EXAMINING THE IMPACT OF MULTIPLE TESTS ON METAMEMORY FOR EMOTIONAL IMAGES

by © Chavon Danial Gonsalves

Thesis submitted to the School of Graduate Studies

in partial fulfillment of the requirements for the degree of

Master of Science/Psychology/Science

Memorial University of Newfoundland

October 2023

St. John's, Newfoundland and Labrador

Abstract

Individuals engage in metamemory when they perceive, control, and monitor their memories. Metamemory research often focuses on participants' accuracy in predicting future memory performance, commonly referred to as judgements of learning (JOLs). Participants often show higher JOLs for emotional (especially negative) content than for neutral content, but their recognition accuracy is often contradictory to these predictions, with better performance for neutral content. As JOLs may be the result of misunderstanding test conditions, participants may benefit from experience with test conditions to calibrate themselves better and adjust their JOLs to match their recognition accuracy. In the present study, participants studied a list of positive, negative, and neutral images while providing JOLs, and then completed an old/new recognition test. They then completed a second block of the same procedure, but with new images. JOLs were highest for negative emotional images in the first block, but recognition accuracy was highest for neutral images. JOLs were even less accurate on the second block, again showing the highest JOLs but the lowest recognition accuracy for negative images; experience did not calibrate participants to provide more accurate JOLs. Theoretical implications regarding the impact of emotion on metamemory for images and future directions of research are discussed.

General Summary

The ability to introspect and assess our own memories is referred to as metamemory. When predicting how well one will remember certain information in the future, individuals rely on several different sorts of cues, such as emotional content. Individuals often predict better memory for emotional content than for neutral content, but this pattern is only found in some types of memory tests, and the opposite pattern is found with other tests. Such a finding could be the result of a lack of experience with certain types of tests, and therefore experiencing such a test may help improve prediction accuracy. Thus, the present study consisted of two tests where participants studied emotional and neutral images before completing a memory test. Participants' predictions were highest for emotional images, but their memory in both tests was highest for neutral images. Implications from these findings about metamemory for emotional content are discussed.

Acknowledgements

I would like to thank my supervisor, Dr. Kathleen Hourihan for guiding me throughout the extensive research involved in producing this Master's thesis. I would also like to thank committee members Dr. Darcy Hallett and Dr. Jon Fawcett for their insightful and constructive comments on this work. Lastly, I am grateful to my mother for her unwavering support throughout this entire program.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Metacognition involves the scientific study of the human mind's ability to monitor and control itself or, in other words, the study of our ability to know about our knowing. Therefore, metamemory is the study of how individuals perceive, control, and monitor their memories or, in simpler terms, the study of remembering and reflecting on our memories (Dunlosky & Bjork, 2008). Much metamemory research focuses on the accuracy of participants' prediction of future memory performance on a subsequent memory test, commonly referred to as judgements of learning (JOLs). In a classic JOL paradigm, participants are required to memorize stimuli, often words or word pairs, and then predict the likelihood that they will remember each item in a subsequent memory test. JOLs are typically provided at the time of or immediately after an item has been studied. Metacognitive monitoring, such as a JOL, can be viewed as being integral in guiding an individual's learning behaviour (e.g., Finn & Metcalfe, 2007). Metacognitive monitoring is comparatively accurate but there are various instances in which there is a discrepancy between metacognitive monitoring (e.g., JOLs) and memory performance (e.g., Benjamin et al, 1998; King et al., 1980; Koriat & Bjork, 2005).

Much research has explored when JOLs for stimuli (e.g., words, faces and images) are accurate and inaccurate in terms of their correspondence with memory performance (e.g., Hourihan & Bursey, 2017; Nomi et al., 2013; Zimmerman & Kelley, 2010). Individuals make JOLs by an inferential process that is predicated on various types of information available at the time of judgements (Koriat, 1997). For instance, emotion is one cue that may be used when making JOLs. There has been an increase in research on how emotion influences metamemory in the past decade, and it has been shown that judgements are responsive to emotion, but accuracy varies (e.g., Hourihan, 2020; Hourihan & Bursey, 2017; Tauber et al., 2017). The current thesis

will examine the impact of immediate test experience on the accuracy of metamemory monitoring, measured by JOLs and recognition, for emotional images.

1.1 Memory Monitoring

In most studies examining metamemory, participants are presented stimuli which they are instructed to study in sets of trials. During the study phase, stimuli are often presented in a random order and individually in formats that vary depending on the experiment. For instance, stimuli are often presented on a computer screen or paper sheet. Following the presentation of each stimulus, participants then provide a JOL for the respective stimulus then complete the same task for the following stimulus until all study items have been presented. Memory performance, often measured by recognition or recall accuracy, is then compared with participant JOLs; researchers examine the correspondence or levels of a discrepancy between JOLs and memory accuracy, and/or whether differences in JOLs of varying stimuli show the same pattern as memory performance.

Koriat (1997) proposed a cue-utilization theory to explain how JOLs are made. The cue-utilization view assumes that JOLs are inferential and may be based on one or more of three types of cues that lead subjects to said inference: intrinsic, extrinsic, and mnemonic. Intrinsic cues involve study item characteristics that assess a subjective level of ease or difficulty in learning. In the case of paired associates, for instance, degree of associative relatedness within a pair may strongly predict memory performance and related pairs are accurately predicted to be better recalled than unrelated pairs (e.g., Rabinowitz et al., 1982). Extrinsic cues pertain to learning conditions such as presentation time, which has been shown to affect JOLs as a function of memory fluency (Mazzoni et al., 1990), and the number of times an item has been studied, which has been shown to improve JOL accuracy (Lovelace, 1984). Mnemonic cues may signal to

the participant the extent of memory fluency for an item, such as the ease with which an item comes to mind (Kelley & Lindsay, 1993) and ease of processing (Benjamin & Bjork, 1996). One or more cues of the varying types may be used in either an implicit or explicit manner to arrive at an actual JOL. For instance, both intrinsic and extrinsic factors can directly affect JOLs by an explicit application of theoretical principle.

In Koriat's (1997) supporting experiments, participants were instructed to study paired associate words for a memory test administered later in the study. Each pair contained a cue word (stimulus term) and target word (response term). Following each study trial, participants were then asked to predict the likelihood in which they would remember the paired associate. In the test phase, participants were presented with each of the cue words in turn, and their recall of the corresponding target words was tested. Recall was then compared with participant JOLs. Koriat (1997) sought to examine the impact of multiple study-test trials on JOL and recall accuracy and accordingly repeated the aforementioned study-then-test procedure with the same paired associates. JOLs and recall accuracy were positively correlated and both measures significantly increased across trials.

When comparing JOLs to memory, there are two ways of considering the accuracy of the correspondence between predictions and performance. Calibration refers to the degree to which the average magnitude of JOLs corresponds to the actual magnitude of memory performance (Dunlosky & Metcalfe, 2009). An individual would have perfect calibration if they predicted 75% recall across stimuli and recalled 75% percent of stimuli; their average JOLs were identical to their average performance. Calibration indicates whether an individual can estimate their actual level of test performance. The other measure of metamnemonic accuracy is resolution, which indicates the extent to which a participant's individual JOLs predict the memory

performance outcome (i.e., remembered or not remembered) in one item relative to another (Dunlosky & Metcalfe, 2009). Resolution is utilized to assess whether items are more likely to be correctly recalled at test. Resolution indicates whether a person can discriminate between varied memorability of stimuli, whereas calibration indicates whether a person can estimate the actual level of performance.

Nelson and Dunlosky (1991) discovered the delayed-JOL effect, where delaying a JOL by several intervening trials after the item's initial presentation drastically improves resolution. They constructed a metamemory test for word pairs. Items were split into two blocks where participants would assign immediate JOLs to half of the items in one block and a delayed JOL to the other half of the same block, in a random order. Nelson and Dunlosky found that resolution was significantly greater for delayed-JOL items than for immediate-JOL items. Mnenomic cues, such as the feeling of retrieval fluency attempts when making a JOL, may be highly important for accurately predicting future memory for individual items.

1.2 Effects of Test Experience on Calibration

Many other experimental factors have been shown to influence calibration, including test experience. For example, JOLs may demonstrate systematic biases such as what Koriat et al. (2002) coined the *underconfidence-with-practice effect (UWP)*. The UWP effect is observed when participants shift their overconfidence (participant recall is lower than mean JOL magnitude) to underconfidence (participant recall is higher than mean JOL magnitude) in subsequent learning phases; indeed, the UWP effect is characterized by JOLs that underestimate recall performance in subsequent tests. Such a finding was also observed in the work of Koriat (1997) when participants were moderately overconfident in the first trial and then underconfident

in subsequent trials; their JOLs were higher than recall for the first trial, then the reverse was observed in subsequent trials.

Koriat et al. (2002) instructed participants to study a list of word pairs for two tests and found that participants displayed slight overconfidence for the first presentation. However, they significantly underestimated JOLs for future recall performance on a repeated presentation of the study materials. Recall increased by 23% between the first two presentations of word pairs but JOLs only increased by 6% and participants continued to underestimate their recall following the third and fourth presentations of word pairs. The UWP effect has also withstood multiple experimental manipulations (e.g., when participants are provided feedback about the correctness of their responses during a test; Dunlosky & Bjork, 2008), displaying its robustness and more of a reason to investigate the effects of subsequent testing in our proposed study. The UWP effect was also demonstrated to occur for both easy and difficult items (Finn & Metcalfe, 2007; Koriat et al., 2002).

The work of Koriat (1997) and Koriat et al. (2002) showed a significant UWP effect regardless of whether participants received feedback. However, there still exists a possibility that UWP may be mediated by feedback when considering confidence-recall research. Theoretically, feedback may produce a general improvement in metamemory accuracy by enabling participants to learn from discrepancies between confidence judgements and actual performance. A participant may learn to reduce overconfidence in some items, and more importantly, reduce underconfidence in other items if provided with item-by-item feedback. Kulhavy and Stock (1989) distinguished between two forms of feedback: verification and elaboration feedback. Verification feedback indicates whether a response was correct. Elaboration feedback provides information in addition to verification, such as what the correct answer was. Both forms of

feedback may be presented as a summary across items, such as percentage correct, or on an item-by-item basis as a test phase progresses (Thompson, 1998).

Thompson (1998) implemented item-by-item verification feedback to address how confidence-recall accuracy is affected by recall performance across varying sets of test items. Participants were presented with multiple sets of general knowledge questions which could be correctly answered with one word, with varied levels of difficulty. A level of confidence was requested when a participant responded to a question. Participants were placed in two conditions, where they either received or did not receive immediate verification feedback following each response and each group were presented two sets of questions. Participants in the feedback condition were more confident, and demonstrated higher metamemory accuracy, than the no-feedback condition. General metamemory accuracy did not improve with feedback. Rather, improvement in metamemory accuracy was specific to the questions for which feedback was provided; metamemory accuracy improvement was not observed in the no-feedback condition. We hope that, for the present study, participants only require the immediate test experience to be able to make more informed JOLs and recognition decisions on the proceeding study-test block. This coincides with the finding that the UWP effect occurs without feedback, which suggests that participants have some idea of how they are performing without feedback.

Participant JOLs are often predicated on the outcome of retrieval attempts on the previous test trial, which is a strong predictor of performance on the next trial (Finn & Metcalfe, 2007; Vesonder & Voss, 1985). This information is known as the memory for past-test heuristic, and it is believed to partially account for improvements in the accuracy of participants' JOLs in subsequent study phases in metamemory experiments (Serra & Ariel, 2014). When devoid of superior diagnostic information, the past-test heuristic postulates that participants may rely on

their recollection of previous test performance when making their JOLs (Finn & Metcalfe, 2007). Finn and Metcalfe (2007) tested whether the memory for past-test heuristics could provide an explanation for changes in JOL resolution across study-test blocks and found that past-test heuristics are one account for the UWP effect. Performance on subsequent blocks of the study was influenced by its predecessors.

Past research has demonstrated that test experience can also affect the way in which participants interpret cues when making JOLs, as well as their memory performance. For instance, Benjamin (2003) examined the impact of word frequency (WF) on *pre*dicting (a judgement made before test) and *post*dicting (a judgement made after a test response) metacognitive judgements. The effect of word frequency on old/new recognition may be best explained as a mirror effect, where the condition that elicits a higher hit rate (when a studied item is correctly identified as studied) also elicits a lower false alarm rate (when a participant incorrectly believes a new item was studied). Within the context of word frequency, the mirror effect is observed when low-frequency (LF) words are more likely to be recognized than high-frequency (HF) after study, and are less likely to be falsely recognized if they were not studied. Theoretically, studied low-frequency (LF) words benefit from exposure due to more efficient coding which could be a function of their distinctiveness. Conversely, a lower false-alarm rate elicited by LF words may be because they are less familiar pre-experimentally, and consequently, less memorable than high-frequency (HF) words (Glanzer & Adams, 1990; Glanzer & Bowles, 1976; Schulman, 1967).

It is suggested that participants rely on different cues when making metacognitive judgements during study from when they are making judgements in a recognition test (Benjamin, 2003; see also Glanzer & Bowles, 1976; Gorman, 1961). In three experiments, participants

studied a set of words and predicted the likelihood in which they would remember each word in a subsequent old/new recognition memory test. At test, participants then provided a postdiction for words they judged to be new (i.e., they were asked to predict the hypothetical likelihood of recognizing the word if it had been studied). In Experiment 1, participants provided higher predictions for HF words than for LF words, but higher postdictions for LF words than for HF words. Participants predicted the incorrect pattern for WF at study, but postdicted the correct pattern for WF at test.

Experiment 2 further examined whether the effects of making test trial postdictions would transfer to future JOLs and recognition. Participants in Experiment 2 completed two study-test phases, and participants were separated into two conditions: postdiction and no postdiction. Participants in the no postdiction condition completed the recognition test without an opportunity to make metacognitive judgements following answering whether they recognized an item in the test phase. Participants in the postdiction group made a metacognitive judgement following each test trial where they called the word "new" (i.e., their belief that they would have recognized a particular word had it been studied). Participants in both conditions replicated the findings of Experiment 1 in the first study-test phase, where HF words elicited higher predictions than LF words at study, but higher recognition accuracy for LF words than for HF words. On the first test, the postdiction group showed higher postdictions for LF words than for HF words, again replicating the first experiment. In the second study phase, the non-postdiction group showed the same pattern as in the first study phase: higher predictions for HF than LF words. Conversely, the postdiction group showed a reversal of this pattern at study and LF words received higher predictions than HF words. Both groups exhibited higher recognition for LF words than for HF words and did not significantly differ from each other.

These findings suggested that participants in the postdiction group learned something about predicting their future recognition, which Benjamin (2003) examined in Experiment 3. It was possible that making postdictions during a recognition test led participants to believe that common words are less memorable on a recognition test than anticipated. Alternatively, they may have instead formed the incorrect belief that uncommon words are more memorable than common words regardless of the specific test format. If this were the case, theoretically, participants may then subsequently mispredict better *recall* of uncommon words, following an experience in which they found uncommon words to be easier to *recognize* than common words. However, if participants correctly learned that it is the nature of old/new discrimination that benefits uncommon words relative to common words, then they should still be able to make correct predictions about recall of common words. Therefore, participants completed two study and test phases in Experiment 3. The first was identical to Experiment 2, where participants, comprising the control group, completed a single study and recall test phase. The experimental group completed two study and test phases, with the first test consisting of old/new recognition with postdictions, as in Experiment 2. Following the first test, they were informed of basic free recall instructions, and were then presented a new list to study while again making predictions. They then completed the free recall task.

It was found that postdictions of recognition performance during the recognition test did not significantly affect predictions of recall—the experimental group showed higher JOLs for common than uncommon words, just as the control group did. Therefore, it may not be inferred that participants incorrectly learned that uncommon words are more memorable than common words, but rather, participants learned novel information about the demands of a recognition test. Specifically, participants learned that they are better at discriminating uncommon words than

they are at common words. Such a finding demonstrates that participants can learn of the difficulties in recognizing specific types of items on recognition tests. They apply this new-found knowledge to subsequent JOLs for the same type of items. Moreover, these findings demonstrate that participants did not overgeneralize the idea that LF words are easier to remember than HF words for all test types and realized that this idea only applies to recognition. To our knowledge, no one has completed multiple study-test cycles with emotional information, so it is yet unknown whether participants can learn to accurately predict the effects of emotion on recognition memory when given test experience.

1.3 Memory and Metamemory for Emotional Information

This thesis will only examine episodic memory for emotional stimuli, not emotional states, nor autobiographical experience. Emotion, within the context of this thesis, refers to the respective categories of stimuli (i.e., neutral, positive, negative) which depend on various levels of valence and arousal. Valence refers to the degree to which an item is pleasant or unpleasant, and it is often divided into positive, negative, and neutral categories. Norming work on emotional items involves participants rating on a scale that ranges from unpleasant at the low end, through neutral, to pleasant at the high end. For instance, many studies on emotional images curate study and test lists based on the International Affective Picture System (IAPS; Lang et al., 2008) (e.g., Hourihan, 2020; Tauber et al., 2017), where images are split into emotional conditions based upon valence ratings**.** Arousal refers to the degree to which an item elicits one's emotions, ranging from calm to exciting (Bradley et al., 1992). A negative item, for instance, possesses relatively low valence and high arousal. Conversely, a positive item will have relatively high valence and high arousal. A neutral item is often neither high nor low in valence or arousal but is typically lower in arousal than either positive or negative emotional items.

Perhaps unsurprisingly, memory for emotional items is often superior to that of neutral items (e.g., Kensinger & Corkin, 2003). Negative emotional stimuli has demonstrated an ability to enhance memory for essential details while positive emotional stimuli may lead to an increased memory for peripheral image details (Kensinger, 2009). Effects of enhanced memory are more pronounced in negative items compared to neutral, rather than when positive items are compared to neutral items.

Research has demonstrated higher JOLs and memory for emotional information. Zimmerman and Kelley (2010) were the first to examine the role of varied levels of emotionality in metamemory. They assessed the effect of emotional content on JOLs for words in cued and free recall paradigms. Participants demonstrated higher JOLs for positive and negative emotional words than for neutral words which was consistent with free recall performance. However, this predicted future memory was not consistent with cued recall of negative emotional words compared to positive and neutral words; rather, positive words were the sole emotional condition to demonstrate higher cued recall than neutral words. Participants overestimated their abilities to recall negative words in a cued recall paradigm.

Zimmerman and Kelley (2010) inferred that valence and task (i.e., the form of memory recall) yield varied effects, such that JOLs for emotional items may be most accurate in the test context that matches participants' expectations at the time of judgement. Memory test performance was measured solely with free recall in the second experiment, while Experiments 1 and 4 implemented cued recall. Experiment 3 was the lone experiment to implement conditions where participants' memory performance was measured with either free or cued recall. Metamemory monitoring of free recall was generally accurate for emotional and neutral words. However, negative word pairs were not more memorable in cued recall conditions which

indicated overconfidence in negative word pairs. Positive pairs, on the other hand, were consistently recalled better than both negative and neutral word pairs in cued recall conditions. This first example of metamemory for emotional information, alone, found that the format in which a test is administered may play an integral role in determining how accurately participants predict future memory.

As further evidence that factors beyond the intrinsic cues of valence and arousal contribute to the accuracy of JOLs for emotional items, Hourihan et al. (2017) conducted a study that comprised three experiments to observe the intricacies of how emotional content influences JOLs. Specifically, they were interested in why and how emotional factors influence JOLs for words. Each experiment examined two different emotional components—valence and arousal—on JOLs. Understanding the influence of these components separately is critical as studies deliberately curate emotional word lists to significantly differ from neutral words in both valence and arousal (Tauber & Dunlosky, 2012; Zimmerman & Kelley, 2010). The first two experiments isolated and separately examined arousal and valence on JOLs and replicated prior findings that participants provide higher JOLs to emotional words than to neutral words even when emotional words differed from neutral words only in valence and not arousal. Such a finding suggests that these effects of metamemory are not solely reflected by physiological emotional arousal experienced at the time of encoding. Words in the third experiment demonstrated variability in levels of valence and arousal, and participant JOLs and recall were not significantly impacted by either emotional component. Such a finding suggests that rather than a physiological factor influencing emotion, a cognitive factor at the encoding level impacts JOLs for emotional content. That is, participants may intentionally assign higher JOLs to emotional words, but only when the study context clearly highlights the fact that some words are indeed emotional.

Tauber and Dunlosky (2012) were the first to explore the monitoring of learning of emotional materials influenced by aging. They examined the accuracy of JOLs for emotional words in young and older adults in terms of JOL resolution and sensitivity to emotional stimuli. Older adults demonstrated JOLs that were higher for negative words than for neutral words which accurately reflected recall performance. These findings did not significantly differ from younger adults. In contrast, older adults' JOLs were less sensitive to words with high positive valence which may be explained by ceiling effects. Moreover, JOL resolution of older adults was at chance level and significantly lower than young adults' resolution. Tauber and Dunlosky (2012) concluded that monitoring of learning emotional materials is generally maintained with healthy aging.

Higher JOLs for emotionally positive faces compared to neutrally emotional faces are well-documented. Nomi et al. (2013) conducted a study in which they examined how emotional facial expressions influence participants' JOLs. Participants were told that they would view a set of faces to study and then be asked to identify the emotional expression in the face displayed for a memory test. Each studied face was denoting one of three emotional expressions: 1) happy (positive), 2) neutral, or 3) angry (negative) facial expression. At study, participants provided JOLs for each image (via computer) on a scale of 50-100% to indicate their predicted level of confidence in which they would be able to select the studied face at test. Confidence judgements in recognition were made at test. Results demonstrated higher JOLs for images of studied faces with positive or negative expressions than for neutrally emotional studied faces. In contrast, participants had higher recognition accuracy for images with emotionally neutral facial expressions than for emotionally positive or negative images. This is the first known instance of

non-word emotional stimuli demonstrating the opposite pattern when comparing JOLs to recognition.

Moreover, Witherby and Tauber (2018) assessed JOL sensitivity for universal emotions and categories of negative valence because there are fewer categories of positive valence than there are of negative valence. For example, the categories of positive valence are generally happy and surprised. On the other hand, negative valence categories comprise angry, afraid, sad, and disgusted (Witherby & Tauber, 2018). However, to avoid cue-overload, Witherby and Tauber (2018) elected to investigate only three negative expression categories (sad, angry, and afraid). Additionally, one of the research goals was to measure the beliefs which participants have about how the aforementioned negative expressions affect their memory which is applicable to JOLs. Similar to the work of Nomi et al. (2013), participants were presented with faces depicting varied emotional expressions (in this instance, neutral, sad, afraid, or angry). Participants demonstrated higher JOLs for images with negative emotional facial expressions than for images with neutral emotional facial expressions. Relative to neutral emotional facial expressions, participants demonstrated higher JOLs for each negative valence expression (i.e., sad, angry, and afraid). However, JOLs did not significantly differ among the various types of negative expressions, and recognition was unaffected by expression type.

1.4 Metamemory for Emotional Images

Beyond the scope of emotional words and faces, higher JOLs for emotional images are well-documented. For instance, Tauber et al. (2017) built upon the aforementioned work of Tauber and Dunlosky (2012) and conducted two experiments to examine whether the age-related difference in metamemory for emotional words extends to images. Tauber et al. (2017) evaluated the effects of valence and arousal on young and older adults' JOLs, recall, and recognition for

emotional pictures. Participants studied and were tested on images that were either neutral or positive and low or high in arousal. Participants provided JOLs immediately following each trial and then were tested on all images. Experiment 1 demonstrated that JOLs were higher for positive than for neutral images; this finding suggests that the magnitude of JOLs were influenced by valence. Experiment 2 factorially manipulated valence and arousal of the to-be-studied images. In both experiments, younger adults demonstrated enhanced memory for high-arousal images relative to low-arousal images. Regardless of arousal, JOLs were higher for images with positive valence than for images with neutral valence. Moreover, recall was higher for images with positive rather than neutral valence. In the present study, we will place emphasis on valence, but are cognizant of the fact that images with high valence also have higher levels of arousal, which likely heavily impacts the cues participants utilize in making JOLs.

Thus far, the literature review in the present study has confirmed that emotion is an intrinsic cue that normally affects memory. JOLs, consequently, should be sensitive to emotion. Moreover, emotional content is a significant cue in pictorial stimuli and may elicit varying physiological responses which may not be replicated when viewing emotionally neutral images (Bradley et al., 2001). Hourihan and Bursey (2017) conducted a study that intended to assess how the emotional content of images impacts recognition and JOLs. They hypothesized that neutral images would be recognized less accurately than positive emotional images and that participants would report lesser JOLs for neutral pictures than for positive emotional pictures. JOLs were substantially higher for positive emotional images than for neutral emotional images in both experiments of the study. Recognition patterns were surprisingly inconsistent and did not correspond to the provided JOLs. Participants demonstrated higher recognition for neutral images than for positive images in Experiment 1 while demonstrating no significant difference

between neutral and positive emotional images in Experiment 2. JOLs being significantly higher for positive images than for neutral images in both experiments is congruent with the theory that participants are reliant upon the intrinsic characteristics of the current stimuli being examined. Participants may have inferred that the happiness they experienced when observing emotionally positive images would predict recognition memory, which may account for the overestimated recognition discrimination for positive images.

It is evident that much of the preceding research examining and comparing JOLs for emotional and neutral images demonstrate reliably higher JOL for emotional images, and in particular, positive versus neutral images (e.g., Hourihan & Bursey, 2017; Tauber et al., 2017). However, often the case is that participant JOLs exceed their recognition accuracy when tested; that is, participants often overestimate their subsequent memory when making JOLs. For instance, as aforementioned, Hourihan and Bursey (2017) found that JOLs for emotionally positive images were reliably higher than for emotionally neutral pictures but produced surprising results; the recognition discriminability was superior for neutral emotional images than for emotionally positive images. It appears that positive emotional stimuli failed to benefit image recognition. Interestingly, Hourihan and Bursey (2017) noted that the memory benefit for negative emotional stimuli is more reliably observed than for positive content and thus set the groundwork for further research examining JOLs for negative emotional images along with positive and neutral emotional images.

There are reasons to believe that JOLs for positive and negative emotional images should differ, such as their respective differences in types of remembered details. As previously stated, negative stimuli has demonstrated an ability to improve memory for essential details. Positive emotional stimuli, on the other hand, may lead to enhanced memory for peripheral image details

(Yegiyan & Yonelinas, 2011). Valence and arousal apparently simultaneously impact JOLs for positive and negative images in different ways. This is congruent with the theory that participants are reliant upon intrinsic characteristics of the current stimuli being examined. Emotional content is a significant cue in pictorial stimuli and may elicit varying physiological responses which may not be replicated when viewing neutrally emotional images. Thus, the way in which participants encode emotional stimuli based on their valence may influence their JOLs and could contribute to the discrepancy between JOL and test performance (Hourihan, 2020). Moreover, JOLs are influenced by subjectivity, and the subjective experience of viewing and attempting to process an image differs between negative and positive emotional stimuli (e.g., Pérez-Mata et al., 2012). In addition, recognition memory accuracy for negative images may surpass recognition for neutral images if negative emotional images are associated with increased memory for central image details.

Hourihan (2020) took these factors into consideration and examined whether positive and negative emotional images differ in terms of metamemory accuracy. Hourihan (2020) followed Hourihan and Bursey (2017) and hypothesized that participants would have higher JOLs of positive emotional images than they would for neutral emotional images and demonstrate equal or higher recognition accuracy for neutral images. Moreover, given the aforementioned information pertaining to negative emotional images, participants were expected to demonstrate higher recognition accuracy for negative emotional images than for their neutral counterparts. Lastly, a more robust correlation between recognition accuracy and preceding JOLs was expected to arise from participants displaying a greater recollection for details of emotionally negative images. Hourihan's (2020) study consisted of a paradigm that followed that implemented by Hourihan and Bursey (2017). In the study phase of the experiment, participants

were asked to study a list of emotional images for a subsequent memory test and provide JOLs immediately following the viewing of each image. Participants demonstrated higher JOLs for negative emotional images than for positive images and each of these conditions was higher than JOLs for emotionally neutral images. However, participant JOLs for emotional images differed from recognition accuracy. On average, participants predicted the highest recognition for negative emotional images. However, negative emotional images garnered the lowest recognition accuracy of all images. These results are similar to those of Hourihan and Bursey (2017) where participants predicted higher recognition accuracy for positive than for neutral emotional images. Participants demonstrated an overestimation of JOLs in both studies. Surprisingly, participants demonstrated the lowest memory accuracy in negative emotional images.

One possible explanation for this result that Hourihan (2020) suggested was that participants did not expect, and consequently failed to appreciate, the demands of a recognition test. It is possible that participants had never completed an old/new recognition test. In such a test, researchers provide participants with each of the previously studied images along with new images from their respective emotional categories which are often varied combinations of positive, negative and neutral (Hourihan, 2020). As JOLs are made during the initial study phase of the experiment, participants may fail to appreciate the challenges in distinguishing stimuli that share semantic and visual information when predicting higher recognition of emotional than neutral images. If participants have limited experience in old/new recognition tests, they would lack necessary skills to make informed and accurate JOLs that account for the challenges of differentiating studied items from new items from the same emotional category. Therefore, the present study will extend the research of Hourihan (2020) and answer the question of whether inaccurately high JOLs for emotional images are a result of a misunderstanding of test conditions

that can be corrected with subsequent testing. In the present study, participants studied negative, positive and neutral emotional images in a set of trials where they provided a JOL for each trial in what comprised the study phase. The test phase consisted of an old/new recognition test, where participants answered whether an image was studied or new and provided a confidence judgement for each trial. Participants then studied and completed an additional old/new recognition test with a new set of images.

While being mindful of the possibility that participants may demonstrate the UWP, we theorize that inaccurate JOLs for emotional images are mainly the results of participants' inability to foresee and therefore appreciate the challenges of old/new recognition test conditions. Hourihan (2020) observed both lower hits and higher false alarms for negative images relative to neutral images. Thus, participants may fail to appreciate the level of difficulty in judging whether related images are old or new. As previously discussed, Benjamin (2003) demonstrated that participants may increase their accuracy in accounting for the effects of word frequency on recognition in subsequent JOLs, after they have had test experience with those items.

The experience which participants acquire from repeated recognition tests may elicit an increased understanding and appreciation for recognition test conditions; this should result in a corresponding change in JOLs with a second opportunity to make predictions. As described above, Benjamin (2003) demonstrated that participants may learn to make more accurate JOLs with word frequency for recognition. With multiple tests, we predict that participants will modify their JOLs to account for the previously unexpected challenges of recognition testing and the discrepancy between JOLs and recognition accuracy should lessen. We especially expect to replicate the findings of Hourihan (2020) on the first block—participants should demonstrate the

highest JOLs for negative, then positive, then neutral images, but recognition discriminability would show the opposite pattern, where participants demonstrate the highest recognition discriminability for neutral, then positive, then negative images. The critical findings lay within the second block. Should participants account for the challenges of a recognition test after immediate experience, they will better calibrate themselves and change the pattern of their JOLs during the second study block, which should better correspond with their second measure of recognition discriminability. Thus, the discrepancy between JOLs and recognition should narrow on the second block.

Chapter 2: Method

2.1 Participants

The target sample size for the present study was determined to be 43 participants, based on the sample sizes used by Hourihan (2020) who examined metamemory for emotional images at three levels of emotion. There was a discrepancy between the number of participants obtained and the target sample size due to the unpredictability of participant sign-up rates to complete the experiment. Forty-six undergraduate and graduate students from Memorial University of Newfoundland enrolled in psychology courses received two course credits for their participation. Twenty-two percent of participants identified as men, and 78% identified as women. The mean age of the participants was 20.8 years (SD = 2.91), and 91% of participants were right-handed. Thirteen of the 46 participants completed the study in-person, and the remaining 33 participants completed the study online.

2.2 Materials

The stimulus pool (see Appendix A) comprised 456 images selected from the International Affective Picture System (IAPS) database (Lang et al., 2008), including 152

positive, 152 negative, and 152 neutral (see Table 2.1). The sets of positive and negative images did not significantly differ from one another on mean arousal (p = .123) but differed significantly in arousal when compared with neutral images (both ps < .001). Additionally, each image set differed significantly from the others in mean valence (all ps < .001). There was no effort made to control for any perceptual factors of the images selected within each emotion condition, such as sharpness, luminance, brightness, complexity, and colour. Assignment of items from each emotion condition to serve as either study items or new test items in one of the two blocks was randomly determined for each participant. At the end of the primary task, participants were shown a series of seven highly positive images obtained from freedigitalphotos.net to counteract the effect of having participants view negative images in the study. Images from this site were selected from the ones used in Experiment 2 in the work of Hourihan and Bursey (2017). Presentation of stimuli and recording of responses were completed with PsychoPy software (v 2022.2.4; Peirce et al., 2019. The study was reviewed and approved by the Interdisciplinary Committee for Ethics in Human Research (ICEHR) at Memorial University of Newfoundland (see Appendix B).

Table 2.1

Mean Valence and Arousal Ratings of Negative, Neutral, and Positive Image Pools

	Negative	Neutral	Positive
Valence (1-9)			
Mean	2.39	5.13	7.25
SD	0.43	0.41	0.43
Range	1.45—3.09	4.37—6.11	6.57–8.34
Arousal (1-9)			
Mean	5.53	3.08	5.5
SD	0.20	0.18	0.20
Range	4.00—6.77	1.72—3.71	4.55—7.35

2.3 Design

The study comprised a 3 x 2 repeated-measures design. The primary independent variable was emotion, with three levels manipulated within-subjects (negative vs. neutral vs. positive). The second independent variable was block (block 1 vs. block 2). Dependent variables included mean JOLs, recognition accuracy (hits and false alarms), *d'* (a computation acquired from hits and false alarms), *c* (response bias), mean recognition confidence, and the correspondence between JOLs and recognition confidence in the form of d_a, a measure of metamnemonic resolution.

2.4 Procedure

The present study largely followed the methodology implemented by Hourihan (2020), but with an additional study and test phase. Data were collected both in-person and online: 13 participants completed the study supervised in-person, while 33 completed the study unsupervised online. Participants who completed the study in-person provided informed consent

via a Qualtrics survey on a computer in the lab. The researcher communicated to the participants that they would be asked to study two lists of emotional images for a memory test, and participants were provided with a brief description of old/new recognition test procedures. They were informed that each image would appear for only a brief period of time in a study phase, and they would be asked to predict future recognition of each studied image immediately after presentation. The online variant of the experiment was nearly identical; the minor differences were that participants received an informed consent form and an overview of the experimental procedures online via Qualtrics and were then directed to complete the online study hosted via Pavlovia (https://pavlovia.org/).

The first study phase comprised 114 randomized trials, including 38 positive, 38 negative, and 38 neutral images. On each trial, a picture was presented for 500 ms, fit to 50% of the full screen, followed by a 250 ms blank screen, before participants were then asked to provide their JOL for the corresponding image. As with all JOLs in the experiment, this was self-paced. The bottom of the screen displayed the rating scale with numbers 1—8, with the verbal labels “I am sure that I will NOT remember this picture” below the “1” and “I am sure that I WILL remember this picture” below the “8”. The instructions remained on the screen until the participant pushed a number key from 1 to 8 to register a response then proceeded to the following image. The next image appeared after a 1000 ms blank screen.

Instructions for the corresponding old/new recognition test appeared following the final JOL. The recognition test consisted of the 228 randomized images (114 old and 114 new) from the first block, each presented one at a time. Participants were asked to indicate whether each image was one they had previously studied (by pressing the “y” key) or one they had not studied (by pressing the “n” key). Keypress labels for “old” and “new” images were situated at the

bottom of the screen, beneath the image for each trial which was presented to fit 50% of the screen. As there was no time limit, the image remained on the screen until a participant pressed one of the response keys. Following each old/new judgement, they provided their confidence in their decision on a scale from 1 (“Not at all confident”) to 8 (“Completely confident”) by pressing the corresponding number keys.

Instructions for the second half of the experiment then appeared. The second study and test phase were identical to the first but utilized the 228 images that did not appear in either the first study or test phase. Following completion of the second and final test, participants were notified that they had completed the study and would be presented with a series of highly positive emotional images, which were presented at 50% of the full screen for 1000 ms each. In-lab participants then completed a demographic questionnaire on Qualtrics and were debriefed by the researcher who remained in the test room throughout the entire experiment to answer any participant inquiries but was faced away from the participant to minimize any demand characteristics and anxiety issues. Following presentation of the positive images, online participants were then re-directed to Qualtrics to complete the demographic questionnaire, and then viewed a debriefing form and were provided with a final opportunity to consent to participate or withdraw from the study.

Chapter 3: Results

Raw data files were exported to Microsoft Excel (2021). Data organization, data cleaning and some analyses were completed with RStudio (2023.03.0 Build 386) and Microsoft Excel (2021). Inferential analyses were conducted with Jamovi (2.2.5.0).

3.1 Analytic Strategy

The goal of the analysis was to examine how the emotion conditions of images influenced participants' memory and metamemory, and whether accuracy of metamemory changed as participants gained immediate experience (i.e., from block 1 to block 2). We first began by analyzing mean JOLs across blocks. We then analyzed recognition accuracy, considering both hits and false alarms. A hit is reached when a participant correctly responds that a previously studied item has been studied. A false alarm occurs when an item has not been studied, but a participant erroneously responds that the item was seen prior to the test phase. We continued by analyzing recognition discriminability using *d'*, a computation derived from hits and alarms that incorporates both the responses to previously studied items (hits) and new items (false alarms) from the same emotion categories. Furthermore, we then analyzed *c* (response bias), a measure which quantifies participants' tendency to respond in a predominantly liberal (tending to call more items "old", indicated by negative values of *c*) or conservative (tending to call fewer items "old", indicated by positive values of *c*) direction on recognition memory tests (Deason et al., 2017). Moreover, we examined participants' mean confidence in correct recognition judgements.

JOLs and recognition confidence (i.e., rated confidence in recognition response, accounting for accuracy) were related via d_a to measure metamnemonic resolution. We used d_a to relate JOLs and recognition confidence. To align the two judgements on the same scale, we transformed raw confidence ratings to include the prior old/new judgement to function on a scale between 1—8 that ranged from 1 = sure new to 8 = sure old. This scale corresponds with the JOL scale of 1 = "I am sure will NOT remember this image", and 8 = "I am sure I WILL remember this image". That is, adjacent pairs of confidence ratings (originally made on a scale

of 1 = “not at all confident” to 8 = “completely confident”) for hits were combined into the highest four ratings on the transformed scale, and adjacent pairs of confidence ratings for misses (i.e., when a participant incorrectly identified a studied item as new) were combined into the lowest four ratings on the transformed scale. For example, an incorrect “new” judgement followed by a confidence rating of “7” or “8” (“completely confident”) was transformed to a rating of “1” (“sure new”); a correct “old” judgement followed by a confidence rating of “1” (“not at all confident”) or “2” was transformed to a rating of “5”, etc.

We decided to forgo gamma correlations as d_a is a more precise measure of resolution. The d_a statistic, a distanced-based metric rooted in signal detection theory, is the distance between the means of multiple normal distributions. On the other hand, *d'* measures the distance between probability distributions scaled by a shared standard deviation, an assumption that has been proven to be inaccurate in many circumstances. Moreover, unlike *d'*, d_a can characterize a receiver operating characteristic (ROC) in a single value (Benjamin & Diaz, 2008; Swets, 1986). Moreover, Gamma correlations, under various conditions, including situations where response bias is prevalent, produce values that can greatly deviate from the actual value. Unlike gamma correlations, d_a will provide a consistent value for both equal and unequal Gaussian evidence distributions (Masson & Rotello, 2009).

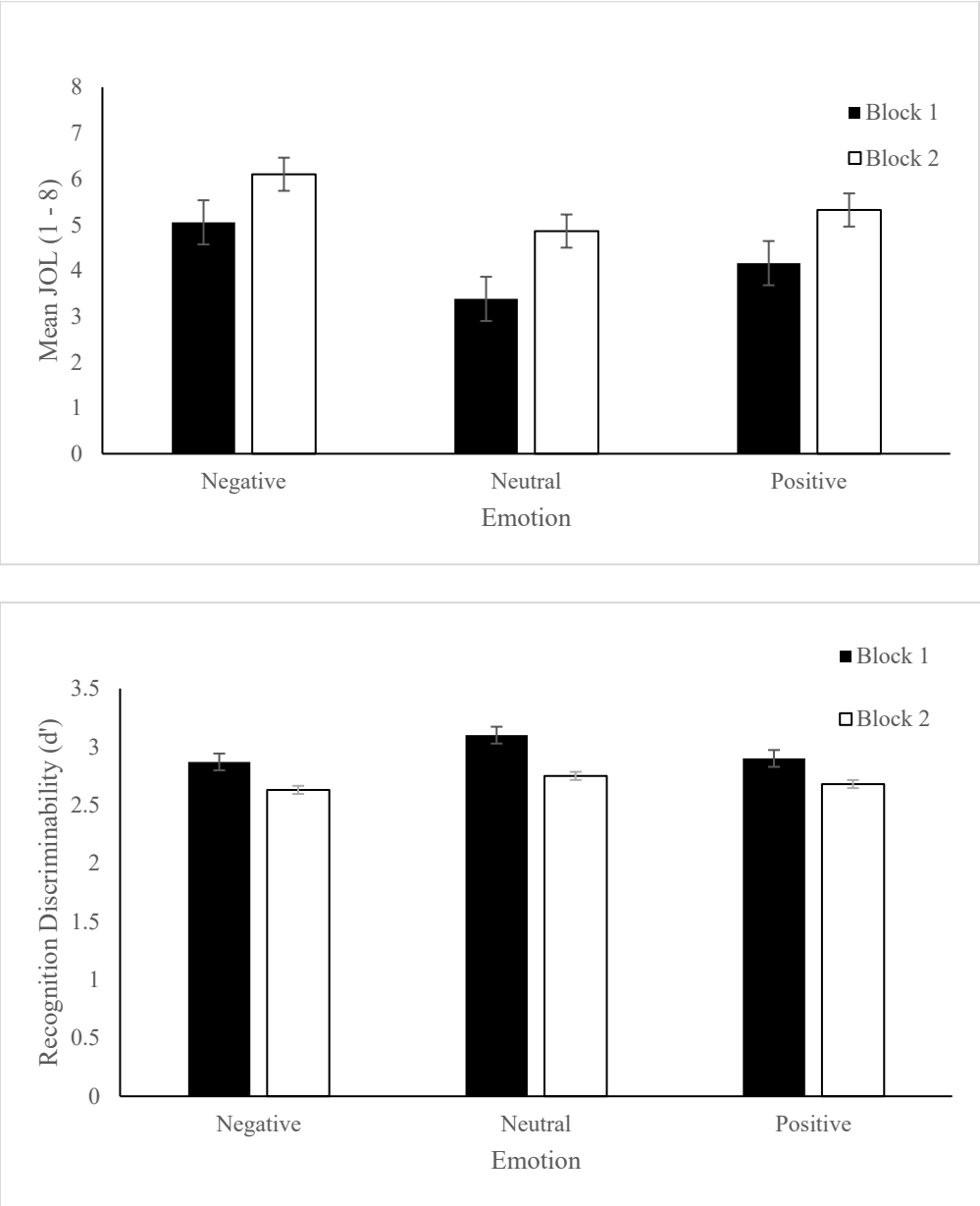
All of the measures described above were analyzed with Analysis of Variance (ANOVA), and planned comparisons reported below use uncorrected *p*-values. For any post-hoc comparisons following up on effects that were not originally anticipated, the Tukey post-hoc correction to the *p*-values was used. The planned comparisons were always to make separate comparisons between each emotion condition to one another.

3.2 Judgements of Learning (JOLs)

Mean JOLs are displayed in Figure 3.1 and were analyzed in a 2 (block: block 1 vs. block 2) x 3 (emotion: negative vs. neutral vs. positive) repeated measures ANOVA. The main effect of block on mean JOLs was significant, $F(1, 45) = 85.70$, $MSE = 1.22$, $p < .001$, $\eta_p^2 = .66$, a large effect size indicating that mean JOLs for block 2 were higher than for block 1. The main effect of emotion on mean JOLs was significant with a large effect size, $F(2, 90) = 56.81$, $MSE = 0.86$, $p < .001$, $\eta_p^2 = .56$. Planned comparisons showed that JOLs were significantly higher for negative images than for positive images with a large effect size, $t(45) = 5.46$, $p < .001$, $d = 0.82$, which, in turn, were significantly higher than JOLs for neutral images with a large effect size, $t(45) = 8.45$, $p < .001$, $d = 0.91$. The overall pattern of JOLs replicated Hourihan (2020) where participants demonstrated the highest JOLs for negative images, followed by positive then neutral images; these effects were statistically large.

Figure 3.1

Mean JOLs (top panel) and Recognition Discriminability Performance (d'; bottom panel) in Blocks 1 and 2



Note. Error bars represent the standard error of the mean.

The interaction between block and emotion was significant with a large effect size, $F(2,90) = 12.12$, $MSE = 0.09$, $p < .001$, $\eta_p^2 = .21$. To further specify the nature of this interaction, a difference score was computed for the increase in JOLs from block 1 to block 2, for the three emotion categories. These JOL increases across blocks were analyzed in a one-way repeated measures ANOVA. This revealed a significant main effect of emotion with a large effect size, $F(2, 90) = 12.10$, $MSE = 0.18$, $p <. 001$, $\eta_p^2 = .21$, with planned comparisons revealing that mean JOL increases across blocks for neutral images were significantly higher (with a small effect size) than those for negative images, $t(45) = 4.04$, $p < .001$, $d = 0.45$, and for positive images (with a small effect size), $t(45) = 4.59$, $p < .001$, $d = 0.31$. However, follow-up comparisons also revealed that mean JOL increases for negative images across blocks did not significantly differ from those of positive images, $t(45) = 1.07$, $p = .291$, $d = 0.10$. JOLs for each condition increased from block 1 to block 2, but JOLs to neutral images increased more across blocks than JOLs to images from both emotion categories. Part of our prediction was reached, as JOLs increased for neutral images, which was a larger increase than for positive and negative images. However, participant JOLs for all emotion conditions still generally increased rather than decreased.

3.2 Recognition Accuracy

As previously stated, recognition accuracy was initially examined by way of hits and false alarms. Hits (see Table 3.1) were analyzed in a 2 (block: block 1 vs. block 2) x 3 (emotion: negative vs. neutral vs. positive) repeated measures ANOVA. The main effect of block on hits was not significant, $F(1, 45) = 2.34$, $MSE = 0.01$, $p = .133$, $\eta_p^2 = .05$. However, the main effect of emotion condition approached significance, $F(2, 90) = 2.75$, $MSE = 0.00$, $p = .070$, $\eta_p^2 = .06$. Although the main effect of condition did not reach conventional levels of significance, planned follow-up comparisons were conducted, which showed that there was a marginal trend within

hits for negative and neutral images where participants demonstrated numerically lower hits for negative images than for neutral images, $t(45) = 1.99$, $p = .053$, $d = 0.15$, lower hits for negative images than for positive images, $t(45) = 2.00$, $p = .052$, $d = 0.12$, and similar hits for positive images as for neutral images, $t(45) = 0.32$, $p = .753$, $d = 0.04$. There was no interaction between block and emotion, $F(2, 90) = 0.86$, $MSE < 0.01$, $p < .425$, $\eta_p^2 = .02$. These findings only partially replicate Hourihan (2020), but, as will be discussed below, participants in the present study demonstrated hit rates which were quite high. Therefore, ceiling effects likely obscured the potential pattern we might have otherwise seen in hit rates.

Table 3.1

Recognition and Metamemory Performance in Blocks 1 and 2

	Block 1			Block 2		
Measure	Negative	Neutral	Positive	Negative	Neutral	Positive
Accuracy						
Hits	.86 (.02)	.88 (.02)	.87 (.02)	.84 (.02)	.86 (.02)	.87 (.02)
False Alarms	.07 (.01)	.06 (.01)	.06 (.01)	.10 (.01)	.09 (.01)	.11 (.01)
Discriminability (*d'*)	2.87 (.11)	3.10 (.12)	2.90 (.10)	2.63 (.12)	2.75 (.13)	2.68 (.13)
Response Bias (*c*)	0.18 (.05)	0.16 (.05)	0.20 (.05)	0.13 (.06)	0.10 (.06)	0.04 (.06)
Confidence						
Old	7.37 (.08)	7.26 (.10)	7.24 (.09)	7.43 (.07)	7.38 (.08)	7.34 (.08)
New	6.65 (.13)	6.52 (1.39)	6.43 (.13)	6.59 (.15)	6.36 (.15)	6.38 (.15)
Metamemory						
Mean JOLs	5.05 (.14)	3.38 (.19)	4.16 (.18)	6.10 (.17)	4.87 (.20)	5.32 (.19)
Resolution (d_a)	0.48 (.07)	-0.27 (.11)	0.22 (.09)	0.61 (.08)	0.25 (.08)	0.57 (.08)

Note: Standard errors are shown in parenthesis beside their respective means.

False alarms (see Table 3.1) were analyzed in a 2 (block: block 1 vs. block 2) x 3 (emotion: negative vs. neutral vs. positive) repeated measures ANOVA. The ANOVA on false alarms showed a significant main effect of block, $F(1, 45) = 16.75$, $MSE = 0.01$, $p < .001$, $\eta_p^2 = .27$; participants had higher false alarms in block 2 than in block 1. Moreover, there was no main effect of emotion condition on false alarms, $F(2, 90) = 1.67$, $MSE < 0.01$, $p < .193$, $\eta_p^2 = .04$ as well as no interaction between block and emotion, $F(2, 90) = 1.02$, $MSE = 0.00$, $p < .365$, $\eta_p^2 = .02$. These findings differ from those of Hourihan (2020), who found the highest false alarms for negative images, which was followed by positive then neutral images. However, participants in the present study demonstrated low false alarms and generally very good performance which is likely explains why the present study fails to replicate Hourihan (2020)'s pattern of false alarms.

Hits and alarms were utilized to compute *d'* (discriminability; see Figure 3.1) and *c* (response bias; see Table 3.1). Perfect hits or false alarms were corrected when necessary, by adding half a trial of error. The ANOVA on *d*' showed a significant main effect of block, $F(1, 45) = 10.83$, $MSE = 0.47$, $p = .002$, $\eta_p^2 = .19$; discriminability was worse in block 2 than in block 1. Moreover, analysis showed a significant main effect of emotion, $F(2, 90) = 3.54$, $MSE = 0.22$, $p = .033$, $\eta_p^2 = .07$. Planned comparisons showed that participants' ability to tell studied items from new items was worse for negative than for neutral items, $t(45) = 2.25$, $p = .030$, $d = 0.22$. However, there was no difference in discriminability when comparing negative and positive images, $t(45) = 0.59$, $p = .827$, $d = 0.05$. Interestingly, discriminability for neutral items also significantly differed from positive items, $t(45) = 2.16$, $p = .036$, $d = 0.17$; participants performed worse at discriminating between studied and new positive items than they did for neutral items. Analysis finally revealed that there was no significant interaction between block and emotion, $F(2, 90) = 0.68$, $MSE = 0.17$, $p = .510$, $\eta_p^2 = .01$. Therefore, recognition discriminability of the

present study only partially replicated Hourihan (2020). As with hits and false alarms, ceiling effects likely played a large part in the observed pattern.

Analysis on *c* (see Table 3.1) showed a significant main effect of block, $F(1, 45) = 7.02$, $MSE = 0.08$, $p = .011$, $\eta_p^2 = .13$. Participants were significantly less conservative in block 2 than in block 1**.** Moreover, the main effect of emotion was not significant, $F(2, 90) = 0.57$, $MSE = 0.05$, $p = .57$, $\eta_p^2 = .01$. Finally, there was no significant interaction between block and emotion condition, $F(2, 90) = 1.68$, $MSE = 0.05$, $p = .192$, $\eta_p^2 = .04$.

Mean confidence at test for correct responses only (see Table 3.2) was then analyzed in a 2 (block: block 1 vs. block 2) x 3 (emotion: negative vs. neutral vs. positive) x 2 (old/new: old vs. new) repeated measures ANOVA. The analysis did not show a main effect of block, $F(1, 45) = 0.00$, $MSE = 0.33$, $p = .969$, $\eta_p^2 = .00$. However, results showed a significant main effect of emotion, $F(2, 90) = 5.58$, $MSE = 0.24$, $p = .005$, $\eta_p^2 = .11$. Follow-up comparisons showed that participants were significantly more confident in recognizing negative images than positive images, $t(45) = 2.77$, $p = .022$, $d = 0.22$, but not compared to neutral images $t(45) = 2.17$, $p = .088$, $d = 0.17$. Moreover, there was no significant difference in confidence in recognizing neutral and positive images, $t(45) = 1.14$, $p = .498$. $d = 0.04$. The finding of a main effect for emotion was also observed by Hourihan (2020; Experiment 1).

The main effect of old/new was significant, $F(1,45) = 57.23$, $MSE = 1.74$, $p < .001$, $\eta_p^2 = .56$. Participants were significantly more confident in their recognition judgements to old items than to new items. The interaction between block and emotion was not significant, $F(2, 90) = 0.37$, $MSE = 0.05$, $p = .690$, $\eta_p^2 = .01$.The three-way interaction between block, emotion and old/new was also not significant, $F(2,90) = 1.53$, $MSE = 0.06$, $p = .222$, $\eta_p^2 = .03$. However, there was a significant two-way interaction between block and old/new, $F(1, 45) = 6.79$, $MSE = 0.16$,

p = .012, η_p^2 = .13. Follow-up analysis for the interaction between block and old/new was carried out by computing the increase in confidence across blocks, and analyzing this difference score in a 3 (emotion: negative vs. neutral vs. positive) x 2 (old/new: old vs. new) repeated measures ANOVA. This analysis showed that participants' confidence in correct recognition of old items increased more across blocks than did their confidence in correct rejection of new items, F(1,45) = 6.79, MSE = 0.31, p = .012, η_p^2 = .131. There was no main effect of emotion, F(2, 90) = 0.37, MSE = 0.09, p = .690, η_p^2 = .008. Lastly, there was no interaction between old/new and emotion, F(2, 90), 1.53, MSE = 0.12, p = .222, η_p^2 = .03.

Table 3.2

Mean Confidence in Correct Recognition Responses in Blocks 1 and 2

	Block 1			Block 2		
	Negative	Neutral	Positive	Negative	Neutral	Positive
Item Type						
Old	7.37 (.08)	7.26 (.10)	7.24 (.09)	7.43 (.07)	7.38 (.08)	7.34 (.08)
New	6.65 (.13)	6.52 (1.39)	6.43 (.13)	6.59 (.15)	6.36 (.15)	6.38 (.15)

Note: Standard errors are shown in parenthesis beside their respective means.

3.3 Metamnemonic Resolution

Finally, we examined metamnemonic resolution as described above. We analyzed metamnemonic d_a (see Table 3.1) in a 2 (block: block 1 vs. block 2) x 3 (emotion: negative vs. neutral vs. positive) repeated measures ANOVA. Findings showed there was a significant main effect of block, F(1, 45) = 17.54, MSE = 0.43, p < .001, η_p^2 = .28. Resolution was significantly higher in block 2 than in block 1. The main effect of emotion was significant, F(2, 90) = 25.50, MSE = 0.30, p < .001, η_p^2 = 0.36. Follow-up comparisons revealed that resolution was

significantly higher for negative images than for neutral images, $t(45) = 6.17$, $p < .001$, $d = 0.96$, but not when compared with positive images, $t(45) = 2.18$, $p = .086$, $d = 0.28$. Moreover, resolution for positive images was significantly higher than for neutral images, $t(45) = 5.03$, $p < .001$, $d = 0.68$. The interaction between block and emotion was also significant, $F(2, 90) = 3.95$, $MSE = 0.22$, $p = .023$, $\eta_p^2 = .08$. Post hoc comparisons showed that there was no difference in resolution for negative images across blocks, $t(45) = 1.26$, $p = .805$, $d = 0.25$. However, resolution for neutral images significantly increased across blocks, $t(45) = 4.04$, $p = .003$, $d = 0.78$. Resolution for positive images also significantly increased across blocks, $t(45) = 3.24$, $p = .025$, $d = 0.63$. This finding is particularly fascinating for multiple reasons. Hourihan (2020) did not observe any significant findings for resolution. However, they only reported gamma correlations rather than d_a. Hourihan and Bursey (2017), on the other hand, reported d_a and found no significant effects of emotion but had only utilized positive and neutral images.

Chapter 4: Discussion

The objective of the present study was to examine how metamemory for emotional images is influenced by multiple tests and specifically whether immediate test experience could improve the accuracy of JOLs on a subsequent study list. It was predicted that participants would report higher JOLs for emotional images than for neutral images in the first block. Moreover, we predicted that participants would report the highest JOLs for negative emotional images, followed by positive emotional images, then neutral images. Hourihan (2020) also found that participants' recognition accuracy displayed a pattern opposite to that of their JOLs: participants had worse recognition accuracy for negative images, followed by positive images, which, in turn,

were followed by neutral images. Thus, we predicted identical findings in the first block of the present study when participants had not yet had test experience.

We had theorized that this discrepancy between JOL and recognition accuracy was attributable to a lack of participant calibration. Participants did not appreciate the demands of an old/new recognition test because they had limited experience completing such tasks. Therefore, we predicted that participants would demonstrate a discrepancy between JOL and recognition accuracy in the first test but narrow said discrepancy between JOL and recognition accuracy with a second study-test phase. An overview of the findings will demonstrate that JOLs increased, recognition discriminability decreased, but surprisingly, resolution improved.

4.1 Summary of Results

Consistent with Hourihan (2020), participants in the present study reported higher JOLs for emotional content than for neutral content. Specifically, participants demonstrated a JOL pattern that is identical to Hourihan (2020)—JOLs were the highest for negative emotional images, followed by positive emotional images, which, in turn, were higher than neutral images. Despite the experience acquired from completing the first block, participants still demonstrated the same overall pattern of JOLs for emotional conditions and increased overall JOLs in block 2.

However, these increased JOLs were not reflected in memory performance measured by discriminability. Participants demonstrated high hit rates across blocks but more false alarms on the second block. Therefore, discriminability decreased across blocks, evidenced by lower discriminability on the second block. Discriminability was lower for positive and negative images relative to neutral images but not when compared to each other. However, performance on the recognition tests was generally higher than is ideal; in addition to a number of perfect hit rates, many were also near ceiling level. This resulted in a ceiling effect in which there was

limited variability as participants demonstrated generally high hit rates across emotional conditions requiring a fairly heavy application of corrections in order to compute *d'*. The false alarm rates across participants were often zero across emotional conditions to the extent that corrections were required to compute discriminability. As with hit rates, a lack of sufficient variability likely affected the overall pattern of memory performance. Therefore, these findings narrowly fail to replicate the findings of Hourihan (2020) as we do not observe the pattern of lowest discriminability for negative images followed by positive, which, in turn, are followed by neutral images. However, numerically, discriminability does replicate Hourihan (2020); participants demonstrated the lowest memory performance for negative images which was followed by positive images, which, in turn, were followed by neutral images.

Participants were more conservative than liberal in terms of response bias across blocks, as suggested by their positive *c* values. Thus, participants required a reasonably high feeling of familiarity in order to judge an image as studied. Participants were comparatively more liberal (but still conservative overall) in block 2 which was reflected in their higher false alarm rates; in block 2, participants were slightly more likely to incorrectly identify new items as old. As noted above, actual hit rates did not differ significantly across blocks. However, it should be noted that the potential to increase number of hits in the second block is significantly affected by participants reaching ceiling performance in the first block. Higher hit rates may have been observed if participants had not demonstrated such high hit rates in the first block. Nevertheless, this relatively liberal shift in response bias on the second block was not related to emotion condition, and thus response bias did not likely contribute to the effects of emotion observed in recognition discriminability.

Participants were more confident in their ability to recognize old items than they were for new items. Confidence in correctly rejecting new items did not change across blocks, but confidence in correctly recognizing old items increased in block 2. The pattern of emotional condition effects on confidence at test is no surprise, as it corresponds with that of participant JOLs across blocks—highest for negative, followed by positive and neutral. Participants predicted that negative images would be most memorable, and they were most confident in their recognition responses for negative images.

Across blocks, JOL resolution was higher for negative images than for positive images, which, in turn, was higher than for neutral images. The interaction between block and emotional condition was such that resolution significantly improved for both neutral and positive images, but not for negative images. However, JOL resolution was still highest for negative images across blocks. That is, at the item level, participants did well at predicting which specific images they would succeed or fail to recognize, and this was most pronounced for negative images. Therefore, participants were surprisingly best at predicting successful and unsuccessful recognition of individual negative images, despite the calibration inaccuracy. This interesting finding will be discussed further below. In summary, these findings generally demonstrate that participants were more confident but performed worse on the second block, with the exception of metamnemonic resolution.

These findings demonstrate that the pattern of participant JOLs did not represent the pattern of memory performance, even on the second block that followed immediate test experience. We may infer that participants were not better calibrated with an additional metamemory test, and in fact, we saw the opposite of our hypothesized narrower discrepancy between JOLs and recognition. Providing participants with additional test experience led them to

be more confident, but this had negative effects on subsequent memory performance. JOLs and recognition accuracy demonstrated opposite patterns: JOLs increased and recognition accuracy decreased across blocks. Here, we discuss possible explanations and future directions to take following these results.

4.2 Emotion and Metamemory

Participants demonstrating higher JOLs for emotional images, and particularly negative images, is consistent with the previous research that participants report higher JOLs for emotional content, including words (e.g., Hourihan et al., 2017; Tauber et al., 2017; Zimmerman & Kelley, 2010), faces (Nomi et al., 2013; Witherby & Tauber, 2018) and images (Hourihan, 2020; Hourihan & Bursey, 2017). The aforementioned works that discuss the influences of valence, arousal, physiological responses and distinctiveness on their findings demonstrate the ways in which these four variables intertwine to elicit higher JOLs for emotional content.

For instance, Hourihan et al. (2017) conducted a three-experiment study to examine why and how emotional content influences participant JOLs for emotional words. In Experiment 1, word lists consisted of neutral valence words, half low-arousal and half high-arousal; Experiment 2 used neutral-arousal words, half negative and half neutral valence. As expected, participant JOLs were higher for emotional words in the first two experiments. Participants provided higher JOLs for emotional words even if they differed from neutral words only in valence and not arousal. Interestingly, this finding suggests that metamemory effects are not solely reliant on physiological emotional arousal, but instead may reflect intentional use of beliefs about how emotion should influence memory, when emotional factors are salient (i.e., in a mixed list that contains two distinct categories of words). The word lists in the third experiment were curated to minimize the distinctiveness associated with valence and arousal by continuously varying their

levels. JOLs were not significantly influenced by valence and/or arousal (only word frequency consistently influenced JOLs); the effects of emotion on metamemory and memory may reflect cognitive factors, such as the relative distinctiveness of stimuli, rather than physiological factors associated with emotion.

Hourihan and Bursey (2017) examined how the emotional content of images influences metamemory by having participants provide JOLs for positive and neutral images and complete a recognition test. Participants provided higher JOLs for positive images than for neutral images across three experiments, which is consistent with the notion that participants rely on intrinsic characteristics of the item in question. Emotional content has been shown to be a highly salient cue: A physiological response is elicited in participants when viewing emotional images which significantly differs from that elicited when viewing neutral images (Bradley et al., 1992; Pérez-Mata et al., 2012). Hourihan and Bursey (2017) inferred that participants likely experienced a subjective feeling of happiness when viewing positive images which increased their perceived likelihood in which they thought they would remember those images. Theoretically, participants in the present study may have experienced a similar feeling when viewing positive images across blocks. Moreover, participants likely experienced feelings of sadness, threat or fear when viewing negative images (Kensinger, 2009), which may have led to an increase in their predicted likelihood of remembering the respective images. We replicated the general JOL finding by Hourihan (2020) that participants provided significantly higher JOLs for emotional images than for neutral images, and similarly, both the work of Hourihan (2020) and the present study curated image lists where emotional images were significantly higher in arousal and significantly different in valence from neutral images. Participants likely accounted for these differences in arousal responses when making JOLs. Higher levels of arousal for emotional images likely

elicited a physiological response that increased JOLs for emotional images. On the other hand, the absence of a salient physiological response to a seemingly benign or innocuous neutral image may have led participants to provide relatively lower JOLs.

The fact that participants demonstrated higher JOLs in the second block suggests that JOLs, and general metamemory, for emotional images, are sensitive to test experience. Moreover, participants demonstrating higher JOLs with more experience is interestingly contradictory to the UWP, which, again, as Koriat et al. (2002) stated occurs when participants shift their overconfidence to underconfidence in subsequent learning phases. JOLs overestimated memory test performance across blocks of the present study, displaying a pattern of confidence that did not coincide with the pattern of accuracy. Thus, the calibration bias that is accrued from the UWP was not observed in the present study despite participants receiving additional test experience. However, this finding may only be applicable when items are the same, and participants can potentially recall prior test outcomes for specific items when making JOLs for a second time (e.g., Finn & Metcalfe, 2007), rather than generalizing to new items of the same general category.

Although calibration was relatively poor in both blocks, the resolution findings in the present study demonstrate an interesting pattern. Although participants' general belief that negative images would be more memorable and consequently more likely to be recognized than positive and neutral images was inaccurate on average, the findings demonstrate that, at the item level, participants were in fact best at predicting which individual negative images would and would not be recognized. That is, for negative images, participants were best at accurately assigning higher JOLs to the specific negative images they would later recognize and lower JOLs to the specific negative images they would later not recognize. Interestingly, resolution improved

for each emotional condition across blocks; improvement in resolution was greater for positive and neutral images, but overall resolution for negative images was highest in both blocks.

It must be acknowledged that metamnemonic resolution only accounts for the correspondence between JOLs for studied items and their subsequent test accuracy, whereas overall recognition discriminability includes responses both to studied items and to novel images presented at test. This correspondence in the present study suggests the highest resolution for negative images. It is possible that JOLs are sensitive to the individual characteristics of images that are likely to lead to successful or failed recognition, but of course JOLs cannot account for future performance on test trials with new items, which influences recognition discriminability. Individual image characteristics, such as valence and arousal have already been addressed, but correlations between both of these two characteristics and the magnitude of JOLs may provide further insight into why this component of metamemory for negative emotional images is more accurate than the pattern of mean JOLs compared to the pattern of mean recognition discriminability.

We conducted post-experiment data analysis to analyze the relationships between valence and JOLs, and between arousal and JOLs, using item-level correlations for each person across the three emotion categories (view Appendix C). We observed a significant negative mean correlation between valence and JOLs for negative images in both blocks (i.e., the more strongly negative an image was, the higher the JOL), which were additionally the largest magnitude correlations for each respective emotional category. Correlations between valence and JOLs for neutral images were significantly different from zero for block 1, but not for block 2. Interestingly, correlations between valence and JOLs were not significantly different than zero for positive items across blocks. This finding serves as evidence of a large difference between the

two emotional categories (i.e., positive compared to negative) in terms of the degree to which valence is used as a cue when making JOLs. The correlations between valence and JOLs for negative images were larger than for positive and neutral images, which, in turn, did not differ from each other.

We observed the highest correlation between arousal and JOLs for negative images, followed by neutral then positive images. The finding of an average negative correlation between arousal and JOL for positive images is interesting, as it shows that more calming positive images elicited higher JOLs than exciting ones. On the other hand, negative images that were highly arousing elicited higher JOLs, as evidenced by the positive correlation between arousal and JOLs for negative images. There was a greater reliance on the intrinsic cue of arousal for negative images than for positive images.

The fact that both JOL resolution and correlational analysis was highest for negative images suggests that the intrinsic cues of valence and arousal are weighed more heavily when making JOLs to negative images when compared to neutral or positive images. This appears to lead to more accurate item-specific predictions. Moreover, there was a tendency for the correlations to be reduced in block 2 relative to block 1, which may indicate a shift to increased reliance on mnemonic cues after the first study-test block.

We also conducted post-experiment data analysis to analyze the relationship between emotion condition and JOL reaction time (RT). We observed that participants showed slower JOL RTs for negative images than positive and neutral images in both blocks (see Appendix D). Additionally, JOL RTs were faster in block 2 than in block 1 across each emotion condition, a finding which is expected as participants had immediate practice by completing the first study-test block. Thus, although negative images had the worst recognition, they had the slowest JOL

RT, highest JOLs, and best resolution. These findings together are clear evidence that opposes the notion that participants spent less time thinking about negative images at the encoding stage of the study.

4.3 Emotion and Memory

The fact that memory performance was the worst for negative images is surprising. Aside from emotional images being associated with higher predicted future memory, behavioural evidence suggests that emotional arousal leads to improved memory, and said benefits are particularly more pronounced for negative emotional stimuli (Kensinger, 2009). One possible explanation is that negative valence may be correlated with JOLs for images, but not necessarily with memory due to the nature of the negative images selected for experimentation. For instance, although the pool of negative images selected from the IAPS database are significantly lower in valence and higher in arousal than the selected pool of neutral items, there is still variability in valence and arousal within an emotional condition, as well as the specific negative emotion depicted. A negative image may depict various emotional states such as anger (e.g., an angry face), sadness (e.g., an emaciated baby), gore (e.g., a decapitated man), disgust (e.g., vomit), death (e.g., an individual who committed suicide by hanging), threat (e.g., an individual pointing a gun), hostility (e.g., a man with an angry face and posing in a threatening posture) or elicit fear from a phobia (e.g., a spider). Moreover, multiple states may be combined into a single negative image. Therefore, participants may potentially have demonstrated better memory performance for negative images depicting particular states, or due to interactions among multiple states.

As previously stated, participants demonstrated undesirably high recognition accuracy across high hit rates and low false alarm rates. Many participants approached near-perfect recognition accuracy which resulted in high scores of discriminability. Ceiling effects may

partially explain why Hourihan's (2020) pattern of worst recognition for negative emotional images, followed by positive emotional images which, in turn, are followed by neutral images, was not fully replicated. The prevalence of ceiling effects in the present study is likely a result of two factors: image duration and the superiority of memory for images. As previously stated, each image in the present study appeared for 500ms, followed by a 250ms blank screen before participants provided a self-paced JOL response before proceeding to the next image. Past research has consistently found that human memory performance for images is remarkably accurate (e.g., Shepherd, 1967) and thus, it is possible that 500ms provided participants with excessive time to encode images and consequently their memory was marginally challenged. Perhaps a follow-up study with a shorter duration in which images are presented across phases and a mask between images would prevent such effects from inflated memory performance

The type of memory test utilized in the present study must also be considered. We used an old/new recognition paradigm (as in Hourihan, 2020) which differs from other studies that demonstrated enhanced memory for emotional images using free recall rather than recognition. Bradley et al. (1992), for instance, implemented a free recall task to measure pictorial memory performance in two experiments and found that memory for high-arousal images was higher than for low-arousal images when measured with immediate free recall. This enhanced memory for high-arousal images also lasted into long-term memory; participants demonstrated the same pattern when asked to complete free recall following a 30-minute retention interval, and even when tested one year after encoding.

Charles et al. (2003) examined age-related differences in free recall and recognition accuracy for positive, negative and neutral images in two-fold study where both recognition and recall were implemented as test measures. As free recall demands more self-directed processing

that is more likely to be influenced by present goals and motivation, it was hypothesized that there would be greater age differences on free recall than recognition. They found that older and middle-aged adults recalled more positive images than negative images, whereas younger adults demonstrated insignificant differences between the two respective emotional conditions. Conversely, younger participants demonstrated greater recognition for negative images than for positive and neutral images whereas middle-aged and older participants demonstrated no differences in recognition regardless of image category. When controlling for attentional processes in Experiment 2, younger adults demonstrated the highest memory performance for negative images on both free recall and recognition tasks, followed by positive then neutral. Older adults, on the other hand, demonstrated no significant difference in recall for positive and negative images, both of which were significantly higher than for neutral images.

Tauber et al. (2017) evaluated young and older adults' monitoring (using JOLs) and memory for positive emotional and neutral images and implemented both free recall and recognition as measures of memory performance. As expected, participants demonstrated higher JOLs for positive emotional images. Interestingly, free recall performance demonstrated enhanced memory for positive images; thus, JOLs accurately predicted performance. However, when recognition was implemented as the test measure, participants demonstrated a similar pattern as the present study, Hourihan (2020), and Hourihan and Bursey (2017): worse recognition discriminability for positive images than for neutral images in Experiment 1; Experiment 2 demonstrated no effects of emotion on recognition. A partial explanation of the inaccuracies observed in the present study may be that participations are approaching the recognition test from the mentality of free recall when these two tests are different and have been shown to yield varied results across different studies.

4.4 Experience and Metamemory

As previously stated, it was hypothesized that an additional test would better calibrate participants and result in a smaller discrepancy between JOLs and recognition accuracy. However, the findings revealed the opposite: participants were more confident but performed worse on the second block. We theorized that one of the reasons for the initial disparities between JOL and recognition accuracy was that participants did not appreciate the demands of a recognition test. Thus, they could better calibrate themselves on a second test, after experiencing a first test. These findings indicate that calibration was not improved, and rather, it was reduced. Participants, therefore, may require additional tests to accrue more experience in a metamemory study to better calibrate themselves. Perhaps a third test in an experiment with three study and test phases would improve calibration. Moreover, the fact that recognition worsened across blocks may suggest that multiple tests may lead to increased confidence (as measured by JOLs and recognition confidence) but reduced recognition accuracy.

As previously discussed, Benjamin (2003) examined the impact of predicting and postdicting metacognitive judgements on how word frequency affects recognition memory. In three experiments, participants were informed that they would need to study a set of words and predict the likelihood in which they would remember each word in a subsequent memory test. Participants predicted, in the form of JOLs, better performance for common words but postdicted superior performance for uncommon words. The group who made a postdiction during their first study-test cycle made the largest gains in calibration on their second study-test cycle. It is suggested that participants rely on cues when making metacognitive judgements while studying items that are different from when they are making judgements in a recognition test (Benjamin, 2003; Glanzer & Bowles, 1976; Gorman, 1961). Such methodology could have been applied to

the present study. Had participants been asked to make postdictions rather than, or in addition to, confidence judgements, they may have learned more during recognition about the challenges in accurately recognizing (rather than recalling) negative images.

Participants demonstrating a significant increase in resolution for neutral and positive images is fascinating. Although participants learned nothing that would facilitate calibration across blocks, they apparently learned to utilize item-specific cues to improve resolution overall. This increase was most pronounced in neutral and positive images, which were substantially lower in block 1 than in block 2. Participants may have unconsciously learned how to utilize item-specific cues, which is a possibility that could be explored in future studies. Perhaps a future study could develop a post-experiment questionnaire where participants describe their strategies for providing recognition and confidence responses relative to emotional condition.

4.5 Limitations and Future Directions

As previously stated, participation in the present study was conducted both online and in person. Although participants who completed the study in person remained in a quiet and uninterrupted environment (i.e., a computer lab room) with a researcher present to respond to questions or address any technical difficulties, it is unknown whether online participants shared such control. It is possible that participants who completed the study online were negatively affected by noise and various potential interferences which may have impacted the attention being provided to the task at hand.

Moreover, due to the nature of available participants for data collection, the present study has an uneven number of online participants compared to those who completed the study in person. Thirty-three participants completed the study online, while only 13 participants completed the study in person. Consequently, we are unable to compare averages between

groups and fail to observe any potential data analysis that would provide more insight into the present study's main findings.

There was also no control for ensuring that variability of the specific contents and details of the images utilized in this experiment was equated across the emotion conditions. However, such a lack of control was a necessity in order to obtain the desired 456 images from three emotional conditions. To obtain a sufficient number of images for two study-test blocks, it was not feasible to obtain equal numbers of categories or category sizes across each emotion condition. Additionally, due to the large numbers of images required from the IAPS, it was not possible to control for numbers of semantic categories across emotion conditions. Hourihan and Bursey (2017) explored differences in image category size as a possible explanation for why recognition for neutral images was superior to that for positive images in Experiment 1. Even when their image sets were more closely matched between emotion conditions in terms of numbers and sizes of categories, the finding of no benefit for recognizing emotional pictures remained constant. Additionally, Hourihan (2020) curated a stimulus set with images from IAPS with the number and size of categories more closely controlled between emotion conditions, and still found that JOLs were highest for negative images but recognition was highest for neutral images. Thus, control for specific details of images in the present study was likely unnecessary.

As described above, it is apparent that experiencing one old/new recognition test was an insufficient amount of experience for participants to better calibrate themselves and adjust their JOLs to match their memory performance in a subsequent block. It would be interesting to conduct an experiment that is identical to the present study but with a third block to examine how JOLs change with a third study and test phase. With an additional block to the present study, participants may finally acquire an adequate amount of study and test experience to better

calibrate themselves; they would correct their JOLs to better match recognition accuracy. If this were the case, then we may be a step closer to discovering a general level of test experience required to benefit participants to the extent to which they can accurately predict the pattern of recognition when making JOLs. However, the main findings of the present study clearly suggest that additional test experience, alone, is insufficient as a manipulation to improve calibration and narrow the discrepancy between JOL and recognition. Therefore, alternative components of the experiment must be manipulated for further examination. For instance, we may consider whether delayed JOLs (Nelson & Dunlosky, 1991) for emotional images may be more accurately calibrated than the immediate JOLs used here.

In the present study, participants had more false alarms and higher levels of confidence in the second block than for the first block. The findings of Benjamin (2003) are applicable to the present study and suggest that participants may have assessed and consequently made better recognition judgements had they received an opportunity to make postdictions at test. For instance, on block 1, participants would have postdicted the likelihood in which they think they would have remembered an item they deemed "new" at test, had it been studied. This postdiction may then lead to reduced JOLs on a future study phase or shift to a more conservative response bias on a future test, which would improve overall memory performance. Further experiments with multiple metamemory tests are required to generalize the current finding to additional forms of emotional stimuli such as emotional words and emotional faces.

As previously stated, the UWP effect has been shown to be robust in conditions where participants receive or do not receive feedback (e.g. Koriat, 1997; Koriat, 2002). However, the materials utilized in these studies were paired associates, action phrases, and general knowledge questions. Additionally, metamemory accuracy was measured with recall; confidence-recall

research has demonstrated that metamemory improvement may increase with item-by-item accuracy feedback (e.g., Thompson, 1998). To our knowledge, no study has examined the effect of item-by-item feedback on metamemory, measured by old/new recognition, for emotional images.

4.6 Conclusions

We built upon the work of Hourihan (2020), which found that participants' predicted memory performance for emotional images did not correspond with their actual recognition memory performance. Participants predicted the highest recognition for negative images which were recognized worse than positive and neutral images. The present study was predicated on the theory that this discrepancy between predicted memory performance and recognition memory was largely due participants' lack of experience in recognition tests. Thus, we predicted that a second opportunity to complete the same task with new stimuli would narrow the discrepancy between predicted and actual memory performance. As with Hourihan (2020), participants demonstrated the highest predicted performance for negative images, but recognition memory demonstrated the opposite—recognition memory was worst for negative images in both cycles. Participants apparently received no calibration benefit from immediate test experience. However, participants demonstrated the best resolution for negative images. Thus, although participants demonstrated the poorest recognition memory for negative images, they were best at predicting which specific negative images would and would not be recognized. Moreover, participants were significantly better at predicting which positive and neutral images would and would not be recognized in block 2 than they were in block 1; we may infer that, although the discrepancy between JOLs and recognition did not shorten with immediate study and test experience, participants may have learned and utilized a cue that guided their ability to discriminate between

individual images across categories. The immediate study and test experience from block 1 appeared to have aided participants in predicting which specific images would and would not be recognized. Participants may have learned a cue (whether consciously or unconsciously) that impacted their perceived level of difficulty in remembering particular items, which may have improved overall resolution. Participants may have implicitly learned how to relate mnemonic cues during the encoding of emotional images to their subsequent recognition ability. However, they would be unlikely to be able to explicitly state why this is the case, or exactly what information they are using to make a JOL. Such a finding has implications for the study of how we consciously and unconsciously attend to specific intrinsic and mnemonic cues when making JOLs for emotional images.

References

Benjamin, A. S. (2003). Predicting and postdicting the effects of word frequency on memory. *Memory & Cognition, 31*(2), 297–305. doi.org/10.3758/BF03194388

Benjamin, A. S., & Bjork, R. A. (Ed.) (1996). *Retrieval fluency as a metacognitive index.* Psychology Press.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology. General, 127*(1), 55–68. doi.org/10.1037/0096-3445.127.1.55

Benjamin, A.S., & Diaz, M. (2008). Measurement of relative metamnemonic accuracy. In J. Dunlosky & R.A. Bjork (Eds). *Handbook of metamemory and memory* (pp. 73-94). Psychology Press.

Bradley, M. M., Codispoti, M., Cuthbert, B. N., & Lang, P. J. (2001). Emotion and motivation i: Defensive and appetitive reactions in picture processing. *Emotion (Washington, D.C.), 1(3*), 276–298. doi.org/10.1037//1528-3542.1.3.276

Bradley, M. M., Greenwald, M. K., Petry, M. C., & Lang, P. J. (1992). Remembering pictures: Pleasure and arousal in memory. Journal of Experimental Psychology. *Learning, Memory, and Cognition, 18(*2), 379–390. doi.org/10.1037/0278-7393.18.2.379

Charles, S. T., Mather, M., & Carstensen, L. L. (2003). Aging and emotional memory: The forgettable nature of negative images for older adults. *Journal of Experimental Psychology. General, 132*(2), 310–324. doi.org/10.1037/0096-3445.132.2.310

Deason, R. G., Tat, M. J., Flannery, S., Mithal, P. S., Hussey, E. P., Crehan, E. T., Ally, B. A., &

Budson, A. E. (2017). Response bias and response monitoring: Evidence from healthy older adults and patients with mild Alzheimer's disease. *Brain and Cognition, 119*, 17–24. doi.org/10.1016/j.bandc.2017.09.002

Dunlosky, J., & Bjork, R. A. (Eds.). (2008). *Handbook of metamemory and memory.* Psychology Press.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Sage Publications, Inc.

Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 33*(1), 238–244. doi.org/10.1037/0278-7393.33.1.238

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 16*(1), 5–16. doi.org/10.1037/0278-7393.16.1.5

Glanzer, M., & Bowles, N. (1976). Analysis of the word-frequency effect in recognition memory. Journal of Experimental Psychology: *Human Learning and Memory, 2*(1), 21–31. doi.org/10.1037/0278-7393.2.1.21

Gorman, A. M. (1961). Recognition memory for nouns as a function of abstractness and frequency. *Journal of Experimental Psychology, 61*(1), 23–29. doi.org/10.1037/h0040561

Hourihan, K. L. (2020). Misleading emotions: judgments of learning overestimate recognition of negative and positive emotional images. *Cognition and Emotion, 34*(4), 771–782. doi.org/10.1080/02699931.2019.1682972

Hourihan, K. L., & Bursey, E. (2017). A misleading feeling of happiness: metamemory for positive emotional and neutral pictures. *Memory, 25*(1), 35–43. doi.org/10.1080/09658211.2015.1122809

Hourihan, K. L., Fraundorf, S. H., & Benjamin, A. S. (2017). The influences of valence and arousal on judgments of learning and on recall. *Memory & Cognition, 45*(1), 121–136. doi.org/10.3758/s13421-016-0646-3

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language, 32*(1), 1–24. doi.org/10.1006/jmla.1993.1001

Kensinger, E. A. (2009). Remembering the details: Effects of emotion. *Emotion Review, 1*(2), 9 9–113. doi.org/10.1177/1754073908100432

Kensinger, E. A., & Corkin, S. (2003). Memory enhancement for emotional words: Are emotional words more vividly remembered than neutral words? *Memory & Cognition, 31*(8), 1169–1180. doi.org/10.3758/BF03195800

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of knowing: The influence of retrieval practice. *The American Journal of Psychology, 93*(2), 329–343. doi.org/10.2307/1422236

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology. General, 126*(4), 349–370. doi.org/10.1037/0096-3445.126.4.349

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 31*(2), 187–194. doi.org/10.1037/0278-7393.31.2.187

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology. General, 131*(2), 147–162.

doi.org/10.1037/0096-3445.131.2.147

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review, 1*(4), 279–308. doi.org/10.1007/BF01320096

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), Gainesville, FL: University of Florida.

Lovelace, E. A. (1984). Metamemory: Monitoring future recallability during study. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 10*(4), 756–766. doi.org/10.1037/0278-7393.10.4.756

Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the goodman–kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(2), 509–527. doi.org/10.1037/a0014876

Mazzoni, G., Cornoldi, C., & Marchitelli, G. (1990). Do memorability ratings affect study-time allocation? *Memory & Cognition, 18*(2), 196–204. doi.org/10.3758/BF03197095

Metcalfe, J., Finn, B. Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review, 15*(1), 174–179 (2008). doi.org/10.3758/PBR.15.1.174

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (jols) are extremely accurate at predicting subsequent recall: The "delayed-jol effect." *Psychological Science, 2*(4), 267–270. doi.org/10.1111/j.1467-9280.1991.tb00147.x

Nomi, J. S., Rhodes, M. G., & Cleary, A. M. (2013). Emotional facial expressions differentially influence predictions and performance for face recognition. *Cognition and Emotion, 27*(1), 141–149. doi.org/10.1080/02699931.2012.679917

Pérez-Mata, N., López-Martín, S., Albert, J., Carretié, L., & Tapia, M. (2012). Recognition of emotional pictures: Behavioural and electrophysiological measures. *Journal of Cognitive Psychology, 24*(3), 256-277. doi.org/10.1080/20445911.2011.613819

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51*(1), 195–203. doi.org/10.3758/s13428-018-01193-y

Rabinowitz, J. C., Ackerman, B. P., Craik, F. I. M., & Hinchley, J. L. (1982). Aging and metamemory: The roles of relatedness and imagery. *Journal of Gerontology (Kirkwood), 37*(6), 688–695. doi.org/10.1093/geronj/37.6.688

Schulman, A. I. (1967). Word length and rarity in recognition memory. *Psychonomic Science, 9(*4), 211–212. doi.org/10.3758/BF03330834

Serra, M. J., & Ariel, R. (2014). People use the memory for past-test heuristic as an explicit cue for judgments of learning. *Memory & Cognition, 42*(8), 1260–1272. doi.org/10.3758/s13421-014-0431-0

Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning & Verbal Behavior, 6*(1), 156–163. doi.org/10.1016/S0022-5371(67)80067-7

Swets, J. A. (1986). Form of empirical rocs in discrimination and diagnostic tasks: implications for theory and measurement of performance. *Psychological Bulletin, 99*(2), 181–198. doi.org/10.1037/0033-2909.99.2.181

Tauber, S. K., & Dunlosky, J. (2012). Can older adults accurately judge their learning of emotional information? *Psychology and Aging, 27*(4), 924–933. doi.org/10.1037/a0028447

Tauber, S. K., Dunlosky, J., Urry, H. L., & Opitz, P. C. (2017). The effects of emotion on younger and older adults' monitoring of learning. *Aging, Neuropsychology, and Cognition, 24*(5), 555-574. doi.org/10.1080/13825585.2016.1227423

Thompson, W. B. (1998). Metamemory Accuracy: Effects of feedback and the stability of individual differences. *The American Journal of Psychology, 111*(1), 33–42. doi.org/10.2307/1423535

Vesonder, G. T., & Voss, J. F. (1985). On the ability to predict one's own responses while learning. *Journal of Memory and Language, 24*(3), 363–376. doi.org/10.1016/0749-596X(85)90034-8

Witherby, A. E., & Tauber, S. K. (2018). Monitoring of learning for emotional faces: how do fine-grained categories of emotion influence participants' judgments of learning and beliefs about memory? *Cognition and Emotion, 32*(4), 860–866. doi.org/10.1080/02699931.2017.1360252

Yegiyan, N. S., & Yonelinas, A. P. (2011). Encoding details: positive emotion leads to memory broadening. *Cognition and Emotion, 25*(7), 1255–1262. doi.org/10.1080/02699931.2010.540821

Zimmerman, C. A., & Kelley, C. M. (2010). "I'll remember this!" effects of emotionality on memory predictions versus memory performance. *Journal of Memory and Language, 62*(3), 240–253. doi.org/10.1016/j.jml.2009.11.004

Appendix A: IAPS Image Numbers for Stimulus Pool

Negative			
1525	3140	9043	9426
2053	3160	9075	9427
2095	3168	9140	9428
2141	3180	9180	9429
2205	3181	9181	9430
2276	3185	9183	9432
2301	3191	9184	9433
2345.1	3215	9185	9435
2375.1	3220	9187	9470
2456	3225	9220	9491
2688	3230	9253	9500
2691	3261	9254	9520
2703	3300	9265	9530
2710	3301	9280	9560
2717	3350	9290	9561
2750	3550	9291	9570
2751	6021	9295	9571
2799	6212	9300	9590
2800	6213	9301	9610
2900	6242	9302	9611
2981	6243	9320	9623
3001	6244	9322	9630
3005.1	6311	9325	9800
3015	6570.1	9326	9830
3016	6571	9330	9831
3017	6825	9332	9832
3019	6831	9340	9900
3051	6834	9342	9901
3059	6838	9400	9902
3061	7359	9405	9903
3062	7380	9412	9905
3063	8230	9415	9909
3064	9000	9419	9911
3068	9006	9420	9920
3100	9007	9421	9922
3101	9031	9423	9925
3110	9040	9424	9927
3131	9041	9425	9941
Neutral			
1333	2516	7004	7140
1670	2518	7006	7150
2002	2570	7009	7160
2020	2580	7010	7161

2026	2593	7012	7165
2036	2595	7014	7170
2038	2620	7016	7175
2104	2720	7017	7179
2190	2745.1	7019	7180
2191	2830	7020	7184
2200	2840	7025	7185
2210	2850	7026	7186
2214	2870	7030	7187
2215	2880	7031	7192
2221	2890	7032	7205
2235	2980	7034	7207
2273	4500	7035	7217
2305	4571	7036	7224
2357	5120	7037	7233
2377	5130	7038	7235
2383	5250	7039	7255
2384	5390	7040	7287
2385	5410	7041	7300
2390	5471	7043	7490
2393	5500	7045	7491
2396	5510	7050	7493
2397	5520	7052	7500
2411	5530	7053	7509
2440	5533	7055	7513
2441	5534	7056	7547
2480	5731	7059	7705
2493	5740	7060	7710
2495	5875	7061	7950
2499	6150	7062	8311
2506	7000	7080	9070
2512	7001	7100	9210
2513	7002	7110	9260
2514	7003	7130	9700
Positive			
1340	2398	5830	8080
1440	2550	5833	8090
1463	2655	5849	8116
1590	4250	5890	8120
1650	4597	5910	8130
1659	4599	5994	8161
1710	4601	7200	8162
1720	4603	7220	8163
1722	4609	7230	8170
1731	4610	7250	8180
1811	4612	7260	8185

1999	4614	7270	8186
2040	4623	7282	8190
2045	4624	7330	8193
2050	4626	7350	8200
2058	4628	7390	8208
2071	4640	7400	8210
2075	4641	7405	8300
2080	5199	7410	8340
2150	5210	7430	8350
2155	5215	7451	8370
2158	5260	7460	8371
2160	5270	7470	8380
2165	5450	7492	8400
2208	5460	7501	8420
2209	5470	7502	8461
2216	5480	7508	8470
2224	5600	7570	8490
2300	5621	7580	8492
2303	5623	7650	8496
2340	5626	7660	8499
2345	5629	8021	8500
2346	5660	8030	8501
2347	5700	8031	8502
2352	5814	8033	8503
2362	5820	8034	8510
2389	5825	8040	8531
2391	5829	8041	8540

Appendix B: Approval from Research Ethics Board

MEMORIAL UNIVERSITY

Interdisciplinary Committee on
Ethics in Human Research (ICEHR)

St. John's, NL Canada A1C 5S7
Tel: 709 864-2561 icehr@mun.ca
www.mun.ca/research/ethics/humans/icehr

ICEHR Number:	**20230604-SC**
Approval Period:	September 15, 2022 – September 30, 2023
Funding Source:	
Responsible Faculty:	Dr. Kathleen Hourihan Department of Psychology
Title of Project:	*Metamemory for Emotional Images*

September 15, 2022

Mr. Chavon Gonsalves
Department of Psychology, Faculty of Science
Memorial University

Dear Mr. Gonsalves:

Thank you for your correspondence addressing the issues raised by the Interdisciplinary Committee on Ethics in Human Research (ICEHR) for the above-named research project. ICEHR has re-examined the proposal with the clarifications and revisions submitted, and is satisfied that the concerns raised by the Committee have been adequately addressed. However, the word "necessarily" must be deleted from the statement "participation is not necessarily anonymous as you are completing this experiment in a lab" in the confidentiality and anonymity section of the PREP consent form.

In accordance with the *Tri-Council Policy Statement on Ethical Conduct for Research Involving Humans (TCPS2)*, the project has been granted ***full ethics clearance*** for **one year**. ICEHR approval applies to the ethical acceptability of the research, as per Article 6.3 of the *TCPS2*. Researchers are responsible for adherence to any other relevant University policies and/or funded or non-funded agreements that may be associated with the project. If funding is obtained subsequent to ethics approval, you must submit a Funding and/or Partner Change Request to ICEHR so that this ethics clearance can be linked to your award.

The *TCPS2* **requires that you strictly adhere to the protocol and documents as last reviewed** by ICEHR. If you need to make additions and/or modifications, you must submit an Amendment Request with a description of these changes, for the Committee's review of potential ethical concerns, before they may be implemented. Submit a Personnel Change Form to add or remove project team members and/or research staff. Also, to inform ICEHR of any unanticipated occurrences, an Adverse Event Report must be submitted with an indication of how the unexpected event may affect the continuation of the project.

The *TCPS2* **requires** that you submit an Annual Update to ICEHR before **September 30, 2023**. If you plan to continue the project, you need to request renewal of your ethics clearance and include a brief summary on the progress of your research. When the project no longer involves contact with human participants, is completed and/or terminated, you are required to provide an annual update with a brief final summary and your file will be closed. All post-approval ICEHR event forms noted above must be submitted by selecting the ***Applications: Post-Review*** link on your Researcher Portal homepage. We wish you success with your research.

Yours sincerely,

James Drover, Ph.D.
Vice-Chair, ICEHR

JD/bc

cc: Supervisor – Dr. Kathleen Hourihan, Department of Psychology

MEMORIAL UNIVERSITY

Interdisciplinary Committee on
Ethics in Human Research (ICEHR)

St. John's, NL Canada A1C 5S7
Tel: 709 864-2561 icehr@mun.ca
www.mun.ca/research/ethics/humans/icehr

ICEHR Number:	**20230604-SC**
Approval Period:	September 15, 2022 – September 30, 2023
Funding Source:	
Responsible Faculty:	Dr. Kathleen Hourihan Department of Psychology
Title of Project:	*Metamemory for Emotional Images*
Amendment #:	***01***

January 9, 2023

Mr. Chavon Gonsalves
Department of Psychology, Faculty of Science
Memorial University

Dear Mr. Gonsalves:

The Interdisciplinary Committee on Ethics in Human Research (ICEHR) has reviewed the proposed revisions for the above referenced project, as outlined in your amendment request dated January 6, 2023. We are pleased to give approval to the proposed online protocols and revised in-person protocols, as described in your request, provided all other previously approved protocols are followed.

The *TCPS2* **requires that you strictly adhere to the protocol and documents as last reviewed** by ICEHR. If you need to make any other additions and/or modifications during the conduct of the research, you must submit an Amendment Request with a description of these changes, for the Committee's review of potential ethical issues, before they may be implemented. Submit a Personnel Change Form to add or remove project team members and/or research staff. Also, to inform ICEHR of any unanticipated occurrences, an Adverse Event Report must be submitted with an indication of how the unexpected event may affect the continuation of the project.

Your ethics clearance for this project expires **September 30, 2023**, before which time you must submit an Annual Update to ICEHR, as required by the *TCPS2*. If you plan to continue the project, you need to request renewal of your ethics clearance, and include a brief summary on the progress of your research. When the project no longer requires contact with human participants, is completed and/or terminated, you need to provide an annual update with a brief final summary, and your file will be closed.

All post-approval ICEHR event forms noted above must be submitted by selecting the *Applications: Post-Review* link on your Researcher Portal homepage.

The Committee would like to thank you for the update on your proposal and we wish you well with your research.

Yours sincerely,

James Drover, Ph.D.
Vice-Chair, Interdisciplinary Committee on
Ethics in Human Research

JD/bc

cc: Supervisor – Dr. Kathleen Hourihan, Department of Psychology

Appendix C: Correlational Analysis

Using the values from the IAPS database (Lang et al., 2008) for mean valence, correlations between valence and JOLs were computed on a per-participant basis and were compared across each emotional condition and the two blocks. The means of these correlations were analyzed in one-sample t-tests, comparing them to zero (see Table C.1). Findings revealed significant correlations between valence and JOLs for negative images in block 1, $t(45) = 11.34$, $p < .001$, $d = 1.67$, and block 2, $t(45) = 8.72$, $p < .001$, $d = 1.29$. These correlations are negative and indicated that higher JOLs tended to be associated with more extremely negative valence. Valence and JOL correlations for neutral images in block 1 were significant, $t(45) = 2.63$, $p = .012$, $d = 0.39$, but not in block 2, $t(45) = 1.85$, $p = .071$, $d = 0.27$. These correlations are very small and only slightly positive. Participants tended to give higher JOLs to slightly more positive items but only in the first block. Lastly, correlations between valence and JOL for positive images were not significant in block 1, $t(45) = 1.54$, $p = .131$, $d = 0.23$, nor block 2, $t(45) = 1.05$, $p = .298$, $d = 0.16$. Therefore, there is no systemic relationship between the extent to which an image was positive and the magnitude of the corresponding JOL.

Correlations between arousal and JOLs were computed in a similar manner (see Table C.1), and were analyzed in one-sample t-tests, again comparing the mean correlations to zero. Findings revealed that the correlations between arousal and JOLs for negative images in block 1 were significant, $t(45) = 9.09$, $p < .001$, $d = 1.34$, as well as for negative images in block 2, $t(45) = 7.48$, $p < .001$, $d = 1.10$. These correlations were positive, indicating that more arousing images tended to receive higher JOLs. Correlations between arousal and JOLs for neutral images in block 1 were significant, $t(45) = 3.05$, $p = .004$, $d = 0.45$, contrasting the insignificant correlations between arousal and JOL for neutral images in block 2, $t(45) = 1.45$, $p = .153$, $d =$

0.21. Correlations were slightly positive across blocks; more arousing images tended to elicit higher JOLs. The correlation between arousal and JOLs for positive images was not significant for block 1, $t(45) = 1.66$, $p = .104$, $d = 0.25$. However, this correlation was significant for block 2, $t(45) = 2.47$, $p = .018$, $d = 0.36$. The correlation in block 2 was slightly negative, such that less exciting images tended to elicit higher JOLs.

Mean correlations between valence and JOLs were analyzed in a 2 (block: block 1 vs. block 2) x 3 (emotion: negative vs. neutral vs. positive) repeated measures ANOVA. There was no main effect of block on correlations, $F(1, 45) = 0.08$, $MSE = 0.04$, $p = .780$, $\eta_p^2 < .01$. However, the main effect of emotion on the valence and JOL correlation was significant, $F(2, 90) = 82.47$, $MSE = 0.04$, $p < .001$, $\eta_p^2 = .65$. Post-hoc comparisons showed that the correlations between valence and JOLs for negative images were significantly greater in magnitude than for neutral, $t(45) = 10.80$, $p < .001$, $d = 1.59$ and positive, $t(45) = 10.17$, $p < .001$, $d = 1.50$, which did not significantly differ from each other, $t(45) = 2.13$, $p = .095$, $d = 0.31$. Moreover, the interaction between block and interaction was significant, $F(2, 90) = 5.68$, $MSE = 0.02$, $p = .005$, $\eta_p^2 = .11$.

Post-hoc comparisons demonstrated that the valence x JOL correlations for neutral, $t(45) = 0.64$, $p = .987$, $d = 0.09$, and positive, $t(45) = 1.88$, $p = .067$, $d = 0.28$, images did not differ across blocks. The magnitude of the correlation for negative images numerically decreased in block 2 relative to block 1, $t(45) = 2.35$, $p = .195$, $d = 0.35$, and although the Tukey-corrected p-value is not significant ($p = .195$), the uncorrected p-value is ($p = .023$) was. Thus, the nature of the significant interaction is such that correlations between valence and JOLs for neutral and positive images do not differ across blocks, while the correlations between valence and JOLs for negative images decrease, but they significantly differ from neutral and positive.

Mean correlations between arousal and JOLs were also analyzed in a 2 (block: block 1 vs. block 2) x 3 (emotion: negative vs. neutral vs. positive) repeated measures ANOVA. The main effect of block on correlations between arousal and JOL was significant, $F(1, 45) = 4.10$, $MSE = 0.03$, $p = .049$, $\eta_p^2 = .08$. Post-hoc comparisons showed that mean arousal and JOL correlations were significantly higher in block 1 than for block 2, $t(45) = 2.03$, $p < .049$, $d = 0.30$. Furthermore, the main effect of emotion was significant, $F(2, 90) = 59.836$, $MSE = 0.04$, $p < .001$, $\eta_p^2 = .57$. Post-hoc comparisons revealed that correlations between arousal and JOLs for negative images were significantly larger than for neutral images, $t(45) = 6.30$, $p < .001$, $d = 0.92$, which, in turn, were significantly larger than for positive images, $t(45) = 4.18$, $p < .001$, $d = 0.62$. Correlations between arousal and JOLs for negative images were significantly larger than for positive images, $t(45) = 11.47$, $p < .001$, $d = 1.69$. Lastly, the block x emotion interaction was not significant, $F(2, 90) = 0.14$, $MSE = 0.03$, $p = .866$, $\eta_p^2 < .01$.

Table C.1

Mean Within-Participant Correlations between Normed Valence and Arousal Ratings and JOLs

	Block 1			Block 2		
Factor	Negative	Neutral	Positive	Negative	Neutral	Positive
Valence	-.32(.03)***	.07(.03)*	.04(.02)	-.25(.03)***	.05(.03)	-.03(.03)
Arousal	.27(.20)***	.08(.17)**	-.04(.16)	.22(.20)***	.04(.20)	-.07(.20)*

Note: Standard errors are shown in parenthesis beside their respective means.

*$p < .05$. **$p < .01$. ***$p < .001$.

Appendix D: JOL Reaction Time (RT)

Analysis of the JOL RTs was not included in the main text of the present study primarily because participants were not instructed to make JOLs quickly, and varied keyboard locations will undoubtedly affect RT. Therefore, RT in the current study is only a very rough estimate of actual decision time. However, analyzing differences in JOL RT across emotion conditions remains interesting. Mean JOL RTs (see Table D.1) were analyzed in a 2 (block: block 1 vs. block 2) x 3 (emotion: negative vs. neutral. vs positive) repeated measures ANOVA. The main effect of block on RT was significant, $F(1, 45) = 34.72$, $MSE = 0.61$, $p < .001$, $\eta_p^2 = .44$. Post-hoc comparisons demonstrated that mean JOL RT was significantly higher in block 1 than in block 2, $t(45) = 5.89$, $p < .001$, $d = 0.62$. The main effect of emotion on JOL RT was significant, $F(2, 90) = 15.77$, $p < .001$, $\eta_p^2 = .26$. Post-hoc comparisons revealed that JOL RT for negative images was significantly higher than RT for neutral images, $t(45) = 6.34$, $p < .001$, $d = 0.41$, and for positive images, $t(45) = 3.79$, $p = .001$, $d = 0.61$. However, JOL RT for neutral images were not significantly different from positive images, $t(45) = 0.54$, $p = .852$, $d = 0.21$.

The interaction between block and emotion was significant, $F(2, 90) = 6.74$, $MSE = 0.19$, $p = .002$, $\eta_p^2 = .13$. Post hoc comparisons demonstrated that mean JOL RT for negative images in block 1 was significantly higher than negative images in block 2, $t(45) = 4.87$, $p < .001$, $d = 0.72$. Moreover, JOL RT for negative images in block 1 were significantly higher than neutral images in block 1, $t(45) = 5.10$, $p < .001$, $d = 0.42$. JOL RT for negative images in block 1 were also significantly higher than positive images in block 1, $t(45) = 3.46$, $p = .014$, $d = 0.48$. JOL RT for neutral images in block 1 were not significantly higher than those for neutral images in block 2, $t(45) = 4.62$, $p = < .001$, $d = 0.58$. JOL RT for neutral images in block 1 did not significantly differ from that of positive images in block 1, $t(45) = 0.31$, $p = 1.000$, $d = 0.03$. JOL RT for

negative images in block 2 were significantly higher than those for neutral images in block 2, $t(45) = 4.19$, $p = .002$, $d = 0.25$, but not significantly differ from positive images in block 2, $t(45) = 1.59$, $p = .612$, $d = 0.11$ Moreover, JOL RT for neutral images in block 2 did not significantly differ from positive images in block 2, $t(45) = 2.96$, $p = .053$, $d = 0.14$. Lastly, JOL RT for positive images in block 1 were significantly higher than positive images in block 2, $t(45) = 5.89$, $p < .001$, $d = 0.49$.

Table D.1

Mean Reaction Times (in seconds) for Participant JOLs in Blocks 1 and 2

	Block 1			Block 2		
Reaction Time (RT)	Negative	Neutral	Positive	Negative	Neutral	Positive
Mean	2.10	1.59	1.56	1.28	1.10	1.20
SD	1.14	1.01	0.74	0.79	0.65	0.73
Range	0.67—8.96	0.48—5.94	0.55—3.97	0.33—3.34	0.36—3.27	0.44—3.62