have less meaning because there is no 'good/bad stimulus' that starts the pleasant or aversive event to contrast the backward CS against, whereas in the CS-US-CS design, the backward CSs' role in stopping the pleasant or aversive event is more salient when contrasted with the forward CSs' role in starting the USs. The findings mentioned above also confirm previous findings, as Andreatta et al. (2013) demonstrated backward CS contrast effects on explicit valence ratings without relational instructions in a CS-US-CS design, but assimilation effects in a US-CS design (Andreatta et al., 2010; Andreatta et al., 2017). Thus, it appears that having a US that is predicted by a forward CS may be more conducive to backward CS contrast effects, either resulting from relational instructions, or from a painful US.

The idea that changing stimulus arrangements (CS-US-CS vs US-CS) interacts with relational instructions and US properties to change the meaning of the backward CS is supported by the fact that startle modulation indicates backward CS contrast effects, regardless of these factors. As startle modulation occurs reflexively at the time of the probe and does not differ as a function of US onset predictability or instructions, this suggests that any differences on explicit valence ratings may occur as a result of reasoning about other factors that could influence the meaning of the backward CS. These factors include other stimuli presented in the acquisition trial or

relational instructions presented before acquisition, both of which may moderate the perception of the events occurring within the acquisition trial. If the proposal that the meaning of the backward CS changes depending on features of the acquisition phase is correct, then these results support a propositional framework of EC in which stimulus valence changes depending on how participants make sense of the acquisition phase. Future research should directly test the interaction between relational instructions and conditioning design (CS-US-CS vs US-CS) when USs and CSs overlap to confirm the pattern of results and explanations presented above. Moreover, doing so would help to explain how each factor changes the perception of stimulus relationships during acquisition in different circumstances.

6.6 US Intensity

Chapter 3 Experiment 1 used sound USs and relational instructions in a concurrent forward and backward conditioning design (CS-US-CS) to show that US intensity could not account for previous discrepancies on different implicit measures of CS valence. Research has since shown that differences between assimilation effects on implicit reaction time based measures (e.g. Moran & Bar-Anan, 2013) and contrast effects on startle modulation (e.g. Andreatta et al., 2010; 2013) were the result of the task structure used in the reaction time based measures, as opposed to differences that may have been driven by US intensity (Bading et al., 2019). US intensity was also unable to explain startle modulation differences between shock US paradigms that found backward CS contrast effects and a picture-picture paradigm that found startle facilitation to CSs paired with both pleasant and unpleasant USs (Andreatta et al., 2010; 2013; Mallan et al., 2008). The lack of US intensity effects may have been due to the intensity manipulation employed in Chapter 3 being insufficient to match the intensity of the different modality USs in previous research. An alternative is that US modality may actually play a currently undefined role in backward CS acquisition, or that US/CS modality match influences encoding resulting in different backward conditioning effects. While Hofmann et al. (2010) found that US/CS modality did not moderate EC, the acoustic modality was not included in the analysis. Moreover, Gast, Langer, and Sengewald, (2016) found larger EC effects during simultaneous conditioning when sound USs and visual CSs were presented that did not occur when both USs and CSs were visual. Future research should assess the effects of US/CS modality match with auditory and tactile

stimuli to determine whether modality match can explain differing patterns of startle modulation results in backward conditioning designs (CS-US-CS and US-CS).

Explicit valence ratings in Chapter 3 Experiment 1 initially appeared sensitive to US intensity, as backward CS contrast effects were observed in the high intensity group, with no observable backward conditioning in the low intensity group. However, the same low intensity USs elicited backward CS contrast effects in a similar paradigm in Chapter 4, and Chapter 2 showed backward CS contrast effects with picture USs that were arguably lower in intensity than the low intensity sound USs. Taken together, these studies suggest that the influence of US intensity on backward CS valence ratings was negligible. This conclusion was mostly supported by previous research, as backward CS contrast effects were found in picture-picture and picture-sound paradigms using relational instructions (Moran & Bar-Anan, 2013; Moran et al. 2016). An exception to this was Andreatta et al. (2013), who found backward CS contrast effects without relational instructions when using shock USs. It is plausible that US modality may be responsible, as EC appears larger when using shock USs than USs from other modalities (Hofmann et al., 2010). Although, in this case, a larger assimilation effect would be expected. Thus, it is more likely that something other than US modality influences backward CS contrast effects in shock paradigms.

A plausible factor that may influence backward CS contrast effects in shock paradigms is pain intensity. The shock USs that Andreatta et al. (2013; 2017) employed were reported to be painful, and higher pain intensity has been shown to lead to greater levels of pain relief (Leknes, Brooks, Wiech, & Tracey, 2008). As no other concurrent forward and backward conditioning (CS-US-CS) studies have employed painful USs, is it plausible that only painful USs elicit relief learning observable on self-report measures. While relief learning has been shown on startle modulation without painful USs (Chapters 3 and 4), this may be due to startle assessing the emotional state at the time of the probe, which does not carry over to explicit valence ratings when the USs are not painful. A review by Biggs et al. (2020) concluded that the neural networks underlying conditioning with painful and non-painful stimuli utilise similar but distinguishable networks, supporting the notion that relief learning may only be observable on self-report measures when using painful USs. Another alternative is that shock USs are different from picture and sound USs as they involve physical contact with an aversive stimulus. They are also

more difficult to avoid as participants would have to unwrap the bandage and remove the electrode, whereas closing the eyes or taking off the headphones are easy in comparison. Future research should investigate whether modality effects are present between shock USs and other types of non-tactile stimuli, between painful and non-painful shocks, and between easy and hard to avoid stimuli, in backward evaluative conditioning. Moreover, future research should assess whether there is a threshold level of pain required for backward CS relief effects to occur on explicit valence ratings.

6.7 Dissociation between Startle Modulation and Valence Ratings

Startle modulation results from the current thesis reliably show relief and disappointment learning as backward CSs following valenced USs always acquired valence opposite to the US they were presented after. This occurred regardless of relational instructions, US intensity, and whether US onset was predictable. Results from explicit valence ratings show that backward CS contrast effects occur only when relational instructions are present. Moreover, US onset predictability, US/CS overlap, and US intensity all influence backward CS acquisition. The end result is a dissociation between startle modulation and explicit valence ratings when relational instructions are not presented (Chapter 4), or during backward conditioning only designs (US-CS; Chapter 3, Experiment 2). It is plausible that this dissociation is the result of dual-process learning where startle modulation represents the association and explicit valence ratings represents propositional inferences about backward CS valence. However, a more parsimonious alternative is that startle modulation and explicit valence ratings are measuring different things, rather than showing a true dissociation.

According to Solomon's (1980) opponent-process theory, a US elicits an emotional reaction, the *a*-process, which arouses an opposing reaction, the *b*-process, which serves to suppress the initial reaction. The summation of these two processes determines the state of the organism. The fact that startle modulation during CSs following valenced USs reliably shows relief and disappointment suggests that startle modulation reflects the emotional state at the time of the startle probe, i.e. the *b*-process. Explicit valence ratings do not reliably conform to the predictions from opponent-process theory (except for Experiment 2 from Andreatta et al., 2013), which could suggest that they not only measure emotional state, but assess other aspects of the acquisition procedure. Measuring startle and valence ratings at

different time points does not invalidate this claim as Luck and Lipp (2017) found the same dissociation between startle and valence ratings when ratings were measured online during acquisition. It seems more plausible that explicit valence ratings reflect the entire acquisition episode, taking into account emotional state, as well as propositions about the meaning of all the relevant elements that may influence backward CS valence, such as relational instructions, US onset predictability, US/CS overlap, and US intensity, among others. This would explain why startle modulation can suggest valence in one direction (i.e. positive) while explicit valence ratings suggest the other (i.e. negative).

Support for the distinction between emotional state and explicit stimulus valence was shown by Weber, Shinkareva, Kim, Gao, and Wedell (2020) who found that EC resulted from the affective valence elicited by the US (i.e. emotional state) rather than the liking of the US (i.e. explicit stimulus valence). Thus, it seems plausible that relational instructions may overpower the EC effect driven by affective valence elicited by the US (emotional state), which reflects propositions about the entire learning episode, while startle modulation still reflects the emotional state. Future research should further examine this possibility by measuring startle responding and explicit valence ratings when using USs that elicit emotion in one direction and liking in another direction, such as using a song that is liked but elicits negative emotion (i.e. a sad song that you love). Moreover, future research should be explicit about whether affective valence (i.e. emotion) or stimulus valence is being assessed in order to delineate boundary conditions between the two. This may help to explain different patterns of results across different measures and paradigms. In addition, understanding more about these different types of outcomes from acquisition procedures (emotion vs liking) has implications not only for EC and related theories, but also for other learning mechanisms based on Pavlovian conditioning principles in general.

6.8 Theoretical Implications

Current theorising about EC revolves around the debate between a single process propositional account and dual-process accounts such as the associative-propositional evaluative (APE) model (Gawronski & Bodenhausen, 2006, 2011, 2018; Mitchell et al., 2009). Single process accounts suggest that a single learning mechanism is responsible for EC, which is based on evaluating the truth value of the relationship between stimuli. Dual-process accounts, such as the APE model, suggest

that an additional associative learning mechanism also influences EC, resulting from a link-based representation of stimulus relations without assessing the truth value of the relationship. Findings from the current thesis could be interpreted as supporting both single process propositional accounts and dual-process accounts of EC. Support for a propositional mechanism is shown by relational instructions eliciting backward CS contrast effects on explicit valence ratings in the majority of cases. Backward CS assimilation effects when using relational instructions were found only in Experiment 2, Chapter 3, which could be considered an associative effect. However, as this only occurred when USs were unpredictable (US-CS instead of CS-US-CS), it seems more plausible that removing the forward CS changed the meaning of the acquisition phase and the effectiveness of the relational instructions, which then changed the relevant propositions drawn from the experience. In this instance, results from explicit valence ratings can be reconciled within a propositional account of EC.

Relational instructions and US onset predictability appeared to have no influence on startle modulation. Moreover, startle modulation consistently showed backward CS contrast effects where the backward CS acquired the opponent emotional response elicited at US offset. This could be taken as support for an associative mechanism of valence acquisition. As explicit valence ratings are influenced by a propositional mechanism, this would suggest that a dual-process account would best explain the current findings. However, as explained previously, it seems more likely that startle modulation is measuring the emotional state at the time of the probe, while explicit valence ratings are measuring both the emotional state and any additional information that may influence the perception of backward CS valence (i.e. relational instructions, US onset predictability etc.). If so, then a dual-process account of EC would be unnecessary to explain the current findings. In any case, strong conclusions about the theoretical implications of the current thesis should be withheld until the dissociation between explicit valence ratings and startle modulation can be clarified.

6.9 Conclusion

The current thesis examined the effects of relational instructions, affective relief, US intensity, US onset predictability, and US offset predictability on the emergence of backward CS contrast effects in evaluative conditioning across four studies containing eight experiments. Minor paradigmatic changes surrounding CS-US timings and presentations that also influenced US offset predictability were found

to be inconsequential. Startle modulation reliably demonstrated relief and disappointment learning, regardless of any instructional or US related manipulations (i.e. US onset/offset predictability and US intensity). The successful demonstration of disappointment learning following appetitive US offset provides a potential avenue in which to further investigate appetitive conditioning in a paradigm that drives physiological responding. Relational instructions led to backward CS contrast effects on explicit valence ratings in CS-US-CS designs, while assimilation effects were found in the US-CS design, suggesting a potential boundary condition in the effectiveness of relational instructions. The predictability of US onset also affected backward CS valence, which suggests that all stimuli in an acquisition phase interact to influence the outcome of backward CS valence acquisition. The dissociations observed between startle modulation and explicit valence ratings suggest that startle modulation may reflect the emotional state during backward CS presentation, while explicit valence ratings reflect the combination of emotional state and all other information present during the acquisition phase (i.e. relational instructions and US onset predictability). The current thesis demonstrates that backward CS valence acquisition occurs through dynamic and interactive processes, and that the examination of backward conditioning is useful to highlight boundary conditions for contemporary evaluative conditioning manipulations and the measurement of CS valence.

6.10 References

Andreatta, M., Mühlberger, A., Glotzbach-Schoon, E., & Pauli, P. (2013). Pain predictability reverses valence ratings of a relief-associated stimulus. Frontiers in Systems Neuroscience, 7(53), 1-12, https://doi.org/10.3389/fnsys.2013.00053

Andreatta, M., Mühlberger, A., Yarali, A., Gerber, B., & Pauli, P. (2010). A rift between implicit and explicit conditioned valence in human pain relief learning. Proceedings of the Royal Society of London B: Biological Sciences. https://doi.org/10.1098/rspb.2010.0103

Andreatta, M., & Pauli, P. (2017). Learning mechanisms underlying threat absence and threat relief: Influences of trait anxiety. Neurobiology of Learning and Memory, 145, 105-113. https://doi.org/10.1016/j.nlm.2017.09.005

Bading, K., Stahl, C., & Rothermund, K. (2019). Why a standard IAT effect cannot provide evidence for association formation: The role of similarity construction. Cognition and Emotion. https://doi.org/10.1080/02699931.2019.1604322.

Brownstein, M., & Madva, A., & Gawronski, B. (2019). What do implicit measures measure? Wiley Interdisciplinary Reviews: Cognitive Science. https://doi.org/10.10.1002/wcs.1501

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, 1-21. https://doi/org/10.1177/1088868320911325

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

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

Galdi, S., Arcuri, L., & Gawronski, B. (2008). Automatic mental associations predict future choices of undecided decision makers. Science, 321, 1100-1102. https://doi.org/ 10.1126/science.ll60769

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. http://dx.doi.org/10.1016/j.actpsy.2016.07.002

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.

Gerber, B., Yarali, A., Diegelmann, S., Wotjak, C. T., Pauli, P., & Fendt, M. (2014). Pain- relief learning in flies, rats, and man: Basic research and applied perspectives. Learning and Memory, 21, 232-252. https://doi.org/10.1101/lm.032995.113

Gibson, B. (2008). Can evaluative conditioning change attitudes toward mature brands? New evidence from the Implicit Association Test. Journal of Consumer Research, 35, 178-188. https://doi.org/10.1086/527341

Herring, D. R., White, K. R., Jabeen, L. N., Hinojos, M., Terrazas, G., Reyes, S. M., … Crites, S. L., Jr. (2013). On the automatic activation of attitudes: A quarter century of evaluative priming research. Psychological Bulletin, 139, 1062-1089. https://doi.org/10.1037/a0031309

Heycke, T., & Gawronski, B. (2020). Co-occurrence and relational information in evaluative learning: A multinomial modelling approach. Journal of Experimental Psychology: General, 149, 104-124. https://doi.org/10.1037/xge0000620

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. Psychological Bulletin, 136, 390-421. https://doi.org/10.1037/a0018916

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, 17-32. https://doi.org/10.1177/0146167216673351

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, 61-68. https://doi.org/10.1177/1948550615599237

Kukken, N., Hütter, M., & Holland, R. W. (2020). Are there two independent evaluative conditioning effects in relational paradigms? Dissociating the effects of CS-US pairings and their meaning. *Cognition and Emotion, 34,* 170-187. https://doi.org/ 10.1080/02699931.2019.1617112

LeBel, E. P., & Campbell, L. (2009). Implicit partner affect, relationship satisfaction, and the prediction of romantic breakup. Journal of Experimental Social Psychology, 45, 1291-1294. https://doi.org/10.1016/j.jesp.2009.07.003

Leknes, S., Brooks, J. C. W., Wiech, K., & Tracey, I. (2008). Pain relief as an opponent process: A psychophysical investigation. European Journal of Neuroscience, 28, 794-810. https://doi.org/10.1111/j.1460-9568.2008.06380.x

Luck. C. C., & Lipp, O. V. (2017). Startle modulation and explicit valence evaluations dissociate during backward fear conditioning. Psychophysiology, 54, 673-683. https://doi.org/10.1111/psyp.12834

Mallan, K. M., Lipp, O. V., & Libera, M. (2008). Affect, attention, or anticipatory arousal? Human blink startle modulation in forward and backward affective conditioning. International Journal of Psychophysiology, 69. 9-17. https://doi.org/10.1016/j.ijpsycho.2008.02.005

Matsuda, K., Garcia, Y., Catagnus, R., Brandt, J. A. (2020). Can behavior analysis help us understand and reduce racism? A review of the current literature. *Behavior Analysis in Practice, 13,* 336-347. https://doi.org/10.1007/s40617-020-00411-4

Moran, T., and Bar-Anan, Y. (2013). The effect of object-valence relations on automatic evaluation. Cognition and Emotion, 27, 743-752. https://doi.org/10.1080/02699931.2012.732040

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2016). The assimilative effect of co-occurrence on evaluation above and beyond the effect of relational qualifiers. Social Cognition, 34, 435-461. https://doi.org/101521soco2016345435

Sheeran, P., Maki, A., Montanaro, E., Avishai-Yitshak, A., Bryan, A., Klein, W. M. P., … Rothman, A. J. (2016). The impact of changing attitudes, norms, and self-efficacy on health-related intentions and behaviour: A meta-analysis. *Health Psychology, 35,* 1178-1188. https://doi.org/10.1037/hea0000387

Solomon, R. L. (1980). The opponent-process theory of acquired motivation: The costs of pleasure and the benefits of pain. *American Psychologist, 35,* 691-712. https://doi.org/10.1037/0003-066X.35.8.691

Weber, C. E., Shinkareva, S. V., Kim, J., Gao, C., & Wedell, D. H. (2020). Evaluative conditioning of affective valence. Social Cognition, 38, 97-118. https://doi.org/10.1521/soco.2020.38.2.97

Zbozinek, T. D., Hermans, D., Prenoveau, J. M., Liao, B., & Craske, M. G. (2015). Post-extinction conditional stimulus valence predicts reinstatement fear: Relevance for long-term outcomes of exposure therapy. *Cognition and Emotion, 29*(4), 654-667. https://doi.org/10.1080/02699931.2014.930421

Appendix A. Copyright Permissions

JOHN WILEY AND SONS LICENSE
TERMS AND CONDITIONS

Jul 15, 2020

This Agreement between Mr. Luke Green ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	4870091070941
License date	Jul 15, 2020
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Psychophysiology
Licensed Content Title	How disappointing: Startle modulation reveals conditional stimuli presented after pleasant unconditional stimuli acquire negative valence
Licensed Content Author	Luke J. S. Green, Camilla C. Luck, Ottmar V. Lipp
Licensed Content Date	Mar 13, 2020
Licensed Content Pages	1
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Print and electronic

Portion	Full article
Will you be translating?	No
Title	An investigation of assimilation and contrast effects in backward evaluative conditioning
Institution name	Curtin University
Expected presentation date	Aug 2020
Requestor Location	Mr. Luke Green Curtin University Kent Street Bentley, WA 6102 Australia Attn: Mr. Luke Green
Publisher Tax ID	EU826007151
Total	0.00 AUD

Terms and Conditions

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a"Wiley Company") or handled on behalf of a society with which a Wiley Company has exclusive publishing rights in relation to a particular work (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your RightsLink account (these are available at any time at http://myaccount.copyright.com).

Terms and Conditions

- The materials you have requested permission to reproduce or reuse (the "Wiley Materials") are protected by copyright.

- You are hereby granted a personal, non-exclusive, non-sub licensable (on a stand-alone basis), non-transferable, worldwide, limited license to reproduce the Wiley Materials for the purpose specified in the licensing process. This license, **and any CONTENT (PDF or image file) purchased as part of your order,** is for a one-time use only and limited to any maximum distribution number specified in the license. The first instance of republication or reuse granted by this license must be completed within two years of the date of the grant of this license (although copies prepared before the end date may be distributed thereafter). The Wiley Materials shall not be used in any other manner or for any other purpose, beyond what is granted in the license. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Wiley Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Wiley Material. Any third party content is expressly excluded from this permission.

- With respect to the Wiley Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Wiley Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Wiley Materials without the prior permission of the respective copyright owner.**For STM Signatory Publishers clearing permission under the terms of the STM Permissions Guidelines only, the terms of the license are extended to include subsequent editions and for editions in other languages, provided such editions are for the work as a whole in situ and does not involve the separate exploitation of the permitted figures or extracts,** You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Wiley Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Wiley Materials on a stand-alone basis, or any of the rights granted to you hereunder to any other person.

- The Wiley Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc, the Wiley Companies, or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Wiley Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Wiley Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto

- NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED BY YOU.

- WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.

- You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or

threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.

- IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

- Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

- The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.

- This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.

- Any fee required for this permission shall be non-refundable after thirty (30) days from receipt by the CCC.

- These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.

- In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.

- WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

- This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.

- This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

WILEY OPEN ACCESS TERMS AND CONDITIONS

Wiley Publishes Open Access Articles in fully Open Access Journals and in Subscription journals offering Online Open. Although most of the fully Open Access journals publish open access articles under the terms of the Creative Commons Attribution (CC BY) License only, the subscription journals and a few of the Open Access Journals offer a choice of Creative Commons Licenses. The license type is clearly identified on the article.

The Creative Commons Attribution License

The Creative Commons Attribution License (CC-BY) allows users to copy, distribute and transmit an article, adapt the article and make commercial use of the article. The CC-BY license permits commercial and non-

Creative Commons Attribution Non-Commercial License

The Creative Commons Attribution Non-Commercial (CC-BY-NC)License permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.(see below)

Creative Commons Attribution-Non-Commercial-NoDerivs License

The Creative Commons Attribution Non-Commercial-NoDerivs License (CC-BY-NC-ND) permits use, distribution and reproduction in any medium, provided the original work is properly cited, is not used for commercial purposes and no modifications or adaptations are made. (see below)

Use by commercial "for-profit" organizations

Use of Wiley Open Access articles for commercial, promotional, or marketing purposes requires further explicit permission from Wiley and will be subject to a fee.

Further details can be found on Wiley Online Library
http://olabout.wiley.com/WileyCDA/Section/id-410895.html

Other Terms and Conditions:

v1.10 Last updated September 2015

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Psychophysiology

Published by Wiley on behalf of Society for Psychophysiological Research (the "Owner")

COPYRIGHT TRANSFER AGREEMENT

Date: August 11, 2020

Contributor name: Luke John Stanley Green

Contributor address:

Manuscript number: PsyP-2020-0237.R2

Re: Manuscript entitled Startle during backward evaluative conditioning is not modulated by instructions (the "Contribution")

for publication in Psychophysiology (the "Journal")

published by Wiley Periodicals LLC ("Wiley")

Dear Contributor(s):

Thank you for submitting your Contribution for publication. In order to expedite the editing and publishing process and enable the Owner to disseminate your Contribution to the fullest extent, we need to have this Copyright Transfer Agreement executed. If the Contribution is not accepted for publication, or if the Contribution is subsequently rejected, this Agreement shall be null and void.
Publication cannot proceed without a signed copy of this Agreement.

A. COPYRIGHT

1. The Contributor assigns to the Owner, during the full term of copyright and any extensions or renewals, all copyright in and to the Contribution, and all rights therein, including but not limited to the right to publish, republish, transmit, sell, distribute and otherwise use the Contribution in whole or in part in electronic and print editions of the Journal and in derivative works throughout the world, in all languages and in all media of expression now known or later developed, and to license or permit others to do so. For the avoidance of doubt, "Contribution" is defined to only include the article submitted by the Contributor for publication in the Journal (including any embedded rich media) and does not extend to any supporting information submitted with or referred to in the Contribution ("Supporting Information"). To the extent that any Supporting Information is submitted to the Journal, the Owner is granted a perpetual, non-exclusive license to publish, republish, transmit, sell, distribute and otherwise use this Supporting Information in whole or in part in electronic and print editions of the Journal and in derivative works throughout the world, in all languages and in all media of expression now known or later developed, and to license or permit others to do so.

2. Reproduction, posting, transmission or other distribution or use of the final Contribution in whole or in part in any medium by the Contributor as permitted by this Agreement requires a citation to the Journal suitable in form and content as follows: (Title of Article, Contributor, Journal Title and Volume/Issue, Copyright © [year], copyright owner as specified in the Journal, Publisher). Links to the final article on the publisher website are encouraged where appropriate.

B. RETAINED RIGHTS

Notwithstanding the above, the Contributor or, if applicable, the Contributor's employer, retains all proprietary rights other than copyright, such as patent rights, in any process, procedure or article of manufacture described in the Contribution.

C. PERMITTED USES BY CONTRIBUTOR

1. Submitted Version. The Owner licenses back the following rights to the Contributor in the version of the Contribution as originally submitted for publication (the "Submitted Version"):

a. The right to self-archive the Submitted Version on: the Contributor's personal website; a not for profit subject-based preprint server or repository; a Scholarly Collaboration Network (SCN) which has signed up to the STM article sharing principles [http://www.stm-assoc.org/stm-consultations/scn-consultation-2015/] ("Compliant SCNs"); or the Contributor's company/ institutional repository or archive. This right extends to both intranets and the Internet. The Contributor may replace the Submitted Version with the Accepted Version, after any relevant embargo period as set out in paragraph C.2(a) below has elapsed. The Contributor may wish to add a note about acceptance by the Journal and upon publication it is recommended that Contributors add a Digital Object Identifier (DOI) link back to the Final Published Version.

b. The right to transmit, print and share copies of the Submitted Version with colleagues, including via Compliant SCNs, provided that there is no systematic distribution of the Submitted Version, e.g. posting on a listserve, network (including SCNs which have not signed up to the STM sharing principles) or automated delivery.

2. Accepted Version. The Owner licenses back the following rights to the Contributor in the version of the Contribution that has been peer-reviewed and accepted for publication, but not final (the "Accepted Version"):

a. The right to self-archive the Accepted Version on: the Contributor's personal website; the Contributor's company/institutional repository or archive; Compliant SCNs; and not for profit subject-based repositories such as PubMed Central, all subject to an embargo period of 12 months for scientific, technical and medical (STM) journals and 24 months for social science and humanities (SSH) journals following publication of the Final Published Version. There are separate arrangements with certain funding agencies governing reuse of the Accepted Version as set forth at the following website: http://www.wileyauthors.com/funderagreements. The Contributor may not update the Accepted Version or replace it with the Final Published Version. The Accepted Version posted must contain a legend as follows: This is the accepted version of the following article: FULL CITE, which has been published in final form at [Link to final article]. This article may be used for non-commercial purposes in accordance with the Wiley Self-Archiving Policy [http://www.wileyauthors.com/self-archiving].

b. The right to transmit, print and share copies of the Accepted Version with colleagues, including via Compliant SCNs (in private research groups only before the embargo and publicly after), provided that there is no systematic distribution of the Accepted Version, e.g. posting on a listserve, network (including SCNs which have not signed up to the STM sharing principles) or automated delivery.

3. Final Published Version. The Owner hereby licenses back to the Contributor the following rights with respect to the final published version of the Contribution (the "Final Published Version"):

a. Copies for colleagues. The personal right of the Contributor only to send or transmit individual copies of the Final Published Version in any format to colleagues upon their specific request, and to share copies in private sharing groups in Compliant SCNs, provided no fee is charged, and further provided that there is no systematic external or public distribution of the Final Published Version, e.g. posting on a listserve, network or automated delivery.

b. Re-use in other publications. The right to re-use the Final Published Version or parts thereof for any publication authored or edited by the Contributor (excluding journal articles) where such re-used material constitutes less than half of the total material in such publication. In such case, any modifications must be accurately noted.

c. Teaching duties. The right to include the Final Published Version in teaching or training duties at the Contributor's institution/place of employment including in course packs, e-reserves, presentation at professional conferences, in-house training, or distance learning. The Final Published Version may not be used in seminars outside of normal teaching obligations (e.g. commercial seminars). Electronic posting of the Final Published Version in connection with teaching/training at the Contributor's company/institution is permitted subject to the implementation of reasonable access control mechanisms, such as user name and password. Posting the Final Published Version on the open Internet is not permitted.

d. Oral presentations. The right to make oral presentations based on the Final Published Version.

4. Article Abstracts, Figures, Tables, Artwork and Selected Text (up to 250 words).

a. Contributors may re-use unmodified abstracts for any non-commercial purpose. For online uses of the abstracts, the Owner encourages but does not require linking back to the Final Published Version.

b. Contributors may re-use figures, tables, artwork, and selected text up to 250 words from their Contributions, provided the following conditions are met:

(i) Full and accurate credit must be given to the Final Published Version.

(ii) Modifications to the figures and tables must be noted. Otherwise, no changes may be made.

(iii) The re-use may not be made for direct commercial purposes, or for financial consideration to the Contributor.

(iv) Nothing herein will permit dual publication in violation of journal ethical practices.

D. CONTRIBUTIONS OWNED BY EMPLOYER

1. If the Contribution was written by the Contributor in the course of the Contributor's employment as a "work-made-for-hire" in the course of employment, the Contribution is owned by the company/institution which must execute this Agreement (in addition to the Contributor's signature). In such case, the company/institution hereby agrees to the terms of use set forth in paragraph A above and assigns to the Owner, during the full term of copyright, all copyright in and to the Contribution for the full term of copyright throughout the world as specified in paragraph A above.

2. In addition to the rights specified as retained in paragraph B above and the rights granted back to the Contributor pursuant to paragraph C above, the Owner hereby grants back, without charge, to such company/institution, its subsidiaries and divisions, the right to make copies of and distribute the Final Published Version internally in print format or electronically on the Company's internal network. Copies so used may not be resold or distributed externally. However, the company/institution may include information and text from the Final Published Version as part of an information package included with software or other products offered for sale or license or included in patent applications. Posting of the Final Published Version by the company/institution on a public access website may only be done with written permission, and payment of any applicable fee(s). Also, upon payment of the applicable reprint fee, the company/institution may distribute print copies of the Final Published Version externally.

E. GOVERNMENT CONTRACTS

In the case of a Contribution prepared under U.S. Government contract or grant, the U.S. Government may reproduce, without charge, all or portions of the Contribution and may authorize others to do so, for official U.S. Government purposes only, if the U.S. Government contract or grant so requires. (U.S. Government, U.K. Government, and other government employees: see notes at end.)

F. COPYRIGHT NOTICE

The Contributor and the company/institution agree that any and all copies of the Final Published Version or any part thereof distributed or posted by them in print or electronic format as permitted herein will include the notice of copyright as stipulated in the Journal and a full citation to the Journal.

G. CONTRIBUTOR'S REPRESENTATIONS

The Contributor represents that: (i) the Contribution is the Contributor's original work, all individuals identified as Contributors actually contributed to the Contribution, and all individuals who contributed are included; (ii) if the Contribution was prepared jointly, the Contributor has informed the co-Contributors of the terms of this Agreement and has obtained their signed written permission to execute this Agreement on their behalf; (iii) the Contribution is submitted only to this Journal and has not been published before, has not been included in another manuscript, and is not currently under consideration or accepted for publication elsewhere; (iv) if excerpts from copyrighted works owned by third parties are included, the Contributor shall obtain written permission from the copyright owners for all uses as set forth in the standard permissions form or the Journal's Author Guidelines, and show credit to the sources in the Contribution; (v) the Contribution and any submitted Supporting Information contain no libelous or unlawful statements, do not infringe upon the rights (including without limitation the copyright, patent or trademark rights) or the privacy of others, do not breach any confidentiality obligation, do not violate a contract or any law, do not contain material or instructions that might cause harm or injury, and only utilize data that has been obtained in accordance with applicable legal requirements and Journal policies; (vi) there are no conflicts of interest relating to the Contribution, except as disclosed. Accordingly, the Contributor represents that the following information shall be clearly identified on the title page of the Contribution: (1) all financial and material support for the research and work; (2) any financial interests the Contributor or any co-Contributors may have in companies or other entities that have an interest in the information in the Contribution or any submitted Supporting Information (e.g., grants, advisory boards, employment, consultancies, contracts, honoraria, royalties, expert testimony, partnerships, or stock ownership); and (3) indication of no such financial interests if appropriate.

Wiley reserves the right, notwithstanding acceptance, to require changes to the Contribution, including changes to the length of the Contribution, and the right not to publish the Contribution if for any reason such publication would in the reasonable judgment of Wiley, result in legal liability or violation of journal ethical practices.

H. USE OF INFORMATION

The Contributor acknowledges that, during the term of this Agreement and thereafter, the Owner (and Wiley where Wiley is not the Owner) may process the Contributor's personal data, including storing or transferring data outside of the country of the Contributor's residence, in order to process transactions related to this Agreement and to communicate with the Contributor, and that the Publisher has a legitimate interest in processing the Contributor's personal data. By entering into this Agreement, the Contributor agrees to the processing of the Contributor's personal data (and, where applicable, confirms that the Contributor has obtained the permission from all other contributors to process their personal data). Wiley shall comply with all applicable laws, statutes and regulations relating to data protection and privacy and shall process such personal data in accordance with Wiley's Privacy Policy located at: https://www.wiley.com/en-us/privacy.

[X] I agree to the COPYRIGHT TRANSFER AGREEMENT as shown above, consent to execution and delivery of the Copyright Transfer Agreement electronically and agree that an electronic signature shall be given the same legal force

as a handwritten signature, and have obtained written permission from all other contributors to execute this Agreement on their behalf.

Contributor's signature (type name here): Luke John Stanley Green

Date: August 11, 2020

SELECT FROM OPTIONS BELOW:

[X] Contributor-owned work

[ ] U.S. Government work

Note to U.S. Government Employees

*A contribution prepared by a U.S. federal government employee as part of the employee's official duties, or which is an official U.S. government publication, is called a "U.S. government work", and is in the public domain in the United States. If the Contribution was not prepared as part of the employee's duties, is not an official U.S. government publication, or if at least one author is **not** a U.S. government employee, it is not a U.S. government work. If at least one author is **not** a U.S. government employee, then the non-government author should also sign the form, selecting the appropriate ownership option. If more than one author is not a U.S. government employee, one may sign on behalf of the others.*

[ ] U.K. Government work (Crown Copyright)

Note to U.K. Government Employees

For Crown Copyright this form should be signed in the Contributor's signatures section above by the appropriately authorised individual and uploaded to the Wiley Author Services Dashboard. For production editor contact details please visit the Journal's online author guidelines. *The rights in a contribution prepared by an employee of a UK government department, agency or other Crown body as part of his/her official duties, or which is an official government publication, belong to the Crown and must be made available under the terms of the Open Government Licence. Contributors must ensure they comply with departmental regulations and submit the appropriate authorisation to publish. If your status as a government employee legally prevents you from signing this Agreement, please contact the Journal production editor. If this selection does not apply to at least one author in the group, this author should also sign the form, indicating transfer of those rights which that author has and selecting the appropriate additional ownership selection option. If this applies to more than one author, one may sign on behalf of the others.*

[ ] Other

Including Other Government work or Non-Governmental Organisation work

Note to Non-U.S., Non-U.K. Government Employees or Non-Governmental Organisation Employees

For Other Government or Non-Governmental Organisation work this form should be signed in the Contributor's signatures section above by the appropriately authorised individual and uploaded to the Wiley Author Services Dashboard. For production editor contact details please visit the Journal's online author guidelines. *If you are employed by the Australian Government, the World Bank, the World Health Organization, the International Monetary Fund, the European Atomic Energy Community, the Jet Propulsion Laboratory at California Institute of Technology, the Asian Development Bank, the Bank of International Settlements, or are a Canadian Government civil servant, please download a copy of the license agreement from http://www.wileyauthors.com/licensingFAQ and upload the form to the Wiley Author Services Dashboard. If your status as a government or non-governmental organisation employee legally prevents you from signing this Agreement, please contact the Journal production editor.*

Name of Government/Non-Governmental Organisation:

_____

[ ] **Company/institution owned work (made for hire in the course of employment)**
For "work made for hire" this form should be signed and uploaded to the Wiley Author Services Dashboard.
For production editor contact details please visit the Journal's online author guidelines. *If you are an employee of Amgen, please download a copy of the company addendum from http://www.wileyauthors.com/licensingFAQ and return your signed license agreement along with the addendum. If this selection does not apply to at least one author in the group, this author should also sign the form, indicating transfer of those rights which that author has and selecting the appropriate additional ownership selection option. If this applies to more than one author, one may sign on behalf of the others.*

Name of Company/Institution:

_____

Authorized Signature of Employer: _____

Date: _____

Signature of Employee: _____

Date: _____