Examining Cultural Differences in Recognition Memory Response Bias: An Extension of the MBBE

by

Majd Z. Hawily
BA, from Lebanese American University, 2020

A Thesis Submitted in Partial
Fulfillment of the Requirements
for the Degree of

MASTER OF SCIENCE

in the Department of Psychology

Examining Cultural Differences in Recognition Memory Response Bias: An Extension of the MBBE

by

Majd Z. Hawily
BA, Lebanese American University, 2020

Supervisory Committee

Dr. D. Stephen Lindsay,
Supervisor Department of
Psychology

Dr. Michael E. J. Masson,
Departmental Member
Department of Psychology

Dr. Adam Krawitz,
Departmental Member
Department of Psychology

Abstract

According to signal detection theory, people tested on an old/new recognition memory test adopt a liberal, conservative, or neutral response criterion. Several prior studies in our lab demonstrated that subjects showed a clear conservative bias when presented with complex images (e.g., paintings, photographs of scenes) as stimuli. When stimuli were English words, bias tended to be liberal or neutral. The reasons for these materials-based differences in response bias remain ambiguous. Our efforts have focused on understanding response bias variation across materials and individuals. Specifically, we have explored whether Canadian and Japanese participants show differences in response bias for new materials called “diffeomorphs”. We conducted an earlier study with Lebanese participants with a smaller sample and materials size that served as a pilot study for our later studies. The materials-based bias effect cannot be applied to all visual stimuli because, even though both pictures and diffeomorphs are visual stimuli, the response bias for each is different. For example, we found that while Japanese elicit a conservative bias for diffeomorphs, Canadians have a neutral response bias. Besides the observed cross-cultural difference in response bias, this work refuted the hypothesis that novelty, colorfulness, and richness are behind the MBBE. It seemed that neither the semantics (line drawings) nor the colorfulness (diffeomorphs) of the stimuli appear to generate a bias towards conservatism. The MBBE and its cross-cultural generalization have been better understood because of the fresh insights offered by this thesis.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I want to thank my mentor, Dr. Steve Lindsay, first and foremost for his assistance with my MSc thesis and program. Steve always challenged me to critically analyze my outcomes and deliver the clearest picture to my audience. His wise counsel and depth of research knowledge have been extremely beneficial. This thesis is noticeably stronger because of his ongoing guidance to carefully consider my research topics and statistical techniques and to respect the values of open science. Additionally, I want to thank Drs. Mike Masson and Adam Krawitz, who were on my committee. Having such perceptive committee members who come from other fields than Steve and I, and their viewpoints and criticism, has been a wonderful delight. Their irreplaceable contribution has aided in my comprehension of the benefits of multidisciplinary cooperation and, perhaps, helped me connect my study to broader issues. I want to express my gratitude to my graduate student cohort, who helped me complete my degree by offering resources for my research as well as mental and emotional support during trying periods. Finally, I'd like to express my gratitude to UVic's amazing Cognition and Brain Sciences group, whose suggestions helped shape and improve the analyses and interpretations of my findings.

Introduction

Understanding how we distinguish between stimuli that we have experienced in a particular context and those that we have not encountered is necessary. This cognitive process, known as recognition memory, is involved in all areas of our everyday lives, such as visiting special locations or figuring out whether certain supermarket goods are available in a shop. To evaluate recognition memory, a variety of techniques have been developed. One widely used technique consists of two independent phases. Subjects learn the stimuli during the initial phase, and then during the test phase, they are shown a sequence of both studied and unstudied stimuli and are asked to determine whether they have studied the stimuli or not. Their judgment can be drawn from a straightforward yes-or-no response or by choosing an answer from a confidence-weighted scale.

Recognition Memory and Signal Detection Theory

Various theoretical and modeling perspectives have been proposed to analyze recognition memory data. Utilizing signal detection theory to conceptualize the data in a structured way has been predominant for two-phase recognition experiments (Green & Swets, 1966). Originating in psychophysics, SDT aims to describe the processes by which humans decide whether a signal is present under noisy conditions, such as in the case of detecting tumors through medical examinations. Notably, using signal detection theory for analyzing recognition memory data has impacted the way researchers describe measures of performance pertinent to such tasks. SDT suggests that performance in tasks relies on two independent processes: response criterion and discrimination.

In a recognition memory test, sensitivity (sometimes called discrimination) refers to the degree to which a subject is successful in endorsing old items and rejecting new ones (Macmillan & Creelman, 2005). In other words, it could be thought of as the diagnostic accuracy of an

individual. In studies of recognition memory, sensitivity, as well as the factors influencing it, is most often the measure of primary interest. For reasons that shall be mentioned in the following chapters, response bias, which is the other derived score, shall also be considered. It characterizes a subject's tendency towards calling items "studied" or "unstudied".

Response bias refers to the decision criterion that reflects a subject's willingness to endorse an item as studied or unstudied. In other words, it is the threshold beyond which a stimulus is classified as a signal. To elaborate, a subject can have a neutral response bias, whereby they show no consistent inclination toward choosing one response over the other. Contrastingly, a subject could have a liberal bias whereby they refer to items as "studied". In this case, their responses would consist mostly of hits and false alarms. Also, there is a third category, a conservative bias, referring to the subject's tendency toward rejecting items and producing mostly correct rejections and misses. The direction and extent of response bias can differ among individuals, experimental conditions, and item classes.

The two measures are independent and distinguishable. To clarify, modifying the response bias of a procedure modifies the probability of identifying the stimuli, but modifying discrimination leads to deviations in the accuracy of recognition. Finding the response bias and sensitivity relies on understanding the possible outcome categories that an experimental recognition test produces. SDT studies decision-making when there are only two choices to make with four possible outcomes: hits, misses, correct rejections, and false alarms, also known as the confusion matrix (Abdi, 2010). More specifically, using signal detection theory to analyze the data involves calculating participants' hit (correct "studied" responses) and false alarm (incorrect "studied" responses) rates.

There are two possible correct responses: hits, which describe when a subject successfully recognizes a studied item, and correct rejection, wherein a subject correctly rejects a non-studied

item. The other two false responses are false alarms (FA), which are incorrect endorsements of unstudied items as old, as well as misses, which are when subjects fail to identify old items as old. These possible outcomes are conceptualized and analyzed with signal detection theory (Macmillan & Creelman, 2005). Scientists studying recognition memory often report subjects' hit and false alarm rates. But it is also common to use the hit and false alarm rates to calculate indices of subjects' sensitivity (the ability to discriminate old items from new items) and response bias (the tendency to classify items as old versus new).

According to signal detection theory, operators set a response criterion used for decision-making. This suggests that in a target detection task, such as an old/new recognition test, subjects decide whether they did or did not study the item by comparing the strength given by the current test stimulus to a decision criterion that they set for themselves (Benjamin et al., 2009). If a test probe gives rise to at least that amount of evidence of oldness, then the operator will classify the item as old; if not, they will classify it as new. The claim seems to be that this threshold or criterion is established before the first test item is presented and is maintained unless some aspect of the procedure changes (such as a manipulation of payoffs to favor hits or CRs).

It has been argued that setting a response criterion is subject to various sources of noise that produce variability in the criterion across trials (Kellen et al., 2021). How the criterion is established prior to the first test probe is not clear, but global matching models of recognition suggest some clues (Osth & Dennis, 2020). It may be that subjects generate expectations about the probability distribution of various amounts of evidence of oldness based on the subjective memorability of the studied materials. The current criterion location is set in relation to various factors, including subjects' experience, implicit and explicit payoffs (such as the reliance on confidence ratings), and involvement thus far in the test (Benjamin et al., 2009). Criterion placement could be further affected by retrieval errors that could result from intervening or

unexpected tasks or events that disrupt the normal pace or rhythm of the test.

Factors that Affect Response Bias

Experimental Factors:

Experimental manipulation influences response bias in many ways. One way an experimenter could do so is by asking participants to be cautious or flexible in endorsing old items. Another way response bias is influenced by experimental manipulation is by varying test list compositions by modifying the proportion of old items, which influences responding (Van Zandt, 2000). A third method that could affect responding is manipulating the payoff structure of the experiment by, for example, altering the reward/punishment system, such as offering participants a greater reward for hits than correct rejections, thus encouraging a liberal response (Van Zandt, 2000).

Individual Differences:

The deviation of response bias affected by factors is not generalizable across groups of individuals. In fact, while response bias is calculated by finding the mean across participants, some evidence hints at a between-subjects variability in bias. For instance, Aminoff and colleagues (2012) reported individual differences in response bias with respect to the manipulation of the proportion of old and new items on the test. Normally, group averages are used to examine response bias, but there may be important individual differences at play. This suggests that some people want more memory proof before classifying something as "old." Whether recognition response bias is partially a trait-like predisposition was examined in four experiments (Kantner & Lindsay, 2012). In the first experiment, bias was found to be highly correlated between two recognition study-test cycles that were separated by ten minutes. When tasks were separated by one week in Experiment 2, a roughly comparable correlation was seen.

Even though the stimuli between the first and second study-test cycles changed noticeably, Experiment 3 showed strong bias correlations. In general knowledge tests, there was no link between response strategies and bias (Experiments 2 and 4); however, bias only marginally predicted the frequency of false recall in the Deese/Roediger-McDermott (DRM) paradigm. These results show a trait-like stability in response bias, indicating that recognition memory theories would need to take this feature of bias into consideration in addition to its modulation by task factors.

Affect and Recognition Memory Response Criterion:

In recognition memory experiments, item attributes affect response bias. In the memory literature, researchers have shown that emotion plays a role in recognition memory processes. Speculations on the effects of emotion on sensitivity have not led to consistent conclusions, even though there seem to be response bias differences because of emotional effects (Dougal and Rotello, 2007). To illustrate, more liberal responses to emotional stimuli compared to neutral stimuli were evident, particularly when the emotional stimuli were associated with negative arousal (Dougal & Rotello, 2007). Another study highlighted the tendency to respond liberally to emotionally salient stimuli (Windmann & Chmielewski, 2008). However, in the same study, when the photographs depicted emotional and neutral facial expressions, the same effect was not detected (Windmann & Chmielewski, 2008).

Response bias Differences Depending on Stimuli

Words and Pictures:

In a series of experiments, Lindsay and Kantner sought to investigate the influence

of accuracy feedback on recognition memory for complex and relatively novel stimuli, specifically poetry, paintings, and Korean melodies. While the results generally showed null effects of accuracy feedback on recognition accuracy, tests consistently revealed a conservative response bias (Lindsay & Kantner, 2011). Among all the stimuli used in this study, the reported results were remarkably consistent for paintings. A consistent conservative bias was evident in 26 of 32 tests and in both feedback and control groups, and in 22 of these cases, this bias differed significantly from zero (Lindsay & Kantner, 2011). It is important to note that even though feedback seemed to increase conservative bias for poetry, paintings, and Korean melodies, this effect was not replicated. On the other hand, a conservative response was marked for paintings, for which response bias was conservative in all experiments for both feedback and control groups.

In subsequent studies, Lindsay et al. (2011) tried to compare response bias for paintings among subjects who did versus did not get trial-by-trial accuracy feedback. These studies were specifically designed to observe the consistency of the response bias with these stimuli, but a robust effect was not observed. The forest plot in Figure 1 displays response bias meta-analyses for five experiments which used the paintings as materials such that mean (c) here is collapsed across feedback and control groups.

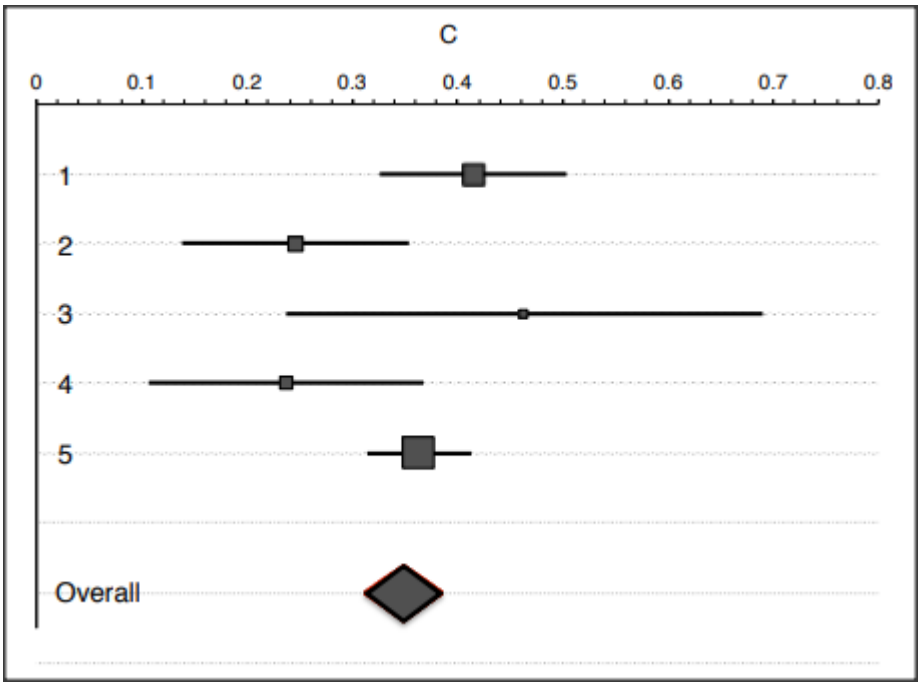


Figure 1: Forest plot depicting mean response bias (c) values obtained by Lindsay and Kantner (2011)

Based on the related variance and sample size, the size of each square reflects the weighting of that mean in the meta-analysis computation. Bigger squares, then, reflect means that were more strongly weighted in the final computation and often originate from bigger samples with low variation. 95% confidence intervals are shown in error bars. Also, the diamonds at the bottom of each plot represent the result of the associated meta-analysis and include the corresponding 95% CI. Experiments that compared response bias for words versus paintings, the trend of bias towards paintings was held consistent as Figure 2 shows. More specifically, this data was obtained from nine within-subjects experiments in which subjects studied and were tested on mixtures of both types of items.

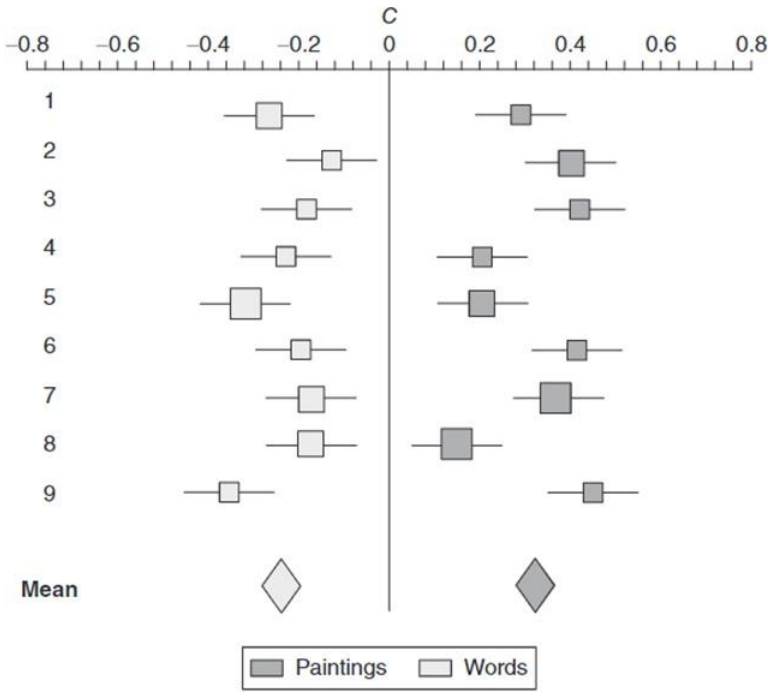


Figure 2: This figure displays a forest plot of estimates of response bias for recognition paintings and words (as constructed by Dr Kaitlyn Fallow). This data was obtained from nine within-subjects experiments in which subjects studied and were tested on mixtures of both types of items.

It is noted that each test probe was presented one at a time in a random order for a confidence-weighted judgment as shown in Figure 3. This figure depicts the study and test phases wherein the stimulus used here is a painting.

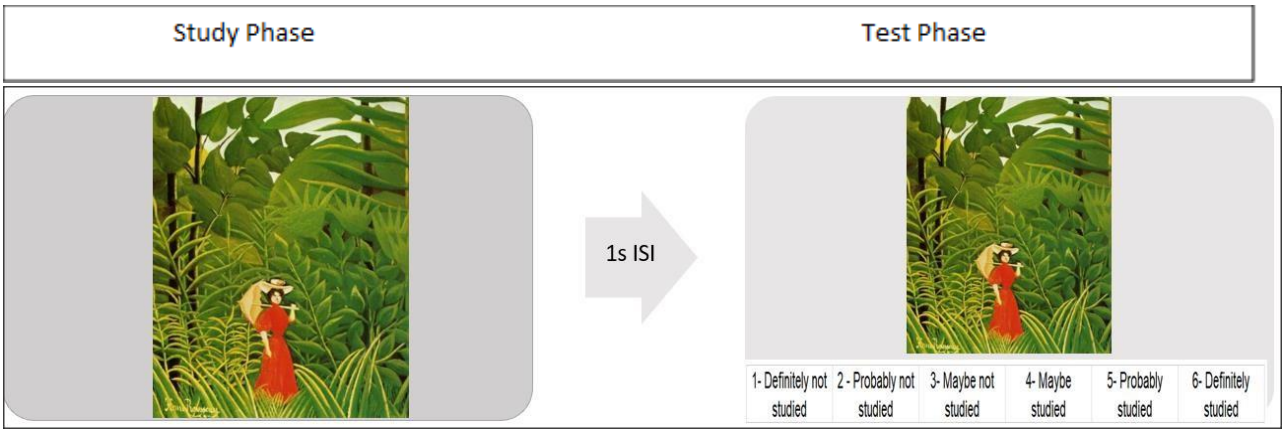


Figure 3: This figure displays the experimental procedure for recognition in the case of a painting stimulus.

The Importance of Considering Response bias

Although response bias had not often been the focus of interest in recognition memory studies, findings reported by Lindsay and Kantner (2011) illustrated the importance of considering response bias. This unexpected finding inspired Lindsay and Fallow to compare response bias directly for paintings and words in a within-subjects design. In several experiments, subjects studied a mixed list of unrelated words and paintings and were later tested on old items (in the same format as the study) mixed with new words and paintings. While paintings consistently triggered a conservative response bias, words usually yielded a liberal one. As for sensitivity, it tended to be higher for paintings than words in this study (as per the picture superiority effect, Paivio, 1968). But Lindsay et al. found that they could obtain better discrimination on words than on paintings by deliberately selecting high-frequency words. Whether discrimination was better or worse for paintings than for words, response bias tended to be more conservative for paintings than for words (Lindsay, Kantner, & Fallow, 2014).

The finding that subjects revealed a conservative bias when they studied and got tested on paintings in comparison to words has been replicated in a series of follow-up experiments. This steady finding has been termed the Materials-Based Bias Effect (MBBE) (Lindsay, Kantner, & Fallow, 2014). The name of the effect suggests that it could apply not only to paintings but also to other stimuli that could be similar or different to paintings in terms of characteristics.

Why is Response bias Conservative for Paintings?

A key question arises here: Why do subjects tend to exhibit a conservative bias when the study and test materials are paintings? Several hypotheses regarding the origin of conservatism in paintings have been tested, but none has received strong support. Perhaps the nature of the painting stimuli, and specifically their richness and distinctiveness, led subjects to expect that they would have strong and clear recognition responses to the studied paintings, which in fact they did. Inarguably, paintings are perceptually complex and diverse, which has documented memory implications for recognition performance. Further, paintings often include multiple concepts and themes, signaling various semantic meanings.

One way of explaining the MBBE is that subjects may tend to have an exaggerated expectation of the memorability of paintings. In theory, that would lead them to set a higher response criterion for paintings than for words. An unduly high response criterion for paintings would produce a conservative response bias: more misses than false alarms. This is known as the subjective memorability hypothesis, which claims that people expect paintings to be more easily retrieved than words (Brown, Lewis, & Monk, 1977). It is suspected that subjects expect a sense of familiarity or oldness during the test phase for items that they judge to be memorable. The expectation of the possible familiarity of a stimulus could be influenced by a wide range of variables, such as personal relevance or repeated presentations, which would lead subjects to expect a strong sense of "oldness" for such items at the test phase.

Similarly, subjects could adopt a stringent criterion and think that a conservative approach elicits an accurate rejection of new items. In this sense, the memorability of a stimulus is not reliant on the objective properties pertinent to the item but rather tied to the

participant's pre-determined subjective criteria. A series of experiments have hypothesized that subjects would expect to remember more paintings than words, and the extent of this tendency would be related to the extent of the difference in response bias between paintings and words (Lindsay & Kantner, 2011). The memorability hypothesis was not supported despite the significant difference in bias between paintings and words. Expecting better memory for paintings than words did not predict more conservative responses to paintings as compared to words, given that participants did not tend to anticipate better memory for paintings than words. This is evident from the fact that there was no correlation between differences in memorability estimates and response bias. Within this perspective, based on the results of these four studies, researchers could not support the notion that the subjective memorability hypothesis could explain the response bias differences (Lindsay & Kantner, 2011).

A different account of the MBBE has to do with the fact that the paintings used in these studies were unknown to the subjects, whereas the words were common English words that the subjects had encountered many times in everyday life. Thus, pre-experimental familiarity would be lower for paintings than for words. To the extent that pre-experimental familiarity contributed to evidence of oldness on a recognition memory test, hit and false alarm rates would be higher for words the two types of stimuli were intermixed at study and test and the same response criterion was used for both. That is, the difference in c values for words versus paintings would not be because of a difference in response criterion, but rather because of greater extra-experimental familiarity for words than paintings. But this account would not explain the fact that conservative response bias is observed on paintings when they are studied and tested alone.

Besides the contribution of pre-experimental familiarity in providing an

explanation, the reminding hypothesis claims that if certain items are more conducive to study phase retrieval, such items may produce confusion when encountered on the test (Hintzman, 2009). Based on this hypothesis, it could be that paintings, being arguably more visually striking than words, more often remind people of previously viewed paintings or increase the likelihood that people will become aware of such events when they occur. This could lead subjects to become cautious when endorsing paintings at the test and thus adopt a conservative criterion. This hypothesis was not supported in a series of experiments conducted by Lindsay et al. (2014).

Another possible explanation could be that people demand more or qualitatively different evidence to endorse an item as "studied" when it belongs to a more distinctive category (Schacter et al., 1999). This is known as the distinctiveness heuristic. But photos of familiar locations are more often correctly recognized and rejected than photos of unfamiliar locations, perhaps due to the higher conceptual distinctiveness of well-known scenes, suggesting that a mirror effect rather than conservative response bias arises from greater distinctiveness (Dobbins & Kroll, 2005).

Using a recognition memory procedure, Fallow and Lindsay (2022) found that response bias was neutral on average for both the words and line drawings in the Snodgrass and Vanderwart (1980) stimulus set. Akin to paintings in their sense of novelty, line drawings depicting familiar objects are new to participants. This shows that novelty, as it is present in recognition memory studies, is not the main factor that could account for a materials-based bias effect. This does not mean that novelty, complexity, distinctiveness, and pre-experimental familiarity are disregarded for forming a better conceptual understanding of the materials-based differences.

The tendency for subjects to be conservative when recognizing visual stimuli such

as paintings has been a mystifying maze. There have been extensive efforts to understand the reasons behind the memorability of visual stimuli, but the mystery remains. Developing an understanding of how and why response biases are affected by materials will help us understand the basic mechanisms underlying those biases. More importantly, a clarified understanding of recognition memory response bias may enhance the understanding of target-detection tasks in other fields, such as the medical field.

After naming a materials-based bias effect, a skeptical researcher cannot help but ask about the generalizability of the materials-based bias effect to all visual stimuli, or, in other words, do other visual stimuli exhibit a conservative bias? If yes, then what are the characteristics that end up triggering such a bias in participants? As mentioned above, our lab has preliminary evidence that suggests that something like visual complexity plays an important role (hence, for example, bias for simple line drawing was neutral and bias for photos of those same objects was conservative; further, bias for color photos of scenes was more conservative than bias for gray-scale versions of those scenes).

MBBE and Cultural Differences

Another interesting question is whether the MBBE interacts with cultural differences. In prior studies with paintings, the paintings have all been by Western painters, and most subjects have been Westerners. Most of the early studies by Lindsay, Kantner, and Fallow (2014) tested undergraduate psychology students at the University of Victoria, most of whom were Canadian. Do people in other parts of the world, such as Japan, have the same response bias triggered by paintings as Westerners? Asking about cross-cultural generalizability does not merely address the consideration of generalization across cultures but could tap into further understanding the mechanisms and dynamics that lead to such a bias.

A Review of the Literature Pertaining to Cultural Differences in Memory

So far, several studies have shown consistency in revealing a conservative response bias when pictures are the stimuli in the experiment. The theory cannot be generalized, as most research has been conducted with Western participants. This highlights the necessity of testing the hypotheses on non-Western populations.

In the studies that we will outline in this section, non-Western populations have been tested, providing fruitful insights on cross-cultural cognitive similarities and differences. In one study that aimed to examine the cultural variations in memory, researchers searched for distinct items and backgrounds that had been integrated together in a single image (Wong et al., 2018). An image recognition paradigm was then used to test 36 Caucasian Canadians in Toronto and 36 Chinese in Beijing. Participants watched 60-line drawings of objects and scenes in the background while encoding. The participants were then asked to identify (through making remember/ know/ sure/ guess judgments) these items and backdrops separately before reporting subjectively whether they focused on the objects, the backgrounds, or both during encoding. Note that while there were two groups of participants, one paying attention only to objects, and the other paying attention both to objects and backgrounds, both were viewing objects and scenes in the background in the same image.

In general, Wong et al. (2018) found little evidence of cultural variation in how people remember solitary items. Canadian subjects had noticeably superior background memory compared to their Chinese counterparts. Canadians scored better than Chinese in background recognition, but there was no cultural difference in object recognition. Using the Stanislaw and Todorov (1999) method, the response bias measure was generated to investigate memory performance variations. This study found that participants were more

reserved in their replies to backgrounds than to objects (i.e., more likely to say "no" to backgrounds than to objects). Since neither the culture main effect nor the culture by picture type interaction were statistically significant, it is likely that response bias was the same for all groups. The findings of this study provide insights into the ways that memory for certain visual features varies between cultures.

Participant Subgroups	Canadians *M* (*SD*)	Chinese *M* (*SD*)
Attention to objects		
n	21	27
Background recognition	0.24 (0.20)	0.19 (0.12)
Object recognition	0.49 (0.23)	0.50 (0.18)
Attention to both objects and backgrounds		
n	14	9
Background recognition	0.35 (0.17)	0.17 (0.11)
Object recognition	0.49 (0.13)	0.43 (0.27)

Figure 4: This figure represents a table taken from Wong et al. (2018) which outlines the hits minus false alarms scores for participants in each subgroup (who only paid attention to objects versus participants who paid equal attention to objects and backgrounds during encoding).

Researchers in this study did not find a cultural difference in participants' memories of objects when these objects were separated from their associated backgrounds. However, they found that Canadian participants recognized more background scenes than Chinese participants. This could be due to a cultural difference in the binding processes during encoding. More specifically, perhaps Chinese participants may be more likely than Canadian participants to process information holistically if they intend to process both objects and backgrounds. In an earlier study conducted by Millar et al. (2013), 32 American and 32 Asian students participated in a computerized recognition test to determine if cultural variations in attention and memory have an impact on the specificity of memory (such as greater memory for visual detail). Participants were instructed to categorize 144 pairs of color photos of familiar objects as being the same,

similar, or new.

To evaluate both particular and general memory for each of the three conditions (oneself, one's mother (close other), and Bill Clinton (unknown other), researchers computed six memory scores for each participant. General recognition was the proportion of either correct 'same' or incorrect ‘similar' responses provided to previously shown items, whereas specific memory was the proportion of correct 'same' responses given to previously viewed objects. With culture as a between-subjects component, memory type (specific, general), and condition (self, mother, Bill Clinton) as a within-subject factor, response accuracy was examined. The findings revealed that culture, memory type, and condition all had a major influence. Most remarkably, culture and memory type showed a significant interaction. Follow-up tests indicated substantial cultural differences in specific memory with Americans performing significantly better than Asians. No significant differences between cultures in general memory were detected.

A follow-up experiment was created to test particular and general memory scores for both focus objects and background pictures presented as composite scenes to investigate the role of background context on memory specificity. Materials included 66 isolated object pairs, 66 isolated background pairs, and 44 composite scene pairings. Participants were asked to indicate whether, if confronted with each scene in real life, they would move closer, go farther away, or remain at the same distance. A 15-minute retention period was followed by a memory test. Participants in the recognition test were shown just isolated photos of items or backgrounds, 44 of which were the same as the things or backgrounds shown in the encoded scenes, 42 of which were comparable to the objects or backgrounds, and 43 of which were completely different backgrounds or settings The cultural difference for specific memory was bigger than that for general

memory, and the interaction between culture and memory type was hardly significant. According to these two studies, there are cultural disparities in the accuracy of certain memory tests for visual information, with Americans performing better than East Asians. Although it does not influence the broad recognition of pictures, this is probably due to a Western-favored style of visual feature processing.

Although in Millar et al. 's study (2013), East Asians seemed to remember images with less perceptual detail than Americans, this result was not replicated. In fact, a later study sought to understand the presence of a cultural variation in memory-related decisions among 19 American and 18 East Asian participants (Paige et al., 2017). The materials came with 252 images of well-known buyable items that were balanced throughout four distinct versions. Participants in Session 1 quickly determined whether they would approach, avoid, or maintain the same distance from each thing if they were to really come across it while they encoded 108 of the object photographs while in an MRI scanner. Participants in Session 2 did a surprise recognition exam on a computer, viewing 144 objects—54 identical, 54 similar, and 36 new—and responding "same," "similar," or "new." Researchers evaluated cultural variations in memory specificity, concluding that culture had no significant main influence. There was no cultural influence on particular memories alone; hence, this conclusion does not repeat earlier literature as there was no cultural effect on specific memory only.

The analysis of discriminability also did not reveal a cultural effect on memory for specific information. They compared the groups on d' to account for false alarms when evaluating memory specificity. The interaction between Americans and East Asians for sensitivity to identical and comparable trials was not significant. There was also a main effect of culture, in which Americans had a higher level of performance than East Asians

across conditions. In other words, Americans performed better than East Asians across conditions, but there was no evidence of greater cross-cultural differences in memory for specific tasks as opposed to the literature. However, regarding differences in specific memory, Americans were less conservative with their ‘‘same’’ responses than East Asians, for both approach and purchase decisions. Given that there was a discernible difference in memory specificity due to culture, the findings suggest that cultural variations altered the brain areas mediating memory specificity since East Asians engaged the left fusiform and left hippocampus more than Americans for particular vs. general memory. More importantly, it seemed that Americans exhibited higher levels of recognition (adopting a liberal criterion) for previously seen “same” items than East Asians. This could be explained by the assumption that Americans could pay more attention to the precise visual details of the item (Millar et al., 2013).

Other studies have tested other populations, such as Australian children. In one such study, Indigenous (Australian) children seemed to follow a conservative decision-making criterion compared to non-Indigenous (Australian) children when completing visual working memory tasks (Freire et al., 2020). To measure whether perceived visual working memory ability differs across cultures, 26 Indigenous and 30 non-Indigenous children were tested through a battery of tasks that were administered individually. Each task comprised 12 randomized practice trials, followed by 64 randomized trials, and the task involved presenting two sets of either colored (color change detection task) or animal (animal change detection task) squares (which were counterbalanced). The first set of stimuli was displayed, followed by the second set of stimuli, whereby a response was made by indicating whether the second set of presented stimuli was the “same” or “different” as the first set. This is somewhat akin to a yes/no recognition test but studying

a list of items and then being tested on a mix of multiple studied and non-studied items is different. In this case, in the visual working memory task, there are multiple study-test cycles in which a to-be-remembered target is presented followed by a test probe.

While researchers were expecting that non-Indigenous children would exhibit weaker visual working memory skills than Indigenous children, the results did not show that. The non-Indigenous group showed a significantly greater perceptual sensitivity for detecting stimulus changes than the Indigenous group. Moreover, there was a significant main effect of group on response bias scores overall. The Indigenous group demonstrated a significantly more conservative response bias than the non-Indigenous group, regardless of task or array size. Notably, both groups showed a conservative response bias. This suggests that Indigenous children demonstrated greater overall mean detection sensitivity, compared to non-Indigenous children, on a change detection working memory task. The researchers posited that the performance disparities between Indigenous and non-Indigenous children could be due to the impact of cultural factors on cognitive and perceptual processing, which subsequently affect how individuals approach cognitive tasks.

In addition to providing literature on the potential cultural variation in memory, it could be relevant to outline what the cognitive-neuroscientific literature posits in this regard. Overall, many papers have reported comparisons of brain activity among people from different cultures. Most, if not all, of these papers show some differences. Often, these studies had modest sample sizes. Few, if any, of the studies pre-registered specific hypotheses and analysis plans. Few have been directly replicated by other labs. So, they must be interpreted with caution. But they are consistent with the possibility that members of different cultures exhibit different patterns of brain activity when performing

comparable tasks (view supplementary material).

The Purpose behind this Research

It could be that prior cross-cultural research has reported cultural variations in memory; the results are clearly inconsistent, limited, and lack replication for confirmation. More thorough research is required, including comparisons of object memory. A thorough search of the literature yielded only a few studies testing for cultural differences in recognition memory sensitivity, or response bias. The limitedness of research on how culture affects psychological processes has led to the approach of this thesis in understanding response bias.

Furthermore, highlighting the importance of considering cross-cultural differences when conducting cognitive research is another goal we hope to target within our research. Evaluating the proposition that cultural differences could affect cognitive task performance takes us closer to a unified understanding of the underlying mechanisms that could explain these differences. Despite the challenges to the study of cultural differences in cognition, we hope that we can address the topic in the most accurate way possible. For purposes of this thesis, the important points are that (a) there does not appear to have been much published research on this question, and (b) the current published papers include hints that there may be cultural differences in recognition memory response bias.

Exploring with a Cross-cultural Lens in Lindsay's Lab

Background

A pattern of materials-based response bias differences has been consistent in a series of more than 20 recognition memory experiments conducted at the University of Victoria (UVic) with undergraduate psychology students. Response bias has varied when the stimuli are words. More specifically, words (4-8-letter, moderate-high-frequency English nouns) have consistently produced a liberal bias in within-subjects designs when words and paintings are intermixed at study and test. However, a statistically neutral bias has been observed in word stimulus-only designs. On the other hand, when images of paintings are used as stimuli materials, the average response bias has always been markedly conservative.

Why were Canadian participants in our prior studies conservative when recognizing paintings? As mentioned earlier, the lab has explored various hypotheses (e.g., that subjects have an exaggerated expectation of the memorability of paintings and hence set their response criterion too high for them, or that the extra-experimental familiarity of words contributes to their recognition), but so far there was not clear and consistent support for any account. Some years ago, Lindsay et al. (unpublished) speculated that it might have something to do with the cultural significance of paintings. These studies avoided particularly famous paintings, but subjects would perceive them as high art, perhaps even as high-brow, and perhaps that, for some reason, makes subjects particularly likely to make a false alarm. If so, people from a different culture would not demonstrate a conservative response bias for the painting materials.

To date, the sole cross-cultural exploration of the MBBE conducted in the lab was led by Kaitlyn Fallow (supervised by Steve Lindsay) in collaboration with Dr. Shinji

Kitagami and his undergraduate students at Nagoya University in Japan. The studies used the same paintings as prior MBBE studies in Lindsay's lab, but the English words were replaced with Japanese words. In the two studies, subjects studied and were tested on a mix of paintings and words. The first study had a within-subjects design wherein subjects studied a mix of paintings and unrelated words and were later tested on those same stimuli randomly intermixed with new paintings and words and presented one at a time for confidence-weighted old/new recognition judgments. The second experiment served as a replication of the first with the exception that it had a between-subjects design.

Although I was not involved in these studies, I include them here because they have not previously been written up in detail. The methods, analysis, and results for both experiments are reported separately. More details regarding the methods of each experiment, including the wording of instructions and experiment-specific manipulations, are available at https://rb.gy/q6limp

Study 1

Methods:

Participants: Participants were 113 undergraduate students at Nagoya University in Japan who participated in exchange for bonus credit in a psychology course. Pre-planning of the sample size was not performed. Rather, the sample resulted from testing as many participants as possible over the course of a single semester.

Materials: The visual stimuli were 204 colour scans of relatively obscure masterwork paintings and 192 high familiarity and moderate- to high-frequency words. The verbal stimuli consisted of Japanese nouns that were taken from Amano and Kondo's (1999) Japanese word database. In brief, the selected words were roughly comparable to the previous stimuli (used in earlier studies) in that they were common nouns that would

be familiar to undergraduate participants. All words were presented in Kanji and composed of two characters. The words that were used were not related to the paintings.

Japanese word (*Kanji*)	Romanization	English translation
自分	*self*	jibun
英語	*English*	eigo
世界	*world*	sekai
空気	*air*	kuki
時間	*time*	jikan
朝食	*breakfast*	chōshoku
日曜	*Sunday*	nichiyō
名前	*name*	namae

Figure 5: This figure shows a sample of Kanji words from studies (1) and (2) with students from Nagoya University.

Procedure: The procedure was like prior experiments in the lab with Western participants. All instructions were presented in Kanji. The program began with a study phase, consisting of three primacy buffers, a 96-item list of randomly intermixed words and paintings, and three recency buffers, each presented one at a time. The studied paintings and words were randomly sampled for each participant. Participants were informed at the beginning of the study phase that they would see a series of paintings and words and were asked to try their best to remember each. The study phase was followed by a brief (one to two minutes) filler activity. The test phase included a randomized test list including 96 studied items mixed with 96 new items. Participants were asked to judge each test stimulus as studied or not studied using a 6-point confidence-weighted scale, with 1 denoting “definitely not studied” and 6 denoting “definitely studied”. During debriefing, participants rated their expertise on paintings on a 5-point scale.

Results: For the first study, following the exclusion criteria, only three participants were excluded from the analysis: two for self-reporting the highest level of art expertise on a 1-5 scale, and one as an outlier on d’ (more than 3 SD above the mean in the

painting's condition). Additionally, ceiling and floor hit (HR) and false alarm rates (FAR) were replaced according to Macmillan and Kaplan (1985) to calculate signal detection measures.

Observing the hit rates of the distribution reveals that they were significantly lower for paintings (0.63) than words (0.77), $t(109) = -11.39$, p .001, $d = -1.08$ (refer to Figure 6).

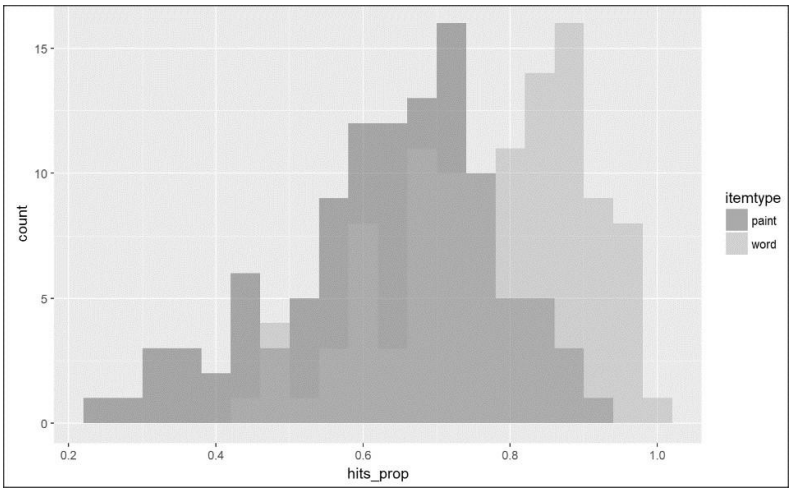


Figure 6: *This figure depicts the distribution of hits constructed by Kaitlyn M Fallow. Hit rates refer to the "hits_prop" in the x-axis. Note that the data presented here represents the Japanese participants' responses from Study 1.*

Similarly, false alarm rates were significantly lower for paintings (0.13) than words (0.19), $t(112) = -4.24$, p <0.001, $d = -0.40$ (refer to Figure 7).

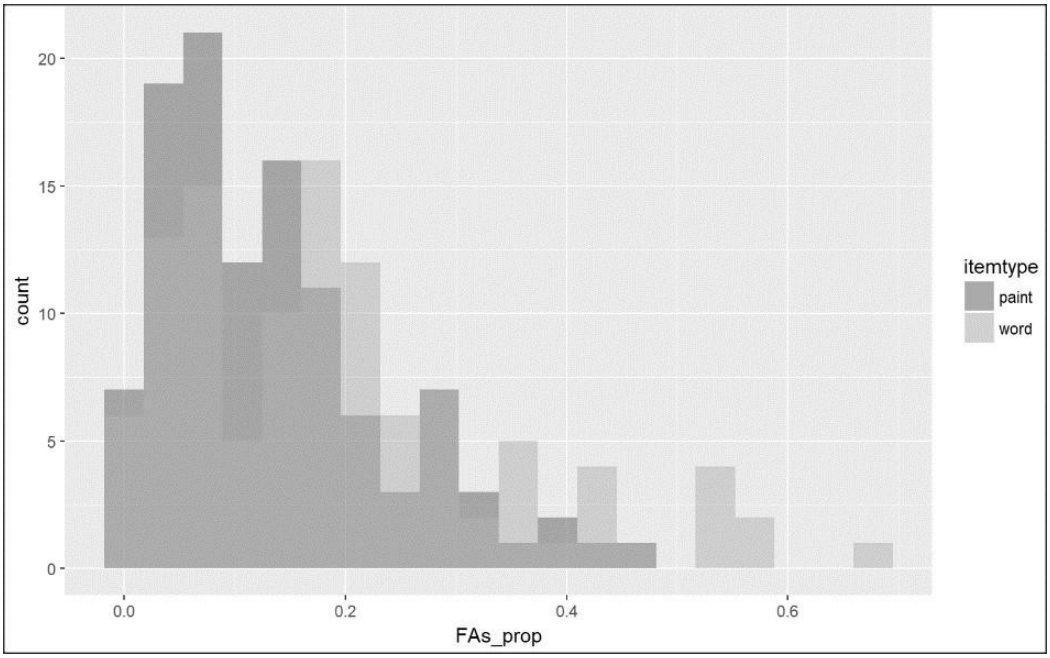


Figure 7: This figure depicts the distribution of false alarms constructed by Kaitlyn M. Fallow. False alarm rates refer to the "FAs_prop" in the x-axis. Note that the data presented here represents the Japanese participants' responses from Study 1.

As for sensitivity (d'), it was significantly better for words (1.87) than paintings (1.62), $t(109) = -3.78$, $p < 0.001$, $d = -0.36$. (1.66), $t(112) = -3.09$, $p = 0.003$, $d = -0.29$ (refer to Figure 8).

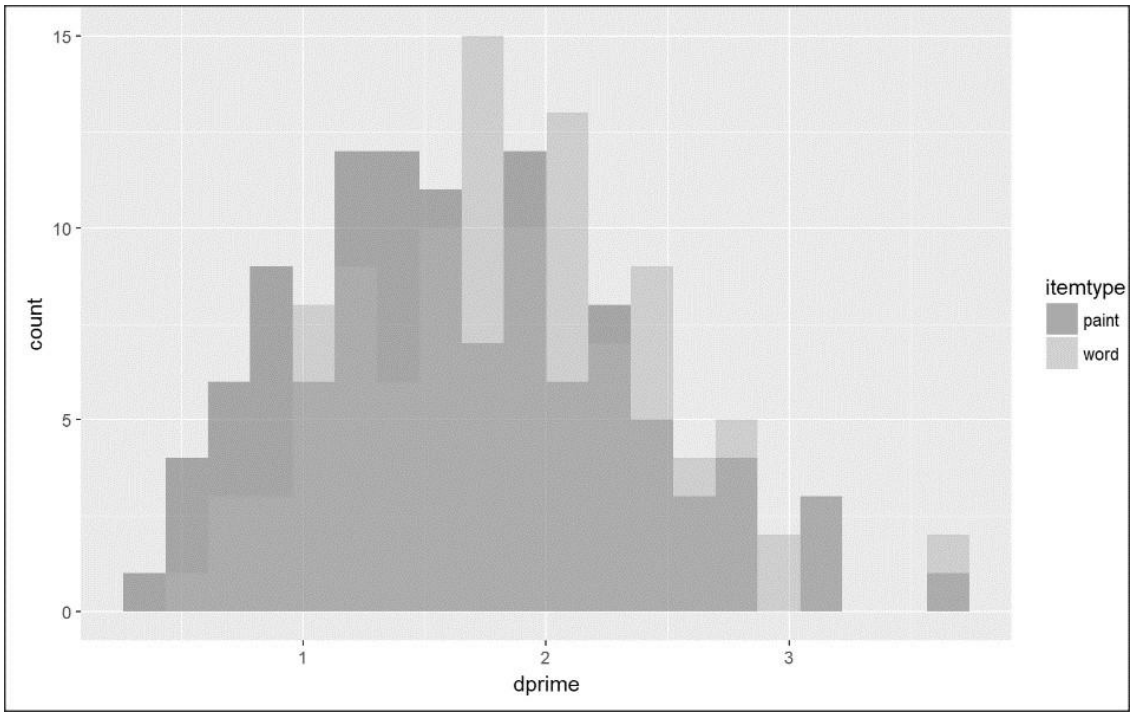


Figure 8: This figure depicts the distribution of sensitivity (d') constructed by Kaitlyn M Fallow Note that the data presented here represents the Japanese participants' responses from Study 1.

Response bias (c) was significantly more conservative for paintings (0.45) than words (0.09), t (109) = 7.49, p <0.001, d = 0.71. Notably, bias for words was also significantly conservative (relative to zero) (t (109) = 2.10, p =0.02, d = 0.20) (refer to Figure 9).

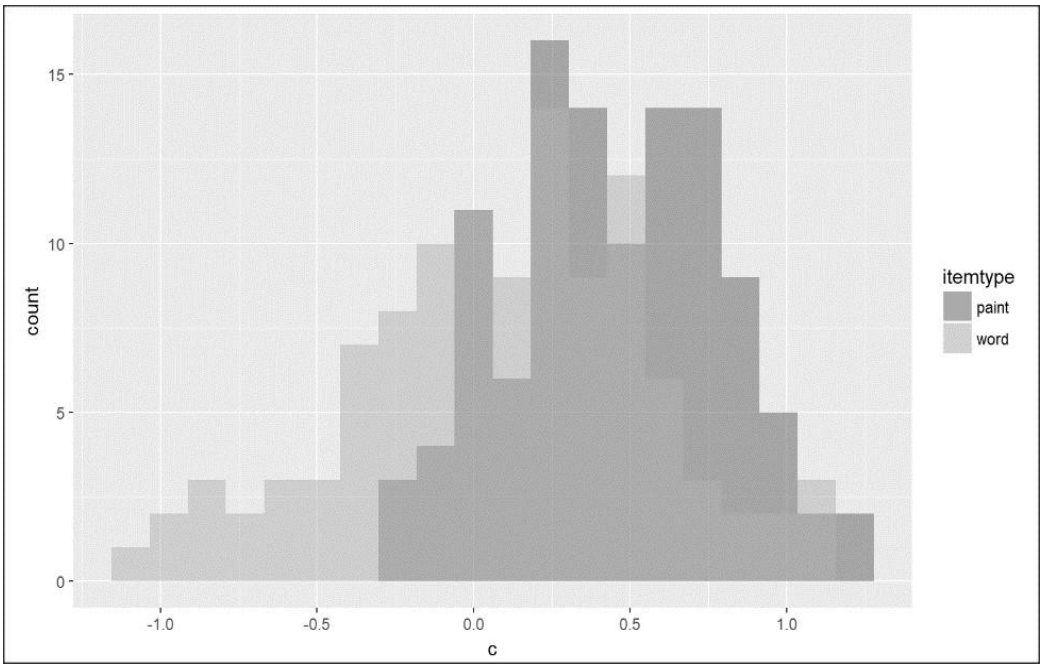


Figure 9: This figure shows the distribution of response bias (c) constructed by Kaitlyn M Fallow. C is used to measure response bias for the data. Note that the data presented here represents the Japanese participants' responses from Study 1.

Quartile-level analyses:

At the quartile level, the test list is divided into quartiles, and hit and false alarm rates were calculated for words and paintings (refer to Figure 10). Response bias (c) for paintings was increasingly conservative across quartiles while remaining roughly stable or fluctuating with no consistent pattern for words. As for sensitivity, it appeared similar in words and paintings (even though it was significantly greater for words). Hit rates showed a decline for both item types, although they seemed, on average, steeper for paintings.

While false alarm rates remained approximately stable for paintings, they increased slightly from the first to the last quartile for words. This pattern is demonstrated below:

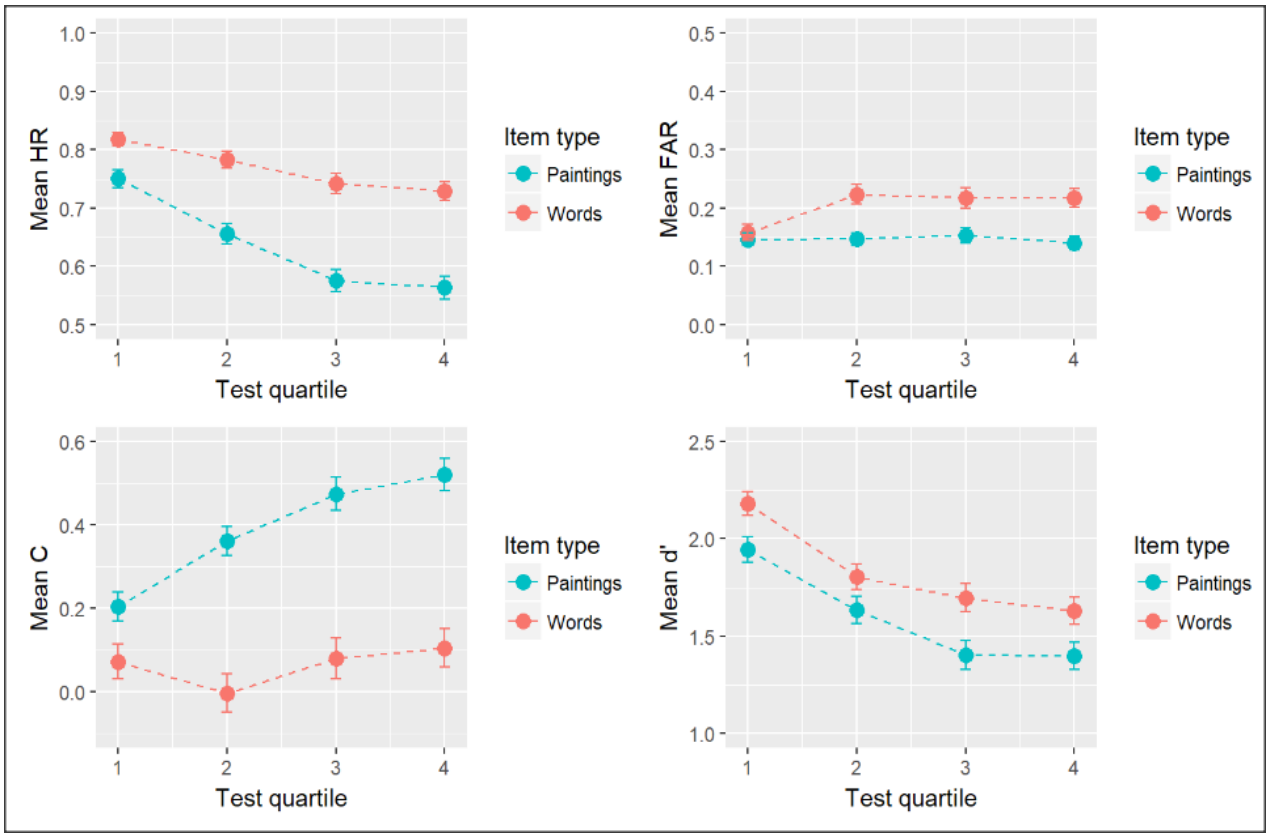


Figure 10: This plot (as constructed by Kaitlyn M. Fallow) shows the means, by test quartile, of hit rates, false alarm rates, response bias, and sensitivity for painting and word stimuli across the first within-subjects experiment conducted at Nagoya University

To illustrate, these are the quartile-level results based on averaging across seven previous experiments with Canadian participants:

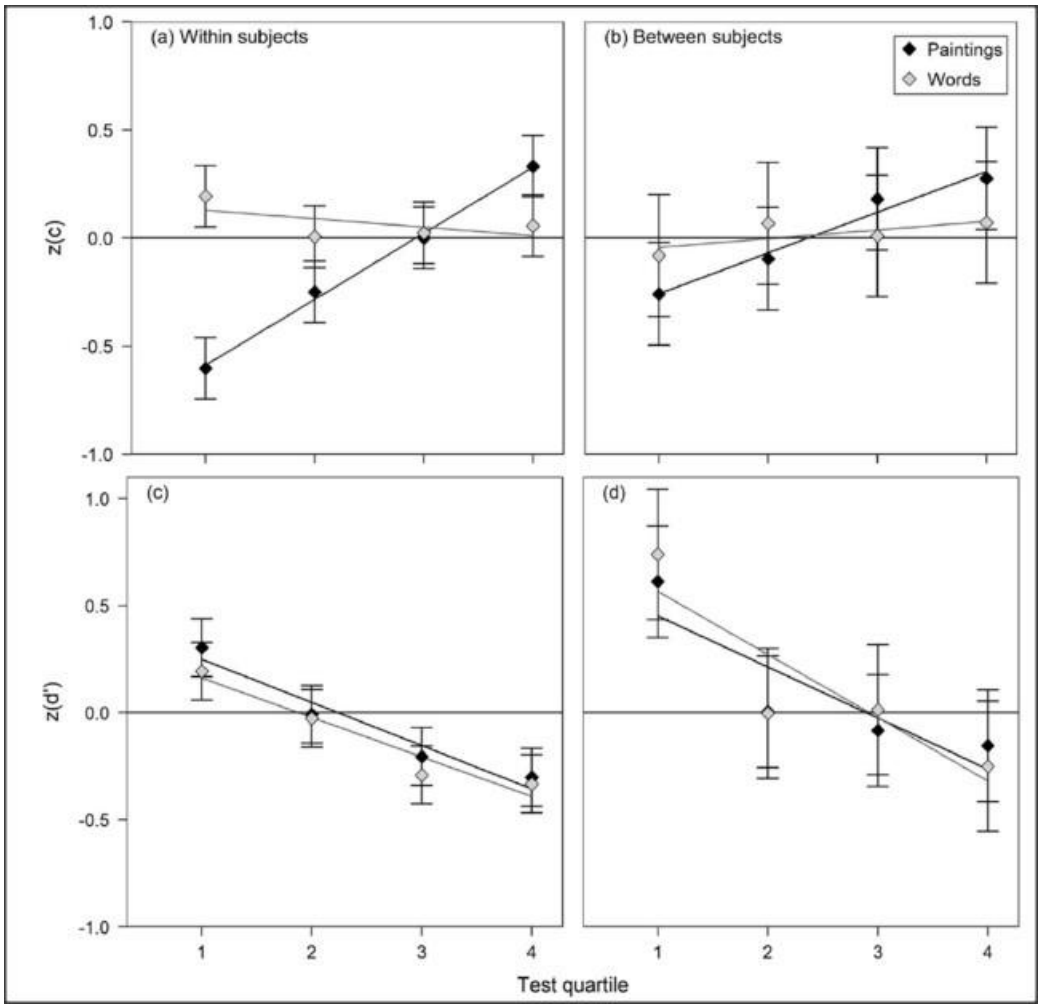


Figure 11: This figure shows the means by test quartile (and accompanying regression lines) of z-transformed response bias (c) and sensitivity (d') scores for painting and word stimuli across seven samples.

Figures 10 and 11reflect the patterns observed while averaging means across quartiles in the current experiment (refer to Figure 10) as well as the patterns observed in

prior experiments (refer to Figure 11). In Figure 11, z-transformed response bias (c) and sensitivity (d') scores for two types of stimuli, paintings and words, are displayed across seven different samples. The z-score transformation standardizes the data and allows for better comparison across quartiles of the experiments, highlighting changes observed over these quartiles. However, it is important to note that the z-score transformation, while useful for identifying changes across quartiles, does not directly reveal the basic direction of the response bias (whether it is conservative or liberal) or the sensitivity scores. It is obvious that the pattern observed in the current experiment is remarkably consistent with these overall trends noted in previous experiments at UVic.

There have consistently been materials-based differences in how responding varies over the course of the test phase in previous experiments with UVic students. The Japanese subjects demonstrated a conservative response bias on paintings, akin to Canadian undergraduates at UVic. However, while in the UVic studies, participants revealed a liberal bias when words and paintings were intermingled, in the Japanese data, the response bias for words was significantly conservative. This shows that although the Japanese subjects were much more conservative on paintings than on words, they were also slightly biased in the conservative direction when tested on words.

ROCs:

Equal variance violations are very common in recognition memory data (Yonelinas & Parks, 2007), and that is evident in the analyses of receiver operating characteristics (ROCs). The PCA-based method described by Vokey (2016) was used for all ROC curve fitting. Note that this method does not assume equal variance. The aggregate plots are shown here:

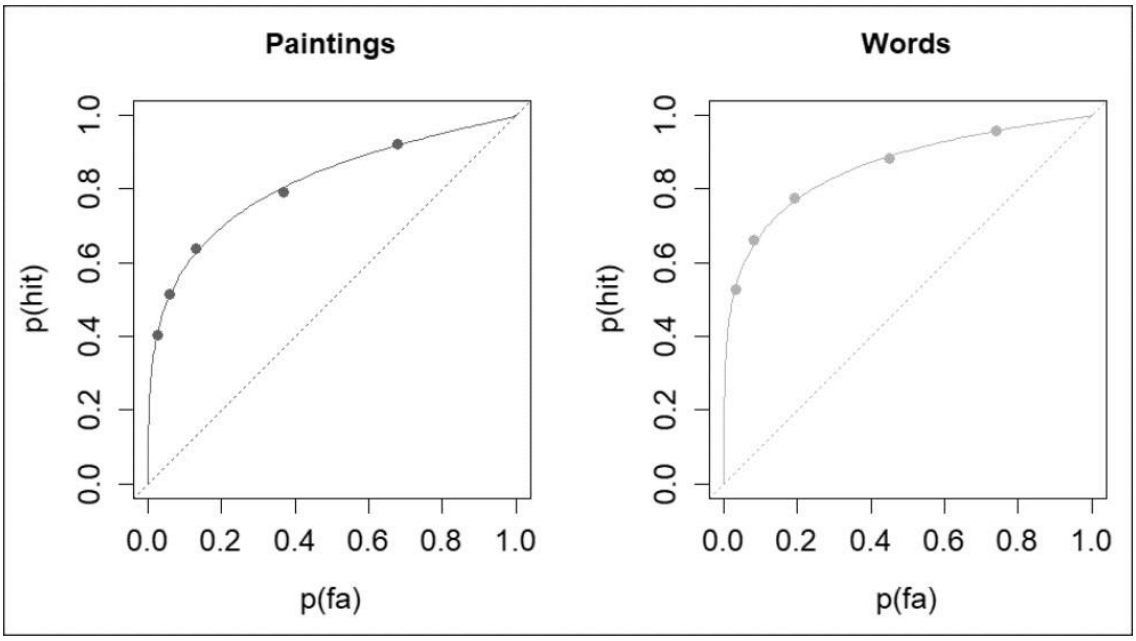


Figure 12: ROC curves for hits and false alarms of paintings and words as constructed by Kaitlyn M. Fallow
Note that the data presented here represents the Japanese participants' responses from Study 1.

In prior analyses of UVic data, the same trend has been constant. ROCs were constructed separately for paintings and words in three different ways to get a sense of the extent of unequal variance in this dataset (refer to Figure 12). In the first way, the data was collapsed across all subjects and items to create a single aggregate curve. While the second way collapsed the data for each subject across items, the third way collapsed the data across subjects for each item. The three calculations yielded similar curves. Figure 12 shows the curves calculated by PCA-based method described by Vokey (2016). Consistent with the d' analysis reported earlier, the area under the curve was greater for Words than for Paintings.

Lindsay and Kitigami and their collaborators conducted a follow-up study of the MBBE with Nagoya University students. Their study used similar materials but a between-subjects design. That is, some subjects studied and were tested on paintings, and others studied and were tested on words.

Study 2

Methods:

Participants: In this design, participants were 112 undergraduate students at Nagoya University in Japan (56 in the words condition and 56 in the paintings condition) who participated in exchange for bonus credit in a psychology course. Akin to the first study, the sample sizes resulted from testing as many participants as possible over the course of a single semester. Also, analyses were based on the full sample, as no participants met the preregistered exclusion criteria.

Materials: The materials were identical to those used in Study 1.

Procedure: Participants were informed at the beginning of the study phase that they would see a series of paintings or words presented one at a time on a computer screen and would be asked to try their best to remember each. Stimulus selection and order were randomized anew for each subject using E-Prime. Subjects received the test instructions followed by the test. In each test trial, a randomly sampled studied or new item from the appropriate set (i.e., painting or Kanji word) was presented for a recognition judgment on the same 6-point confidence scale as used in Study 1. In the paintings condition, during debriefing, participants had to rate their art expertise (from 1 = very below average to 5 = very above average).

Analyses:

To produce hit rates (HRs) and false alarm rates (FARs) for each participant, raw recognition test responses were collapsed into binary old/new responses (1-3 = “new”, 4-6 = “old”). These responses were then used to calculate signal detection measures of sensitivity (d’) and response bias (c).

No participant's performance fell below the exclusion criterion of d' <0.2, so all were included in the analysis. Notably, c and d' were calculated under the assumption of equal variance of the underlying "old" and "new" distributions (known to be regularly violated in recognition memory data; e.g., Yonelinas & Parks (2007)).

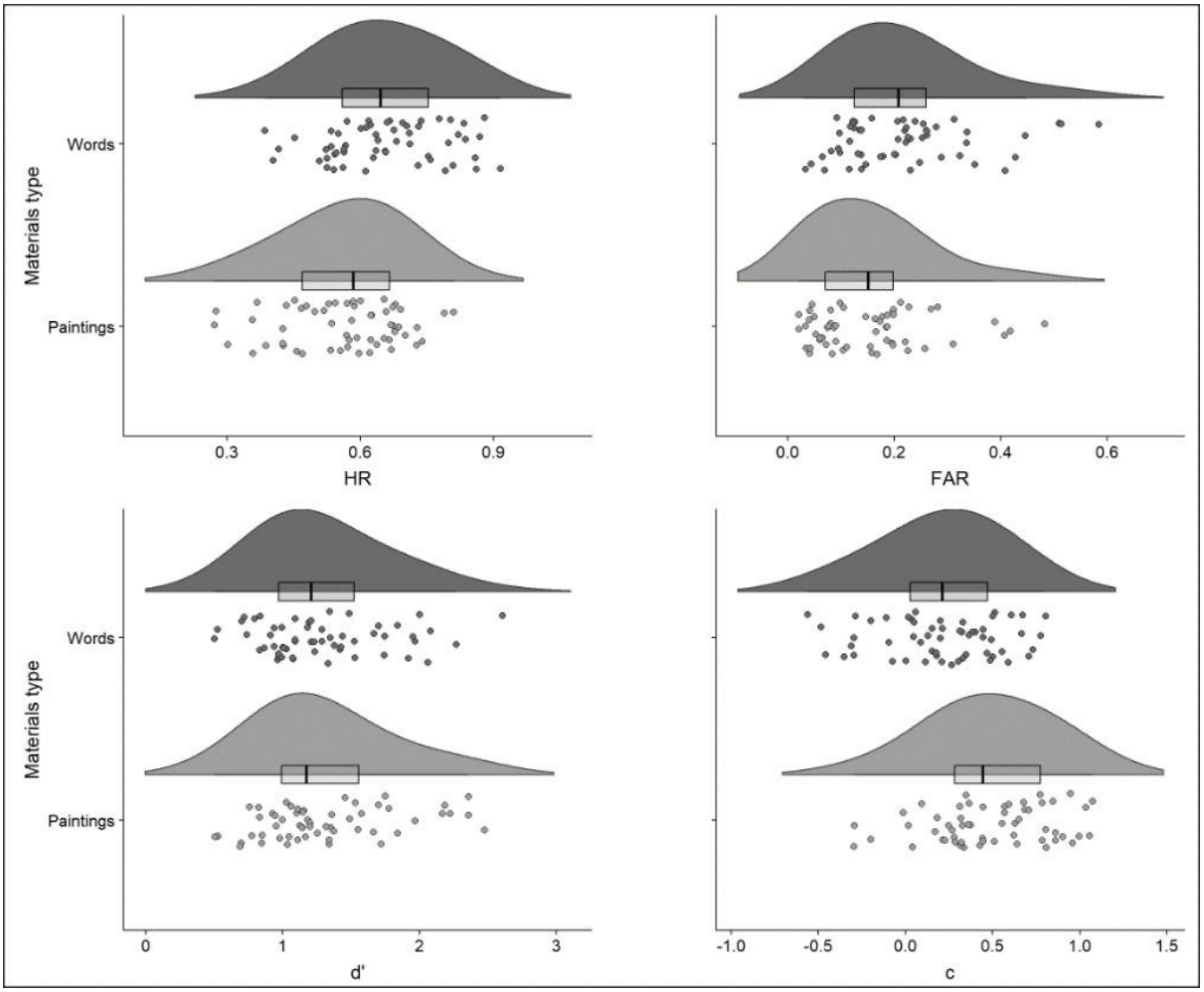


Figure 13: Raincloud plots for hit rates, false alarm rates, and standard equal variance signal detection (EVSD) measures of sensitivity (d') and response bias (c) as constructed by Kaitlyn M. Fallow. Note that the data presented here represents the Japanese participants' responses from Study 2.

Results:

Expectedly, response bias (c) was significantly conservative in the paintings group ($M = 0.48$, $SD = 0.34$; $t(55) = 10.78$, $p < 0.001$, $d = 1.44$ [1.06, 1.81]) (refer to Figure 13). Response bias was significantly more conservative in this group than in the words group, $t(109.99) = 4.32$, $p < 0.001$, $d = 0.82$ [0.43, 1.2]. These results align with the previous within-subjects study with Nagoya students as well as the usual pattern observed at UVic. As the former study showed a significantly conservative response bias in the words group, the same was anticipated here. This was precisely the case ($M = 0.21$, $SD = 0.34$, $t(55) = 4.62$, $p < 0.001$, $d = 0.62$ [0.33, 0.9]), showing a conservative bias in the words group. In the three UVic studies that included a words-only condition, bias was not significantly different from zero, in contrast to the pattern that is currently apparent in both studies conducted with Japanese undergraduates.

ROCs:

ROCs were constructed for the paintings and words groups using the method described by Vokey (2016). The recognition for words appears to be better than for paintings (refer to Figure 14).

Comparing the model for words to the model for paintings, the model performs better generally at distinguishing old and new stimuli, as indicated by the greater AUC for words. For words, the ROC curve is closer to the top-left corner of the graph (where true positive rate is 1 and false positive rate is 0) indicating a higher performance. This shows that for the word-based model, the true positive rate (sensitivity) is high and the false positive rate (1-specificity) is low. In other words, compared to the model for paintings, the model for words is better at properly distinguishing genuine positives and true negatives.

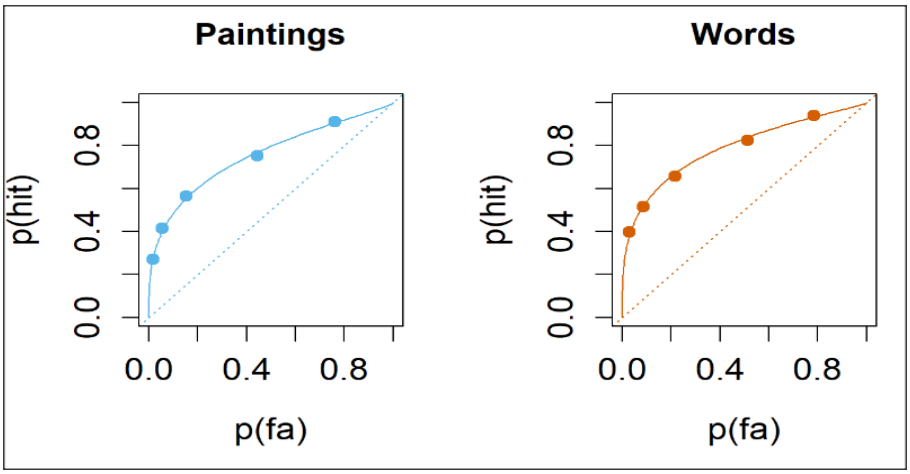


Figure 14: ROC curves for hits and false alarms of paintings and words as constructed by Kaitlyn M. Fallow. Note that the data presented here represents the Japanese participants' responses from Study 2.

Quartile-level analyses:

The 192 test responses were divided into 48-item quartiles to understand how responses varied over the course of the test (refer to Figure 15). Then, ceiling and floor HRs and FARs were replaced according to Macmillan and Kaplan (1985) to calculate sensitivity (d') and response bias (c) at this level.

Quartile-level data (with lines of best fit) is presented below:

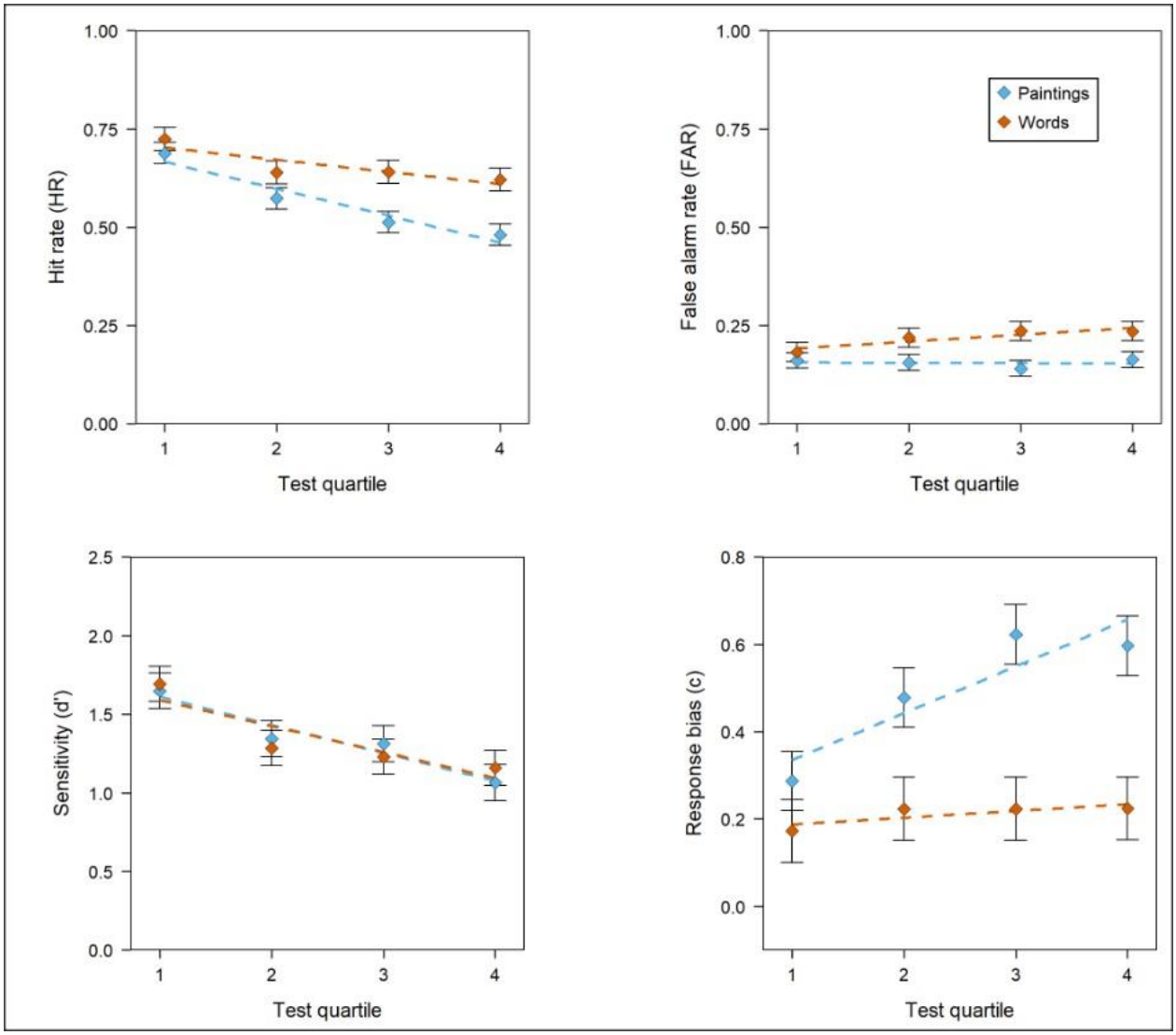


Figure 15: *Plot depicting the quartile-level analyses as constructed by Kaitlyn M. Fallow Error bars are 95% within-subjects confidence intervals (Loftus & Masson, 1994) calculated for each measure via separate one-way repeated measures ANOVAs for the words and pictures.*

The assumptions of researchers which were based on the results evident in the study were consistent with the results shown here. To elaborate, c increased significantly over the course of the test for paintings but not for words. On the other hand, sensitivity decreased similarly for the two material types. To clarify interpretation of these patterns and the results for raw response rates, data were subjected to 2 (materials type) x 4 (test quartile) mixed ANOVAs for the purposes of examining the interaction terms and accompanying effect sizes. Interactions were significant at the 0.05 level for c ($F(3, 330) = 6.7, p = 0.0002, \eta G^2 = 0.02$), HRs ($F(3, 330) = 6.03, p = 0.0005, \eta G^2 = 0.02$), and FARs ($F(3, 330) = 3.73, p = 0.01, \eta G^2 = 0.01$), but not for d' ($F(3, 330) = 1.08, p = 0.3571, \eta G^2 = 0.004$).

Limitations

It should be emphasized that whereas all paintings used in the current study have been previously used in UVic studies, the words in the current study were not carefully matched with those used in previous UVic studies. This is important because of the inconsistency of the effect of words on the variables that affect memory performance, such as frequency and imageability. To aid the interpretation of this difference in results obtained at the two universities, a finer-grained characterization of these word sets (and word-level analyses) may provide more accuracy. All in all, comparing the two studies conducted here, we can see a similar pattern to that observed in the within-subjects design. That is, the Japanese respondents were more conservative with paintings than with words. Interestingly, they nonetheless tended to be conservative, even on words. Canadian subjects, in contrast, have generally demonstrated neutral bias on words in the between-subjects design.

In summary, to understand the possible bias that a different culture might have, a within-subjects study was conducted with undergraduate students at Nagoya University in Japan using paintings and Japanese words as stimuli. These studies produced two clear and interesting results. First, contrary to our speculation, Japanese participants demonstrated a strong MBBE. They were strongly conservative when recognizing Western paintings. These studies were not designed to allow a head-to-head comparison of the Japanese samples to Canadian samples, but conservatism among Japanese participants was strong. Perhaps our idea about "highbrow" art somehow instilling conservative bias was wrong, or perhaps Japanese and Canadian participants viewed the paintings in a similar way. The second striking finding in these two studies was that Japanese participants tended to be conservative when recognizing words. Even in the within-subjects design, which has often yielded evidence of a liberal bias among Canadians, Japanese tended to be conservative.

Experiment 1

Background

The two studies conducted by Lindsay, Kitagami, and collaborators indicate that Japanese undergraduates, like UVic undergraduates, tend to be more conservative when making old/new recognition memory judgments of digital scans of paintings than of words. Thus, the tendency to be conservative when recognizing these Western paintings is not peculiar to Westerners. There was also a suggestion in the data that the Japanese students were more conservative than the Canadian students, both when recognizing paintings and when recognizing words. But the results only hinted at a cross-cultural difference in response bias.

Although the two Japanese MBBE studies were informative, we cannot draw definitive conclusions relying on these two studies alone. A major limitation of these studies is that the cultural comparison was done across experiments by comparing the new Japanese data with findings from earlier studies conducted in our lab. Another weakness is that the paintings were probably differently perceived by Japanese versus Western participants. To resolve this issue, Lindsay and Fallow decided to conduct follow-up studies using new materials that may be portable across cultures in that they represent unidentified objects. With Jamie Lee Gardens' help, Steve and Kaitlyn created new stimuli called "Diffeomorphs" through MATLAB, whereby original photos of everyday objects and settings were transformed into unidentifiable photos (refer to Figure 16). In other words, the stimuli are rich and distinctive, but they do not depict identifiable objects or scenes and therefore seem to be portable across cultures (refer to Figure 17).

Figure 16: This figure illustrated the process of transforming images where the image of a barn is transformed into a corresponding diffeomorph.

Figure 17: This figure illustrates eight diffeomorphs from the sample used in the experiment.

In a preliminary pilot study using this procedure and these stimuli with UVic students, false alarm rates were above floor, hit rates were below ceiling, and overall recognition accuracy was well above chance (but below ceiling). Most interestingly, the average response bias was neutral. This is an important and interesting finding because (like paintings) the diffeomorphs are novel to subjects, are colorful, and possess complex visual patterns. Yet average response bias was, in that small preliminary study, dead on neutral. In Experiment 1, the pilot study was replicated with a sample of undergraduate students from the American University of Beirut. The question to be asked in this study is: will participants from a different culture also show a neutral response bias?

Method

Subjects: In collaboration with Dr. Zahra Hussain at the American University of Beirut, we recruited 123 undergraduate Lebanese or Palestinian/Lebanese students. Recruited from the pool of students registered in introductory psychology courses, most of the students were bilingual (English-Arabic) or trilingual (English-French-Arabic) and they completed the study for optional bonus points in a psychology course. After exclusions, analyses were based on 108 participants (The exclusion criteria consisted of excluding participants with d' less than 0.2, less than 5 hits, and more than 15 false alarms.)

Materials: This study used 28 diffeomorphs, each based on a painting selected from the set used in earlier studies. We attempted to select diffeomorphs that were reasonably distinctive. We used Gorilla (https://gorilla.sc/) to program the experiment. which is accessible at https://app.gorilla.sc/openmaterials/617473.

Procedure: In this experiment, subjects studied and were tested on diffeomorphs. Study and test lists were randomized for each subject with respect to study status and study or test position. Each recognition study list contained 14 of these items, plus three primacy and recency buffers. Each test contained the 14 items from the preceding study list plus 14 new items. During the study phase, paintings were presented for 2 s each, with a 1-s blank interstimulus interval. Test responses were not speeded, and participants were asked to make studied or not studied judgments on a 6-point confidence weighted scale (from 1 = definitely not studied to 6 = definitely studied). This scale remained on screen throughout the self-paced test. At the end of the experiment, participants were asked whether they faced technical difficulties and whether they were focused on the experiment. On average, the duration of the experiment was about 15 minutes.

Design and Hypothesis: The independent variable was the old/new status of the test probe, and it was manipulated within subjects. The dependent variables were the hit and false alarm rates and the SDT measures of response bias (c) and sensitivity (d'). We hypothesized that the responses of subjects would be conservative when tested on diffeomorphs. Below I discuss the analysis, which includes the average measures of sensitivity and response bias for each study; comparable analyses of confidence response rating across trials; and confidence data, including but not limited to receiver operating characteristics (ROCs).

Analysis

Sensitivity and Response bias:

The calculation of SDT aggregates, response bias, and sensitivity was performed through R. All analyses of sensitivity and response bias in this research line have been based on the measures c and d'. With respect to c and d′ specifically, the sole advantage of these model-based measures over simpler ones, namely, the ability to separate the contributions of bias and sensitivity to observed response, only holds under a constrained set of assumptions about the underlying evidence strength distributions.

Descriptive Statistics

Table 1

Descriptive Statistics for c and d' (R)

	n	M	Lower CI	Upper CI
C	108	-0.09	-0.17	-0.01
dprime	108	1.43	1.30	1.56

Table (1) reflect the calculated statistics of the measures c and d'.

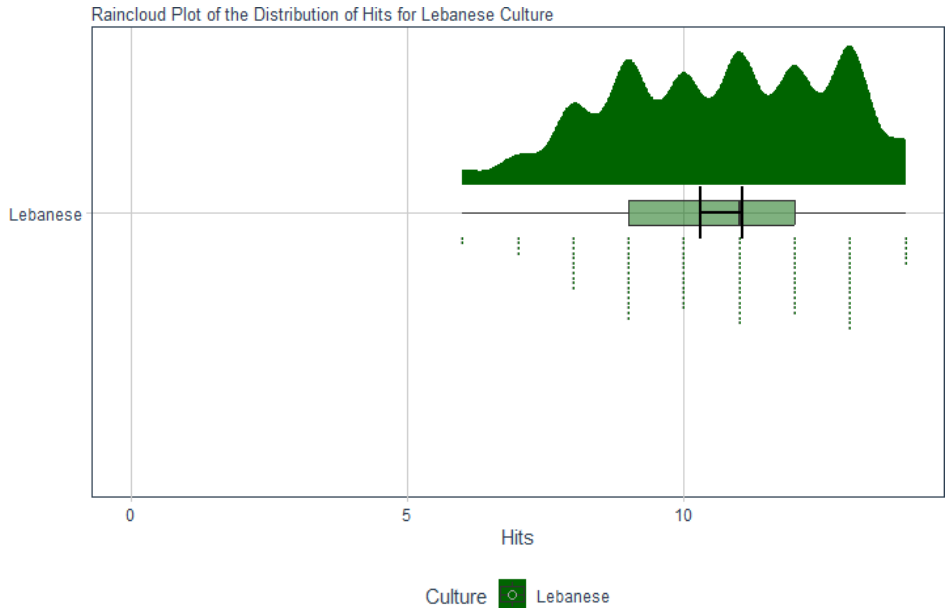


Figure 18: ***Raincloud plot visualizing the measure of hits for the Lebanese sample***

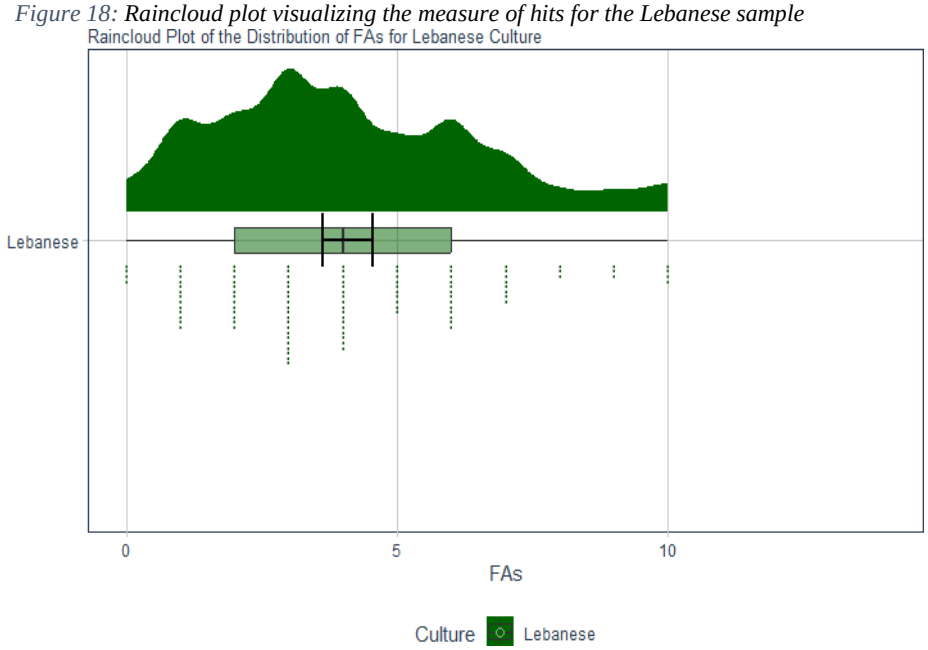


Figure 19: ***Raincloud plot visualizing the measure of false alarms for the Lebanese sample***

Figure 18 shows that subjects have the tendency to respond "yes" more frequently, regardless of whether the signal is present or absent. Figure 19 highlights that Lebanese subjects have a lower threshold for identifying the presence of a signal.

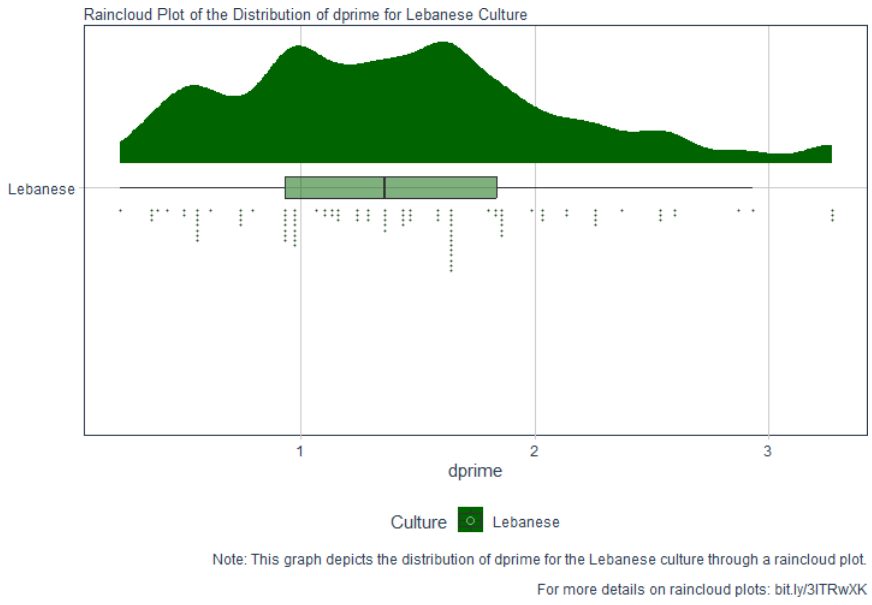


Figure 20: Raincloud plot visualizing the measure of Sensitivity (d') for the Lebanese sample

Figure 20 reveal the subjects' high accuracy- which we assume to be due to the low number of stimuli (this point will be further elaborated on in the next section where we introduce Experiments 2 and 3).

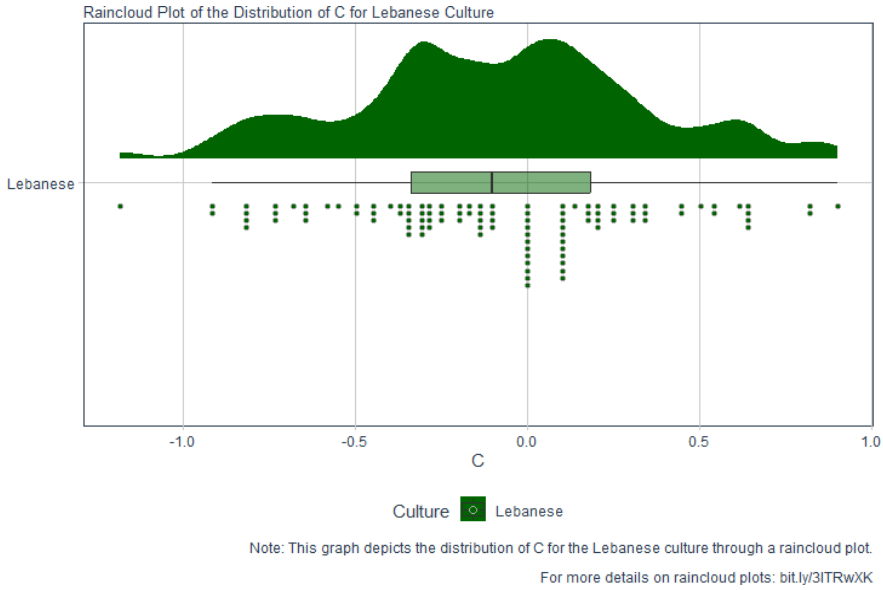


Figure 21: Raincloud plot visualizing the measure of response bias (c) for the Lebanese sample

As shown in Figure 21, we notice that the values of c tended to average somewhat below the point zero, thus indicating that subjects have a liberal response bias. To measure response bias statistically, we ran a one-sample t-test t. The one-sample t-test indicates that (t (107) = -2.13, p = 0.04, 95% CI [-0.17, -0.006], Hedges' g = -0.20, CI [-0.39, -0.01]. This suggests that the observed response bias for the diffeomorphs is statistically different from 0. In other words, response bias for diffeomorphs shows a tendency of being closer to liberal than it is to neutral among Lebanese subjects.

Receiver operating characteristics Analysis

One of the most useful means of presenting confidence data is in the form of receiver operating characteristic curves (ROCs), which allow for several conclusions about the structure underlying the data. The ROC is constructed by plotting hit and false alarm pairs beginning with the most confidently recognized items and then repeatedly

recalculating the values by including the next most confidently recognized items. Then, the slope and intercept of the best-fit line (to the cumulative data expressed in z-coordinates) are estimated.

Methods for obtaining the best-fit line in the z-coordinate space have varied, and one recent one is using PCA (Vokey, 2016). PCA plotting has been resorted to due to the iterative maximum likelihood fitting algorithms implemented in the analysis, which allow for overcoming the inadequacy of least-squares solutions. The code provided by John Vokey (2016) computes the components' fit to the data, and then plots the ROC in the unit-square in the p-coordinate. It is also used to compute the empirical proportional area under the ROC curve (AUC) in a non-parametric way.

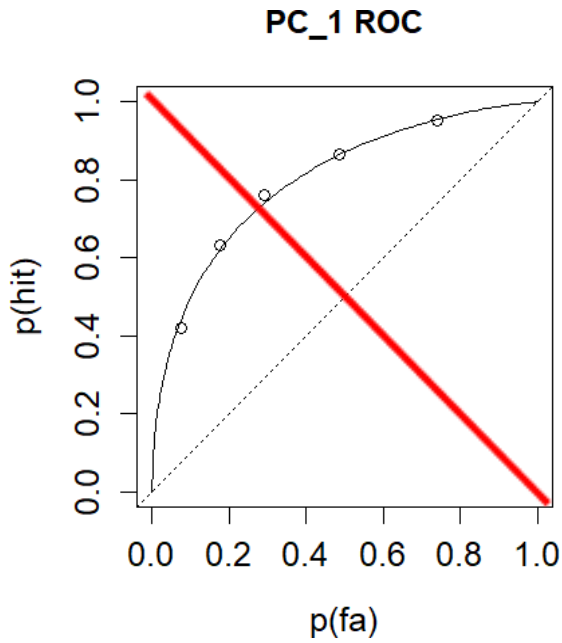


Figure 22: ***ROC plot***

Figure 22 shows the graphical depiction of a binary classifier system's performance when its discriminating threshold is changed in this plotted ROC. For various threshold values, it shows the hit versus the false alarm rates. According to the ROC curve, the hit rate appears to rise as the false alarm rate rises. This shows that if the classification threshold is dropped, the model may properly identify more hits, but at the expense of endorsing more false alarms.

The model showed a moderate degree of predictive power, as evidenced by the area under the curve (AUC) of 0.67. An ROC curve that follows a diagonal where the hit and false alarm rates are equal are typical characteristics of a model that performs no better than random guessing with an AUC of 0.5. When a model's AUC is less than 0.5, it performs worse than random guessing, whereas an AUC of 1.0 indicates flawless prediction. This implies that while the model is superior to random guessing, there is still potential for improvement, as indicated by an AUC of 0.67. The diagonal line in the provided ROC curve is far off the curve, showing that the model is outperforming chance. All in all, the binary classifier has some predictive capacity overall, according to the ROC curve, but there is potential for improvement in terms of lowering the false alarm rate and raising the hit rate. Although it may correctly identify a higher proportion of hits than false alarms, it could also generate false alarms at a relatively high rate. The plotted ROC, through a PCA approach, resembles the indications of the graphical distributions as well as the statistical testing. It showed that the Lebanese subjects had neither a conservative nor a liberal bias. In brief, besides measuring response bias and sensitivity, a refined understanding of performance encompasses plotting an ROC to visualize these distributions.

Experiments 2 & 3

Background

Our work is influenced by prior studies in the lab that sought to construct a conceptual understanding of the mechanisms underlying response bias and accuracy of participants while studying different types of stimuli. In this line of work, I collected data and conducted initial analysis under the title of "Generalizing the MBBE". And what is precisely meant by generalizing is, essentially, generalizing the conservative bias observed for paintings for stimuli like "diffeomorphs" (assuming the similarity between paintings and the diffeomorphs) and more importantly the use of such stimuli for generalizing the effects observed across cultures.

As mentioned in the prior section, I have collected data of 112 Lebanese participants using diffeomorphs as stimuli. While participants showed a liberal bias for the 14 diffeomorphs that they were tested on, this bias tended to lean towards being conservative across trials. To extend my data collection, I recruited Canadian participants who studied the same stimuli (40 diffeomorphs). Notably, in this study, I have modified my experiment by increasing the number of stimuli to 20 at study, and 40 at test to get an accurate estimate of performance while using a fair number of stimuli. Then I recruited Japanese participants who studied the same stimuli (40 diffeomorphs). In Experiment 2, the sample was recruited from the SONA system of psychology students at UVic. In Experiment 3, we collaborated with Dr. Satoshi from Nagoya university to recruit our participants. The questions to be asked in this study are: will Canadians' response bias on diffeomorphs be neutral on average (as it was in a small pilot study with Canadians and a full sample of Lebanese using a smaller set of stimuli)? and (2) will Japanese participants reveal a conservative bias on average, as they did in the prior studies with painting and

Kanji words?

Method

Canadian Subjects: We recruited the maximum number of 181 participants we could get between October and December 2022 from the SONA system at UVic. Subjects were undergraduate students aged between 18 and 25. Recruited from the pool of students registered in introductory psychology courses, most of the students completed the study for optional bonus points in a psychology course. After exclusions, analyses were based on 170 participants. We lack complete data on the cultural/ethnic/national backgrounds of the "Canadian" sample but based on the demographics of the University of Victoria it is likely that most but not all were born in Canada and were of Western European descent.

Japanese Subjects: We recruited the maximum number of 147 Japanese participants we could get between December 2022 and January 2023 with the help of our collaborators, Dr. Moriizumi Satoshi at Nagoya University and Dr. Yukio Itsukushima at Nihon University. Subjects were undergraduate students aged between 18 and 25. Recruited from the pool of students registered in psychology courses, the students completed the study for optional bonus points in a psychology course. After exclusions, analyses were based on 129 Japanese participants.

Materials: This study used 40 total diffeomorphs, each based on a photo selected from the set used in earlier studies. We attempted to select diffeomorphs that were reasonably distinctive. We used Gorilla (https://gorilla.sc/) to program the experiment, which is openly accessible at https://app.gorilla.sc/openmaterials/617473. Additionally, a significant effort was made to ensure that the Japanese version of the materials maintained a high fidelity to the original English version.

To achieve this, we employed meticulous translation processes through local cultural consultation. This helped in ensuring that the nuances of the content remained intact across the two language versions.

Procedure: The same procedure was applied here except that each test contained the 20 items from the preceding study list plus 20 new items. On average, the duration of the experiment was about 25 minutes.

Design and Hypothesis: We hypothesized that the responses of Japanese subjects would be conservative when tested on diffeomorphs, whereas those of Canadians would be neutral. Below I discuss the analysis, which includes the average measures of sensitivity and response bias for each study; comparable analyses of confidence response rating across trials; and confidence data, including but not limited to receiver operating characteristics (ROCs).

Analysis

Exclusions: The total Canadian sample collected was 181 participants. We had pre-set exclusion criteria that focused on excluding participants with d’ less than 0.2 (as pre-registered). After piloting a sample of 20 students, we decided to further exclude participants with fewer than 5 hits and more than 15 false alarms. Kindly note that we included the sample piloted. The total number of participants was 170 after exclusions. Similarly, the total Japanese sample collected was 147 participants. After exclusions, and including the piloted sample, the total number of the Japanese sample was 129 participants.

Descriptive Statistics

Table 2

Descriptive Statistics for c and d'- Canadian Sample (R)

	n	M	Lower CI	Upper CI
c	170	0.02	-0.03	0.08
dprime	170	1.23	1.13	1.32

Table 3

Descriptive Statistics for c and d'- Japanese Sample (R)

	n	M	Lower CI	Upper CI
c	129	0.18	0.29	0.65
dprime	129	1.04	0.94	1.13

The tables (2 & 3) reflect the calculated statistics of the measures c and d' of Canadian and Japanese samples.

Sensitivity and Response bias

Akin to the first experiment, the calculation of SDT aggregates, response bias, and sensitivity was performed through R. All analyses of sensitivity and response bias in this research line have been based on the measures d' and c.

Histograms of hits for Canadian and Japanese samples:

Hit rates tended to be statistically significantly higher in the Canadian than in the Japanese sample: $t(271.5) = 5.02$, $p < 0.001$, CI [1, 2.29] , Hedges' $g = 0.59$, CI[0.35, 0.82] (refer to Figure 23).

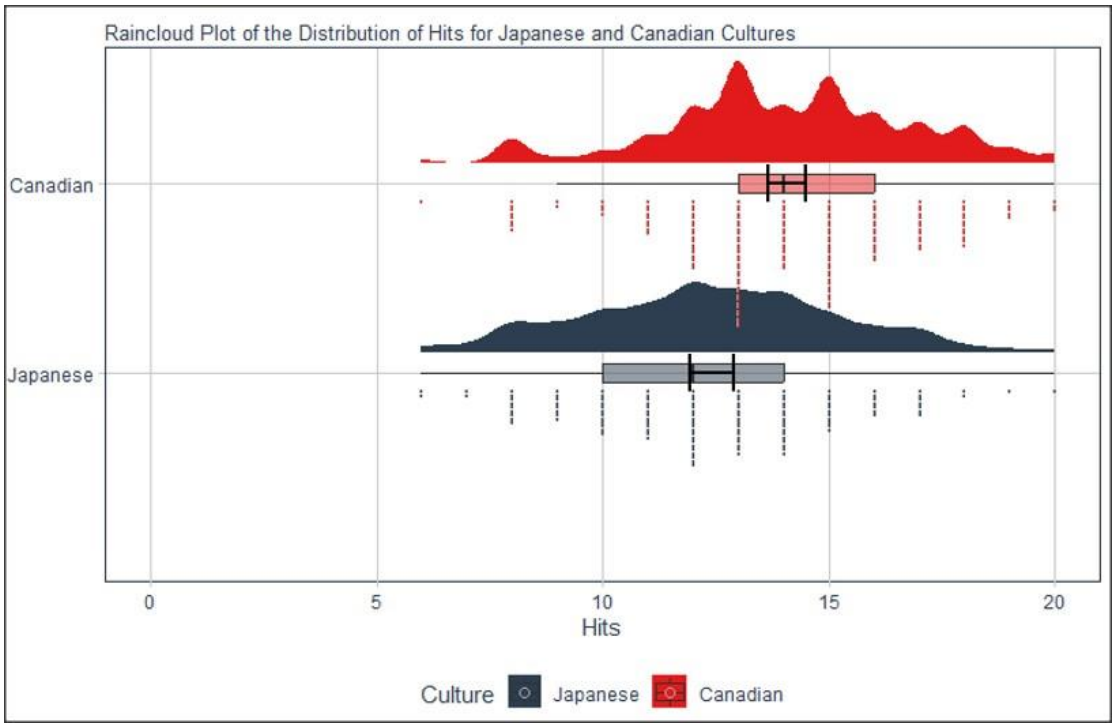


Figure 23: Raincloud plot of hits for Canadian and Japanese Cultures. Note: This graph depicts the distributions of Hits for the two cultures through raincloud plots, including 95% confidence intervals and error bars. For more details on raincloud plots: bit.ly/3ITRwXK

Histograms of false alarms for Canadian and Japanese samples

False alarm rates in the Canadian and Japanese samples did not significantly differ: $t(278.78) = 1.13$, $p = 0.26$, CI [-0.26, 1.06], Hedges' $g = 0.13$, CI [-0.10, 0.36] (refer to Figure 24).

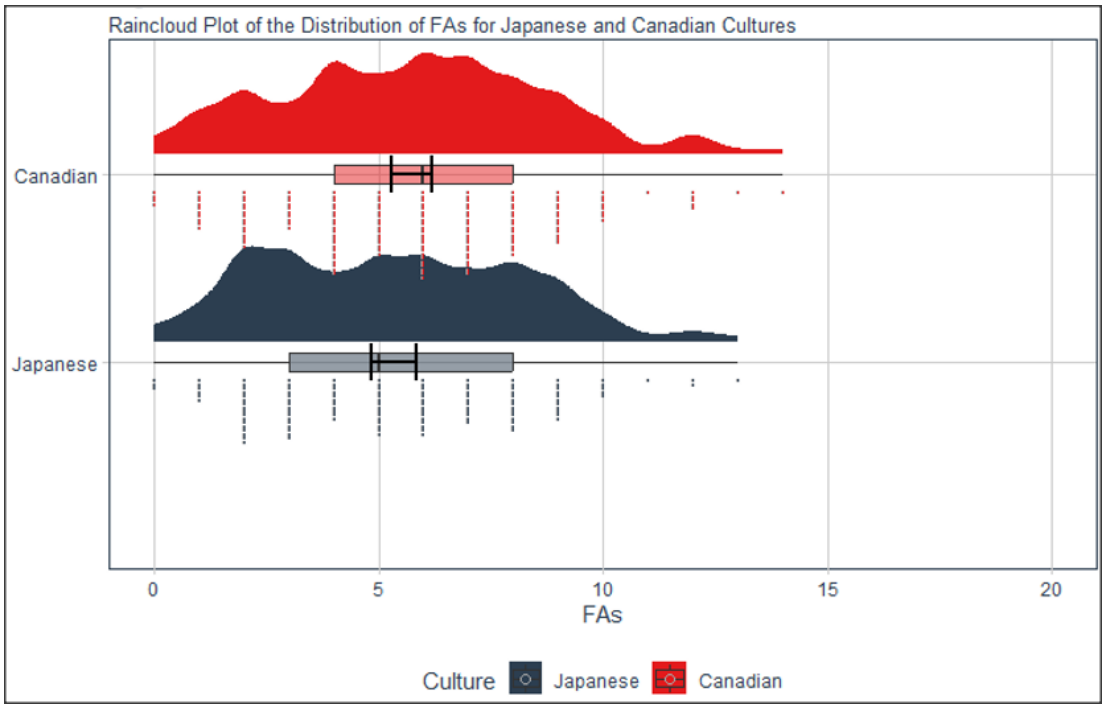


Figure 24: ***Raincloud plot of False Alarms for Canadian and Japanese Cultures***

Histograms of Sensitivity (d') for Canadian and Japanese samples

Recognition accuracy indexed with d' was statistically significantly higher in the Canadian sample than in the Japanese sample: $t(293.6) = 2.81$, $p < 0.001$, CI [0.06, 0.32] , Hedges' $g = 0.32$, CI[0.09, 0.55] (refer to Figure 25).

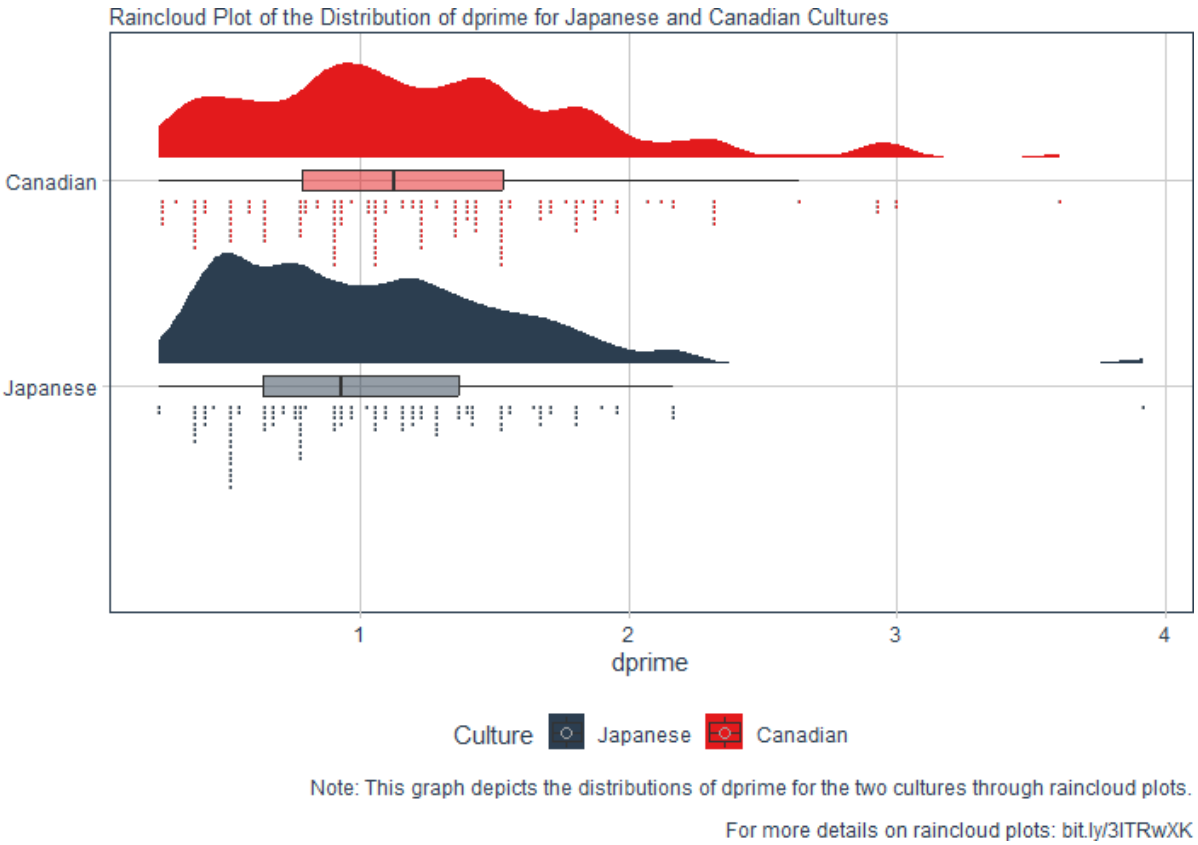


Figure 25: Raincloud plot of Sensitivity (d') for Canadian and Japanese Cultures

Histograms of Response bias (c) for Canadian and Japanese samples

Response bias was statistically significantly greater (ie., more conservative) in the Japanese sample than in the Canadian sample: t (271.68) = 3.58, p <0 .001, CI [0.07, 0.24] , Hedges' g= -0.41 , CI[-0.65, -0.18] (refer to Figure 26).

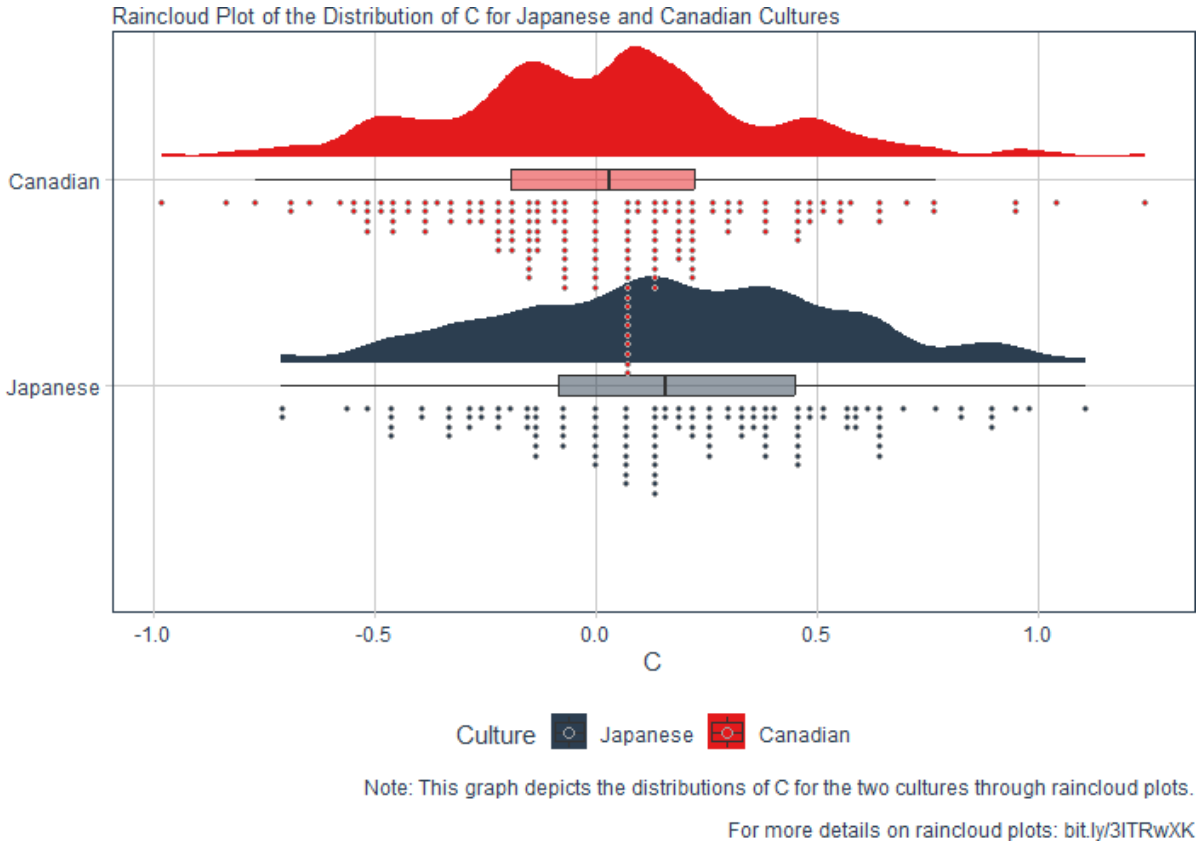


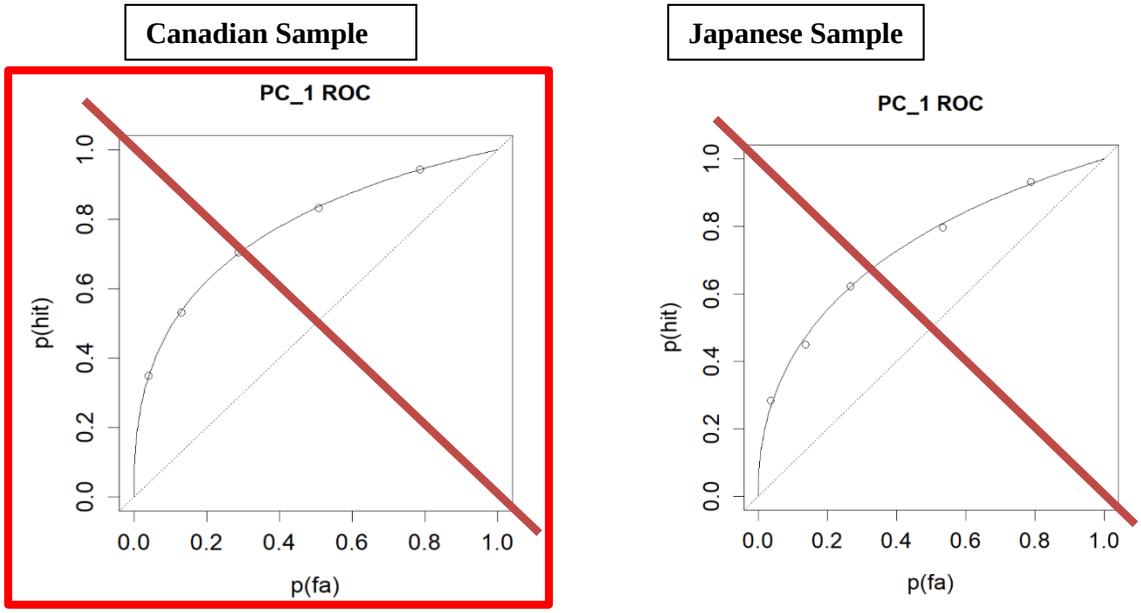
Figure 26: *Raincloud plot of response bias (c) for Canadian and Japanese cultures*

As the figures show, Japanese participants were significantly more conservative than Canadian participants. For the Japanese data, a single-sample t-test indicated a value of c that was significantly greater than zero: $t(128) = 5.38$, $p < 0.001$, CI [0.11, 0.24], mean = 0.18, Hedges' $g = 0.47$, CI[0.29, 0.65]. For the Canadians, in contrast, c did not significantly differ from zero; $t(169) = 0.82$, p = 0.42, CI [-0.03, 0.08], mean = 0.02, Hedges' g = 0.06, CI [-0.09, 0.21].

Compared to Japanese participants, Canadians more often correctly judged studied items as studied, whereas the two groups did not differ significantly in false alarm rate (although the small and nonsignificant difference was in the direction of more false alarms among Canadian than Japanese participants). To the best of our knowledge, this is the first evidence of a cultural difference in recognition memory response bias.

Receiver operating characteristics Analysis

Equivalent to experiment 1, we attempted to visualize the data by carrying out an ROC, with the x-coordinate defined by the false alarm rate, and the y-coordinate defined by the hit rate. Akin to the first experiment, the ROC was constructed by plotting hit and false alarm pairs beginning with the most confidently recognized items and then repeatedly recalculating the values by including the next most confidently recognized items. We also used PCA plotting to overcome the inadequacy of least-squares solutions (Vokey, 2016).



'Figure 27: ROC plot for the samples.

In an ROC, the points represent performance at different levels of the decision criterion. In these plotted ROCs, we used our usual 1-6 confidence ratings as a proxy for different decision criteria. Thus, the very first point on this plot shows what the hit and false alarm rate would be at the most conservative criterion in the experiment, that is, if we only counted a response as "old" if it was given the highest-confidence rating (6 = definitely studied). The more liberal the bias, the farther the first point is from the origin. Overall, the curves do not appear farther up in the upper left. However, they are above chance as evidenced by an AUC of 0.71 for the Canadian sample and 0.67 for the Japanese sample (refer to Figure 27). Concerning symmetry, if we were to fold the curves in half along the diagonal axis, we would notice that they are close to being symmetrical, reflecting the response criterion. The ROC curves suggest that the models have better sensitivity than specificity. In other words, the models are better at correctly identifying hits than they are at correctly identifying misses. The ROCs gave a more accurate visualization of the results of both samples.

In this experiment, we have tried to replicate the same study we did in Experiment 1 on a Canadian and Japanese sample while increasing the number of stimuli. Besides interpreting aggregate measures of signal detection theory like response bias (c) and sensitivity (d'), a multi-level model was constructed to explore the progressive trend of responses (measured using a 1-6 confidence-weighted scale) across several trials. Further, we constructed the ROC for the diffeomorphs following the approach described by Vokey (2016).

Discussion

Introduction

The interest in studying materials-based differences cross-culturally started with the effort of inspecting cultural differences that could exist in recognition memory. The initial investigation in Lindsay's lab began with a collaboration between Dr. Steve Lindsay and Dr. Shinji Kitagami at Nagoya University in Japan. Recruiting undergraduate students showed that respondents from Japan exhibited a conservative response bias for paintings and a more conservative response bias than for words, which is consistent with findings from UVic. Intriguingly, in contrast to within-subjects research carried out at UVic, Japanese students also demonstrated a conservative response bias for words.

Students from Nagoya University participated in a follow-up between-subjects study to investigate material-based disparities. The outcomes were in line with the within-subjects investigation. The word set was not adequately matched with earlier trials, which could have affected memory-related variables and limited the study. For paintings, the study discovered a widespread phenomenon of a conservative reaction bias; however, the results for words demand further study because it is difficult to match psycholinguistically significant characteristics across languages. Those results cannot confirm a cross-cultural difference and further research is needed.

To understand the interactions of elements that give rise to materials-based differences, we aimed to use materials that lack the semantic meaning of paintings but still carry the novelty, complexity, and colorfulness elements, so we chose the diffeomorphs, which are based on distortions of the color versions of photos of Snodvan items. Notably, my colleague, Dr. Fallow, had already tested a Canadian sample's response bias on

diffeomorphs finding that UVic students appeared to be neutral on average for those entirely novel and rich stimuli.

Experiments

We conducted Experiment 1, a pilot study with 123 undergraduate Lebanese students from the American University of Beirut, to investigate whether participants from a different culture would also exhibit a neutral response bias. The stimuli used here were 14 diffeomorphs in the study phase and 28 in the test phase. The analysis focused on the average measures of sensitivity and response bias (d' and c) for each study, and confidence data, including receiver operating characteristics (ROCs). This was aimed at understanding the cultural influence on response bias and potential variations in response patterns.

The results indicated a liberal response bias (c) for diffeomorphs among Lebanese subjects, as the mean was statistically different from 0. In Experiments 2 and 3, 170 Canadian and 129 Japanese participants were recruited to examine whether participants from different cultures would also show a liberal response bias. The hit rates were found to be significantly higher in the Canadian sample compared to the Japanese sample. This showed that Canadian participants were more likely to correctly recognize studied items than Japanese, reflecting differences in response bias between the two cultures. False alarm rates in the Canadian and Japanese samples did not significantly differ.

Recognition accuracy was statistically significantly higher in the Canadian sample than in the Japanese sample, while response bias was statistically significantly greater in the Japanese sample than in the Canadian sample. The response bias measures indicate that Canadian participants showed a neutral response bias, which is consistent with the previous studies piloting Canadian (led by my colleague Dr. Fallow), wherein subjects

exhibited neutral response bias when tested on diffeomorphs. On the other hand, Japanese participants showed a conservative response bias, which does not support the notion that subjects could exhibit a neutral response bias when tested on diffeomorphs. Experiments 2 and 3 provide further evidence that response bias in recognition memory tests can be influenced by cultural differences. A thorough observation of the variance in biases between the two populations revealed that the difference was primarily driven by a difference in hit rates. Canadians more often correctly judged studied items as studied compared to Japanese participants, while the two groups did not differ significantly in false alarm rate. This is the first evidence of a cultural difference in recognition memory response bias. The ROC analysis of the Canadian and Japanese participants showed that the data was better at accurately recognizing hits than at correctly identifying misses. It was clear that the Canadian subjects were neither conservative nor liberal.

Linking my results with present cross-cultural research

The significance of my results hints at cultural differences in recognition memory and provides an enhanced understanding of the MBBE. In attempts to examine cultural variations in memory, researchers have thought that cultural differences in cognition arise from differences in holistic versus analytical styles of information processing and differing levels of attention to the field versus the object (Gutchess & Sekuler, 2019). According to Miyamoto, Nisbett, and Masuda (2006), culture can affect how people perceive their physical surroundings. For example, people from individualist cultures tend to take a more analytical approach, whereas those from collectivist societies prefer to take a more holistic perspective.

The distinction between analytic and holistic processing styles has been suggested to account for cross-cultural differences in attention and memory for complex

information. Evidence suggests that Westerners describe background information less than Easterners (Masuda & Nisbett, 2019). Perhaps the cultural background helps in prioritizing details and objects for Westerners and context and relationships for Easterners. Contrasting research shows that Canadians show better background recognition compared to Chinese (Wong et al., 2018). The difference in preference for detail could be due to response bias differences, wherein the more specific memory of Westerners simply reflects a preference for detail.

Cultural differences in memory for details may be primarily a function of response bias. Evidence suggests that Americans have a stronger tendency to claim that information is presented with the same details as items seen before (Paige et al., 2017). Also, there are cultural differences in the specificity of the memories, with Caucasian Americans tending to recall more specific episodes that occurred at one moment in time, while Asian Americans recall more general memories of recurring events (Wang & Ross, 2005). Our results support the notion that Westerners or individualist cultures show better sensitivity than collectivist cultures, as evident in the distinction of the (d') measure between our Canadian and Japanese samples. To elaborate, Canadians seem to show a significantly higher sensitivity compared to Japanese when providing recognition judgments of diffeomorphs. It is important to point out that mixed findings show that multiple mechanisms contribute to cultural differences in memory specificity, such as perceptual mechanisms and stable cultural values (Gutchess & Sekuler, 2019).

Studies have investigated how cultural differences affect how precisely people recall details about objects. Americans are more aware of the specific visual characteristics of the object (such as specific recall if the exact same picture of a car shows on the recognition test), although East Asians and Americans may both recognize

that they previously examined a picture of a car (Millar et al., 2013). It seems that generic recall for objects is the same throughout cultures. In other words, regardless of the specific perceptual characteristics, Americans and East Asians are equally able to recognize that a car was previously researched.

When accounting for items presented with or without a background (Millar et al., 2013), individual ratings of emotional intensity or congruency of the items and context, as well as information that was neutral or emotional (Steinmetz et al., 2017), and encoding under different instructions (Paige et al., 2017), robust cultural differences in memory specificity emerge. Despite evidence showing that cultural differences only emerge in cases where stimuli are emotionally stimulating or instructions are modified, our results have shown otherwise. In our experiments, we have not looked at memory specificity in the sense that we did not precisely test for cultural differences in the remembrance of details. Even though our review of past research has shown that the general recall of items is universal across cultures, we have seen that there is a clear difference in that; Canadians are neutral while Japanese are conservative when recalling diffeomorphs. This has been supported by Millar et al. (2013), wherein Americans exhibited better levels of recognition for previously viewed "same" things than East Asians. Even though the Lebanese study included only 28 diffeomorphs at the test (in contrast with 40 as in Experiments 2 and 3), the response bias of our participants seemed to be liberal, in contrast with a neutral response in the Canadian sample and a conservative one in the Japanese sample. We would have expected that the Lebanese participants would perform like the Japanese, given that the two cultures are collectivistic. Yet, we cannot draw conclusions based on that because the Lebanese study was a pilot study, and perhaps the smaller number of items led to a great recognition performance for our sample.

Although environmental aspects have not yet been investigated, it is reasonable to assume that environmental influences contribute to the focus on detail. By exposing Americans to densely populated scenes, priming can momentarily refocus attention within situations (Miyamoto et al., 2006). Developmental data and environmental affordances can explain cross-cultural variations in memory specificity. According to a developmental research study, Japanese children demonstrate a greater relational emphasis at the ages of 3 and 4, whereas American children at that age are more object-focused and sensitive to details (Kuwabara & Smith, 2016). It is essential to take environmental affordances into account, for how the physical environment is structured highlights what information is important based on what stands out as novel and distinctive and what blends into the background. The number of objects in outdoor Japanese scenes tended to be higher than in American scenes, according to a comparison of photographs taken in small, medium, and large cities in the United States and Japan (Miyamoto et al., 2006). This research implies that the sheer number of items in the Japanese scenery diverts attention away from specific objects.

Even though there seems to be research supporting the existing contrasts between Westerners' attention to detail and Easterners' heightened attention to background, there seems to be contradicting evidence. For instance, a study that aimed to test memory accuracy across cultures found that memory accuracy was similar across cultures, such as Americans and Turks (Schwartz et al., 2014). Another study studying Australian indigenous and non-indigenous children when completing visual working memory tasks found that, contrary to expectations, Indigenous children demonstrated greater overall mean accuracy and detection sensitivity, compared to non-Indigenous children, on a change detection working memory task (Freire et al., 2020). This shows that even though

there are a limited number of studies on cross-cultural memory differences, the evidence is inconsistent. More importantly, the work presented comparing response bias differences between the Canadian and Japanese samples matches trends in the literature.

Linking my results with present theoretical knowledge on the MBBE

The significance of the results provides an enhanced understanding of the MBBE and cultural differences in recognition memory. The first objective aimed at a better understanding of the cultural differences, and below we outline our efforts in explaining the materials-based bias effect based on the findings. We have discussed elaborately that paintings have complex and varied perceptual and semantic properties, which have been shown to affect recognition performance in memory, thus explaining the conservative bias on an old/new recognition memory test. Theories have been examined to explain the causes of conservatism in paintings.

The subjective memorability theory (Brown, Lewis, & Monk, 1977), the pre-experimental familiarity with the stimuli (Hintzman, 2009), and the reminding hypothesis, which were suggested as explanations for the MBBE, were not supported (Lindsay & Kantner, 2011). The uniqueness heuristic offers another explanation, but it was also disproved in a recent study by Fallow and Lindsay (2022) In these two experiments, 100 subjects (per group) studied and were tested on words, line drawings, or grayscale photos representing the same set of objects (Snodgrass & Vanderwart, 1980). Experiment two was a direct replication of Experiment one. Recognition memory response bias was neutral for words and line drawings but conservative for photos depicting the same common objects. A neutral response bias identified for both words and line drawings raised the possibility that the semantic meaning of paintings alone may not fully explain the materials-based bias effect. Using the diffeomorphs as stimuli for study has

additionally raised an important indication: that colorfulness and novelty cannot account for a conservative bias. Even though we have used rich stimuli with Canadians, they did not show a conservative bias towards these stimuli, which are characterized by patterns lacking any meaning.

Limitations

The MBBE presents crucial issues regarding the effect's applicability to all visual stimuli and its interplay with cultural variations. Our results have contributed to a better understanding of the MBBE as well as new knowledge regarding cultural variations in response bias to different stimuli. Although our results hint at cross-cultural differences in response bias and sensitivity, more investigation is required to determine how culture influences response biases in recognition memory tasks. Our knowledge of the relevant cognitive processes and the possible effects of cultural variations on cognitive performance will be improved as a result.

Future directions

Following our findings, the first step would be to test the replicability of the pattern. In other words, we want to see if new samples of Canadian and Japanese participants with the same materials and procedures yield the same general pattern of results. The demonstrated statistical analyses conducted in this thesis indicate that it would be unlikely to get much evidence of conservative response bias in a Japanese sample compared to a neutral bias for the Canadian sample (using the same materials and procedure). This is merely a speculation that would require further data collection and analysis to be disproved.

Furthermore, we want to confirm if the pattern obtained with new data is replicable. Besides, it would be crucial to see whether the Japanese tendency toward eliciting a conservative response bias for stimuli such as diffeomorphs is robust or not. We already have, from the earlier experiments, indications that a conservative response for Japanese is observed in various kinds of materials (words, paintings, and diffeomorphs), but it would be good to test that more stringently, especially for diffeomorphs. One could also do cultural studies of response bias in domains other than recognition memory (for example, a visual target detection task). Also, one could use payoffs to manipulate response bias. For example, if a hit earns 6 points and a correct rejection earns 4 points, subjects will respond more liberally than if the two kinds of correct responses earn the same reward. How big a difference in rewards for hits versus correct rejections would it take for the Japanese to demonstrate neutral response bias? Furthermore, we have tried new approach for analyzing recognition memory data using multilevel models. This approach aids in looking at trial by trial data of each participant and measure participants' responses without the need to use SDT measures like c and d'.

Additionally, there are new methods that are being used to study the memorability of stimuli, such as the development of computational models that aid in predicting responses based on stimuli manipulations (refer to Supplementary Material). Even though finding collaborators from collectivistic countries is a difficulty many researchers face, it is worthwhile to put in the effort to enable generalizing phenomena such as the MBBE.

Conclusion

It is worthwhile to learn about what makes images memorable. One of the benefits of learning about their qualities is attempting to quantify the properties of images that are

most appealing. This helps in creating memorable images in visualization, photography, and education. But [N7] it would be inaccurate to generalize phenomena tested in one culture to all other populations from distinct cultural backgrounds.

Our study on recognition memory across Canada and Japan compared the recognition memory of samples from different cultural backgrounds, social norms, and customs. By examining recognition memory across these populations, we could determine if cultural factors might influence cognitive processes and memory performance. In other words, this thesis has provided new insights into understanding the MBBE and its generalization across cultures. The materials-based bias effect cannot be generalized to all visual stimuli since response bias for photographs differs from that for difffeomorphs. Even though both are visual stimuli, bias could vary across cultures,as we noticed that each of our samples has a distinctive response bias for diffeomorphs. As for the distinction between stimuli, contrary to expectations, the use of diffeomorphs complicated understanding the MBBE as neither semantics (line drawings) nor colorfulness (diffeomorphs) seem to elicit a conservative bias.

All in all, the differences observed between Japanese and Canadian populations are likely to be attributed to cultural influences rather than racial differences. In our studies, the primary objective was to enhance the understanding of recognition memory for people of different backgrounds. Racial research, on the other side, would seek to understand the genetic and evolutionary underpinnings behind variations in recognition memory, if present.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis.*Wiley interdisciplinary reviews: computational statistics,* 2(4), 433-459.

Amano, S., & Kondo, T. (1999). Nihongo-no Goi-Tokusei (Lexical properties of Japanese).*Sanseido Tokyo.*

Aminoff, E. M., Clewett, D., Freeman, S., Frithsen, A., Tipper, C., Johnson, A., … Miller, M. B. (2012). Individual differences in shifting decision criteria: A recognition memory study.*Memory & Cognition*, 40(7), 1016–1030.

Benjamin AS, Diaz M, Wee S. Signal detection with criterion noise: applications to recognition memory. *Psychol Rev.* 2009 Jan;116(1):84-115. doi: 10.1037/a0014351. PMID: 19159149; PMCID: PMC2862236.

Beth, E. H., Budson, A. E., Waring, J. D., & Ally, B. A. (2009). Response bias for picture recognition in patients with Alzheimer's disease.*Cognitive and behavioral neurology: official journal of the Society for Behavioral and Cognitive Neurology,* 22(4), 229.

Brown, J., Lewis, V. J., & Monk, A. F. (1977). Memorability, word frequency and negative recognition. *Quarterly Journal of Experimental Psychology,* 29(3), 461–473. doi:10.1080/14640747708400622

Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., & Oliva, A. (2015). Intrinsic and extrinsic effects on image memorability. *Vision research*, 116, 165-178.

Dubey, R., Peterson, J., Khosla, A., Yang, M. & Ghanem, B. What makes an object memorable? in *Proc. IEEE Int. Conf. on Computer Vision* (ICCV), 1089–1097 (2015).

Dobbins, I. G., & Kroll, N. E. A. (2005). Distinctiveness and the Recognition Mirror Effect: Evidence for an Item-Based Criterion Placement Heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 31(6), 1186–

1198. https://doi.org/10.1037/0278-7393.31.6.1186

Dougal, S., & Rotello, C. M. (2007). “Remembering” emotional words is based on response bias, not recollection. *Psychonomic Bulletin & Review*, 14(3), 423-429.

Fallow, K.M., Lindsay, D.S. Test position effects on hit and false alarm rates in recognition memory for paintings and words. *Memory & Cognition* 50, 378–396 (2022) https://doi.org/10.3758/s13421-021-01227-5

Fox, J., Dennis, S., & Osth, A. F. (2020). Accounting for the build-up of proactive interference across lists in a list length paradigm reveals a dominance of item-noise in recognition memory. *Journal of Memory and Language*, 110, Article 104065. https://doi.org/10.1016/j.jml.2019.104065

Freire, M.R., Pammer, K. Influence of culture on visual working memory: evidence of a cultural response bias for remote Australian Indigenous children. *J Cult Cogn Sci* 4, 323–341 (2020). https://doi.org/10.1007/s41809-020-00063-4

Glanzer, M., Kim, K., Hilford, A., & Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(2), 500.

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. *John Wiley.*

Gutchess, A., & Sekuler, R. (2019). Perceptual and mnemonic differences across cultures. *Psychology of Learning and Motivation*, 131–174. https://doi.org/10.1016/bs.plm.2019.06.001

Hintzman, D. L. (2009). How does repetition affect memory? *Evidence from judgments of recency. Memory & Cognition*, 38(1), 102–115. doi:10.3758/MC.38.1.102

Iwata, N. (2014). Cultural Distinctiveness in Response Bias. In: Dollard, M., Shimazu, A., Bin Nordin, R., Brough, P., Tuckey, M. (eds) Psychosocial Factors at Work in the Asia Pacific.

Springer, Dordrecht. https://doi.org/10.1007/978-94-017-8975-2_3

Kantner, J., & Lindsay, D. S. (2012). Response bias in recognition memory as a cognitive trait. *Memory & Cognition,* 40, 1163-1177.

Kellen, D., Winiger, S., Dunn, J. C., & Singmann, H. (2021). Testing the foundations of signal detection theory in recognition memory. *Psychological Review,* 128(6), 1022–1050. https://doi.org/10.1037/rev0000288

Kuwabara, M., & Smith, L. B. (2016). Cultural differences in visual object recognition in 3-year-old children. *Journal of Experimental Child Psychology*, 147, 22-38.

Ksander, J.C., Paige, L.E., Johndro, H., & Gutchess, A.H. (2018). Cultural specialization of visual cortex. *Social Cognitive and Affective Neuroscience*, 13, 709 - 718.

Lee, C.-Y., Chen, C.-C., Mair, R. W., Gutchess, A., & Goh, J. O. (2021). Culture-related differences in the neural processing of probability during mixed lottery value-based decision-making. *Biological Psychology*, 166, 108209. https://doi.org/10.1016/j.biopsycho.2021.108209

Lindsay, D.S., & Kantner, J. (2011). A search for influences of feedback on recognition of music, poetry, and art. In P. Higham & J.Leboe (Eds.), *Constructions of remembering and metacognition: Essays in honor of Bruce Whittlesea.* Houndmills, UK: Palgrave Macmillan.

Lindsay, D. S., Kantner, J., & Fallow, K. M. (2014). Recognition memory response bias is conservative for paintings and we don't know why. In Remembering (pp. 237-254). *Psychology Press.*

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review,* 1(4), 476–490. https://doi.org/10.3758/BF03210951

Macmillian, N. A., & Creelman, C. D. (2005). Detection theory: a user's guide, 2nd edn.

Lawrence.

Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin,* 98(1), 185 – 199.

Masuda, T., & Nisbett, R. E. (2001). Attending holistically versus analytically: comparing the context sensitivity of Japanese and Americans. *Journal of personality and social psychology,* 81(5), 922.

Millar, P. R., Serbun, S. J., Vadalia, A., & Gutchess, A. H. (2013). Cross-cultural differences in memory specificity. *Culture and Brain,* 1(2–4), 138–157.

Mickley Steinmetz, K. R., Sturkie, C. M., Rochester, N. M., Liu, X., & Gutchess, A. H. (2017). Cross-cultural differences in item and background memory: Examining the influence of emotional intensity and scene congruency. *Memory,* 26(6), 751–758. https://doi.org/10.1080/09658211.2017.1406119

Miyamoto, Y., Nisbett, R. E., & Masuda, T. (2006). Culture and the physical environment: Holistic versus analytic perceptual affordances. *Psychological Science,* 17(2), 113-119. https://doi.org/10.1111/j.1467-9280.2006.01675.x

Osth, A. F., Dennis, S., & Kinnell, A. (2014). Stimulus type and the list strength paradigm. *The Quarterly Journal of Experimental Psychology,* 67(9), 1826–1841. https://doi.org/10.1080/17470218. 2013.872824

Osth, A. F., & Dennis, S. (2020). Global matching models of recognition memory.

Paige, L.E., Amado, S. & Gutchess, A.H. Influence of encoding instructions and response bias on cross-cultural differences in specific recognition. *Cult. Brain* 5, 153–168 (2017). https://doi.org/10.1007/s40167-017-0055-x

Paivio, A., Rogers, T. B., & Smythe, P. C. (1968). Why are pictures easier to recall than words? *Psychonomic Science,* 11(4), 137-138.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Radvansky, G. A., Doolen, A. C., Pettijohn, K. A., & Ritchey, M. (2022). A new look at memory retention and forgetting.*Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Saraee, E., Jalal, M., & Betke, M. (2020). Visual complexity analysis using deep intermediate-layer features. *Computer Vision and Image Understanding*, 195, 102949.

Schacter, D. L., Israel, L., & Racine, C. (1999). Suppressing false recognition in younger and older adults: *The distinctiveness heuristic. Journal of Memory and Language*, 40, 1–24. doi:10.1006/jmla.1998.2611

Schwartz, A. J., Boduroglu, A., & Gutchess, A. H. (2014). Cross-cultural differences in categorical memory errors. *Cognitive Science,* 38(5), 997–1007. https://doi.org/10.1111/cogs.12109

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning & Memory*, 6(2), 174–215. https://doi.org/10. 1037/0278-7393.6.2.174

STUDENT, A. (2019). CSUNPOSIUM.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior research methods, instruments, & computers*, 31(1), 137-149.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 26(3), 582.

Vokey, J. R. (2016). Single-step simple ROC curve fitting via PCA. *Canadian Journal of*

Experimental Psychology/Revue Canadienne de Psychologie Expérimentale, 70(4), 301–305. https://doi.org/10.1037/cep0000095

Wang, Q., & Ross, M. (2005). What we remember and what we tell: The effects of culture and self-priming on memory representations and narratives. *Memory*, 13(6), 594-606.

Windmann, S., & Chmielewski, A. (2008). Emotion-induced modulation of recognition memory decisions in a Go/NoGo task: Response bias or memory bias?*Cognition and Emotion*, 22(5), 761-776.

Wong, B. I., Yin, S., Yang, L., Li, J., & Spaniol, J. (2018). Cultural Differences in Memory for Objects and Backgrounds in Pictures. *Journal of Cross-Cultural Psychology*, 49(3), 404–417. https://doi.org/10.1177/0022022117748763

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin,* 133(5), 800–832. doi:10.1037/0033-2909.133.5.800

Zhang, W., Andrews-Hanna, J. R., Mair, R. W., Goh, J. O., & Gutchess, A. (2022). Functional connectivity with medial temporal regions differs across cultures during post-encoding rest. *Cognitive, Affective, & Behavioral Neuroscience*, 22(6), 1334–13

Supplementary materials

A- Neuropsychological cross-cultural research in the literature

In addition to providing literature on the potential cultural variation in memory, it could be relevant to outline what the cognitive-neuroscientific literature posits in this regard. An increasing collection of research indicates that culture affects how people see the world (Ksander et al., 2018). By evaluating voxel pattern representations using multi-voxel pattern analysis (MVPA), this study explored whether these cultural variations are evident in a straightforward object-seeing task and the visual cortex. Participants (20 East Asians and 20 Americans) underwent functional magnetic resonance imaging while seeing images of commonplace objects that were matched for cultural familiarity and conceptual agreement. These stimuli were examined using whole-brain searchlight mapping and non-parametric statistical analysis to see if they generated multi-voxel patterns that varied between cultural groups. In Brodmann regions 18 and 19 (in the visual cortex), researchers discovered that individuals' cultural identities could be accurately predicted from stimulus representations. This finding highlights culturally specific visual cortex functioning distinction in the place perception test.

Another study assessed how cultural variations in the analytical or holistic processing of probabilistic outcomes during value-based decisions reflect disparities in economic risk-taking between Westerners and East Asians (Lee et al., 2021). An investigation using the Lottery Choice Task (LCT) of functional magnetic resonance imaging (fMRI) was completed by 27 Americans (US) and 51 Taiwanese (TW) young adults. Participants chose to accept or reject bets with variable odds of gaining or losing

points of various magnitudes. When win probabilities were greater than 0.50, TW individuals accepted more bets, but US participants discriminately decreased their acceptance rates of winning bets as win probabilities increased. Similar numbers of losing wagers (win probability 0.5) were rejected by both groups.

Importantly, responses from the ventromedial prefrontal cortex (VMPFC) indicated stronger discrimination between win probability circumstances in the US and in TW groups. In brief, the findings show that cultural variations in the ability to discern between probabilistic reward outcomes influence brain calculations of risk and prospects through a neurocognitive process in the VMPFC.

Resting brain connectivity can reveal individual variations and influence behavioral outcomes, including memory. In this work, researchers have looked at how cultural differences affect functional connectivity with areas of the medial temporal lobe (Zhang et al., 2022). After encoding images of items, 59 East Asians and 46 Americans participated in this study. The left parahippocampal gyrus (anterior and posterior areas) and left hippocampus were chosen as seed regions to study cross-cultural variations in resting state functional connectivity. These areas were chosen because they had previously been connected to research on how cultural variations affected the success storage age of detailed memories.

The findings showed that East Asians had a better connection with parieto-occipital areas than Americans had stronger connectivity with the left posterior parahippocampal gyrus. People from East Asia and Americans both demonstrated more connection in the left anterior parahippocampal gyrus with the frontal and temporal areas, respectively. Although patterns were related to cultural values, connectedness was not related to memory function. The results are examined in relation to possible cultural

variations in memory consolidation or more general trait- or state-based processes, such as holistic vs. analytic processing.

Researchers in this study hypothesized that cultural variations in how different forms of information are prioritized and how that affects how well that information is remembered. Communication between medial temporal and ventral stream areas suggests that East Asian participants tended to continue mobilizing object-processing regions, while Americans may have kept up their mobilization of spatial processing areas. These networks may have implications for post-encoding consolidation of information in memory or reflect more general cultural differences in cognitive processes (e.g., holistic vs. analytic processing) or network connectivity.

B- An alternative approach towards understanding the MBBE

More recent studies have aimed to develop a method for analyzing visual complexity in images by utilizing deep features extracted from a pre-trained convolutional neural network and to demonstrate its effectiveness for various applications, such as image quality assessment, image retrieval, and image understanding (Saraee & Jalal, 2020). Researchers can learn more about what makes an image memorable or simple to recognize by measuring its visual complexity. The way our brains interpret visual information may make some levels of visual complexity more memorable. Furthermore, unlike words, which rely on semantic and phonetic processing, these techniques can help reveal the underlying patterns and structures in images, making them simpler to recognize and identify. By enhancing the memorability and recognition of visual content, the results of such analyses may be used to improve image-based communication and information retrieval systems.

Other studies have focused on examining how different levels of visual

information, like color texture, and scene content, influence image memorability. More specifically, Bylinskii et al. (2015) have assessed the significance of these elements in affecting image memorability by providing a computational model for predicting the memorability of images. By examining the role of several visual cues in predicting image memorability, the work by Bylinskii et al. (2015) advances our knowledge of the MBBE. The research illuminates the fundamental mechanisms that render images more remembered than words by investigating the effects of color, texture, and scene content on image memorability. According to research by Bylinskii et al. (2015), some visual characteristics are significantly connected with how memorable an image is. Their computer model demonstrated that images with more colors tend to be more remembered, with color information being a strong predictor of visual memorability. Their model also showed that the distribution of edges, textures, and gradients in an image are only a few textural aspects that help make an image memorable. Their model also showed that a key factor in how memorable an image is its scene content (for ex: the presence of things like people, things, and other semantic information). This suggests that by combining parameters related to color, texture, and scene content, it is possible to predict image memorability successfully. This demonstrates the significance of these visual elements in comprehending the MBBE and offers insightful advice for creating more memorable visuals.

Further studies aiming at investigating the visual memorability of natural images have developed a computer model for anticipating image memorability (Khosla et al., 2015). By using a collection of photos labeled with their corresponding memorability scores to train a deep neural network, the study's authors developed a computational model for predicting image memorability. Human participants who were shown a series of

images and then asked to judge how well they remembered each one provided eye-tracking data that was used to create the dataset. Based on an image's visual characteristics, including its color, texture, and composition, as well as its spatial and semantic context, the neural network was trained to predict how memorable an image will be. To optimize a pre-trained deep neural network that had previously been trained on a sizable dataset of photos for object recognition tasks, the authors utilized a method termed transfer learning.

The existence of prominent objects or regions inside an image, according to the study, is the most crucial element in determining visual memorability. Those things or areas that stand out from their surroundings and draw the eye are known as salient. According to the study, photographs featuring conspicuous objects or regions are more likely to stick in people's minds than those without them. The scientists did point out that other visual components, such as color and intricacy, can also influence how memorable an image is. ones with a moderate level of complexity and warm, vibrant colors were found to be more recalled than ones that are very simple or overly complex or that use cool, gloomy colors. The study's conclusions suggest that certain visual characteristics, such as color, brightness, and complexity, might predict how memorable an image would be. Additionally, the study discovered that fixations and eye movements are very important in deciding how memorable an image is. The possibility of applying computational approaches to comprehend image memorability more fully is demonstrated by the authors' suggested computational model's state-of-the-art performance on predicting image memorability. The results of the study are eye-opening in the sense that computational models can be used to better understand and predict image memorability.

C- Multilevel modelling to measure variation in response bias across trials

1- The Lebanese sample (N= 108):

After presenting the overall sensitivity and response bias data, it is relevant to discuss some analyses that show how these measures change over the course of the test. For these analyses, hit and false alarm rates were calculated based on the responses and numbers of old and new words and paintings in each individual quartile. In other words, to understand how the measured responses change across trials and examine if there are any individual changes, a multi-level model would be employed.

Throughout the study, the type of materials might seem distinctive or memorable as the test goes on, leading some participants to adjust their decision criteria. Typically, hit rates decline over the course of the test, while the pattern for false alarm rates varies (Fox et al., 2020). These patterns could arise from the interplay among multiple mechanisms, with stimulus-, participant-, and experiment-level factors potentially influencing the relative contributions of each. Notably, most previous research on materials-based differences in memory has focused on hit rates or accuracy, but work by my colleague (Dr. Fallow) has demonstrated that limiting comparisons to such measures risks missing potentially informative materials-based response bias effects.

Computing a multi-level model, defining measures:

Usually, researchers calculate response bias (c) as an aggregate across a test. Indeed, c cannot be calculated without a hit rate and a false alarm rate across some number of trials. Some researchers have examined changes in bias over the course of a test by dividing the test into portions, such as the first through fouth quartile of test trials. Estimates of c can then be measured in each quartile. Here I introduce a new method that

uses the confidence-weighted scale with which subjects reported their old-new judgments in a multi-level analysis.

After studying the responses of participants, we wanted to account for individual differences in starting points and individual differences in the rates of change as a function of old/new status. Therefore, we added a predictor to level 2 of both equations. Note that after creating the initial structure of the model, we have added a level 1 predictor variable, that is “trial”, to see if subjects’ individual responses change between each trial. Then, we added a time-invariant predictor (item_status), which is the level 2 predictor, that helped us in understanding if the responses differ depending on whether the item is old or new (refer to the R code for a better understanding of the technicalities).

We report fixed and three random effects estimated in the model:

- The fixed intercept ($\gamma 00 = 5.27$) is essentially, with the presence of predictors, the predicted value when all other predictor variables equal zero. In this case, it would be an estimate at baseline (trial= 1) when the item status is old (Item_status = 0).
- This represents the effect of item status when time = 0 ($\gamma 01 = -2.19$),trial = reflecting the effect of item status on intercept. This is the difference between intercepts (baseline levels of 2 item status scores at 0). It represents an effect of item status when trial= 1, so it is an intercept. It is noted to be the hypothesized difference of average initial status between the stimuli that participants studied vs those that they did not study.
- Fixed slope ($\gamma 10 = -0.04$) which is the average rate of change of responses across trials.
- An interaction between trials and item status whereby we ask whether the effect of time differs based on item status. This represents the effect of item status on rate of

change.

- It has a value of ($\gamma 11 = 0.02$). Based on item status, added to the slope, an interaction happens. The purpose is to seek whether the effect of trials' progression differs if stimuli are old or new.
- Random intercept: u0i = 0.41
- Random slope: u1i = 0.02
- Residual: eit = 1.46

I estimated the model that includes 'item_status' as a Level 2 predictor as follows:

- Level 1: Respit = $\beta 0i + \beta 1i trialit + eit$
- Level 2: $\beta 0i = \gamma 00 + \gamma 01 item\_statusi + u0i$
- $\beta 1i = \gamma 10 + \gamma 11 item\_statusi + u1i$

That is, we have individual differences in intercept u0i, individual differences in the rate of change u1i, and the residual eit. Also, we have a fixed intercept $\gamma 00$, an effect of item status when time = 0 ($\gamma 01$), an effect of trials $\gamma 10$, and an interaction between item status and trials $\gamma 11$.

Model 3's Plot:

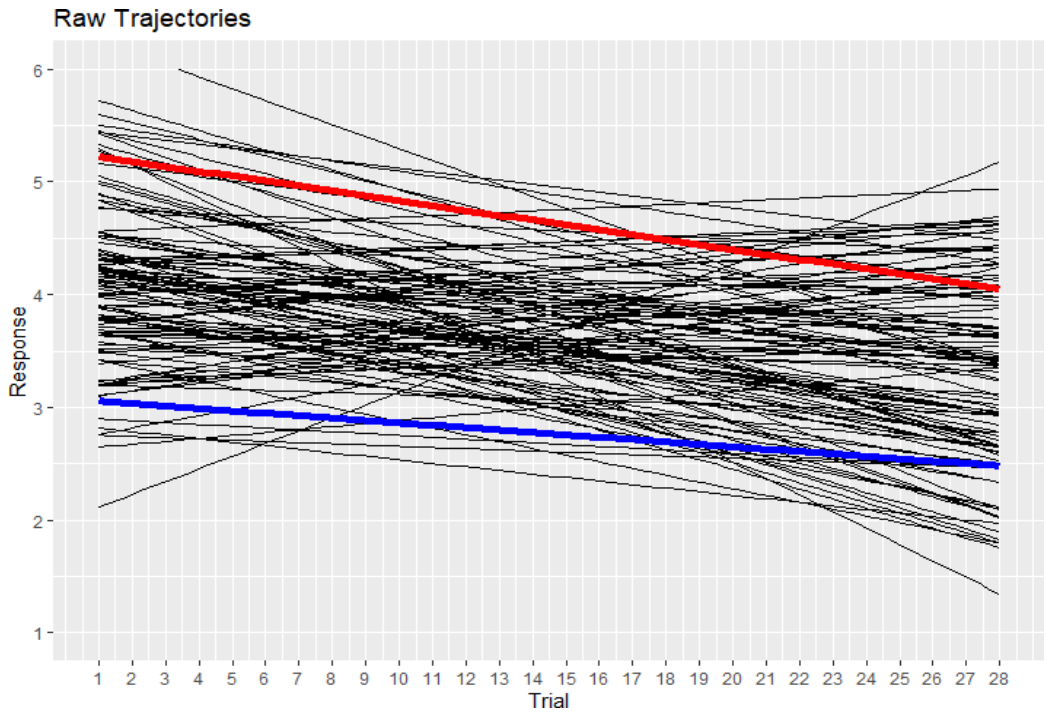


Figure 28: reveals the difference between the estimated confidence ratings for old (red) compared to new (blue) stimuli.

In Figure 28, the black lines represent participants' raw response visualized over the 40 trials of the test phase. A linear function has been applied to them for graphical purposes. As for the red and blue lines, they are considered the estimated trajectories of participants' responses. The red line is the average rate of change in confidence for old items; and the blue line is the average rate of change in confidence for new items. The focus of this analysis was to explore responses' tendency across trials, and understand within or between-subject differences, if any. What was quite novel in this analysis was the distinction between the response ratings' intercepts of old and new stimuli. To be clear, for old stimuli, responses started at 5.27 at the first trial. This means that when subjects saw an old stimulus that they had studied before, their response started at 5.27 for the first trial. On the other hand, the responses for new items started at 3.07 for the first items. This means that when subjects did not study the item, their responses started at the level of 3.07 for the first trial. This shows that there is a significant difference in responses (2.19, $p < 0.005$) between recognizing old and new stimuli. At the beginning of the test, the mean estimated rating (collapsing across old and new items) was about 4.71, more than a few points about the midpoint. So, it appears that at the beginning of the test response ratings were

liberal (because the average across old and new items was greater than 3.5, the mid-point of the scale). At the end of the test, the estimated rating across old and new items had declined to a less liberal or somewhat more neutral response.

2- **The Canadian (170) and Japanese (129) samples:**

Multi-level modeling across trials

To examine how the responses vary over the course of the test and if there are any individual changes, a multi-level model was employed.

Computing a multi-level model, defining measures:

In parallel to Experiment 1, the confidence-weighted scale was used to measure subjects' responses.

We have four fixed and three random effects estimated in the model:

The fixed intercept ($\gamma 00$=4.87) mod 3, is essentially, with the presence of predictors, the predicted value when all other predictor variables equal zero. In this case, it would be an estimate at baseline (trial= 1) when the item status is old (Item_status = 0).

- An effect of item status when time = 0 ($\gamma 01$ = -2.06) reflecting the effect of item status on intercept. The difference between intercepts, baseline levels of 2 item status scores at 0. This represents an effect of item status when trial= 1, so it is an intercept. It is noted to be the between-effect.
- Fixed slope ($\gamma 10$ = -0.024) which is the average rate of change of responses across trials.
    -

An interaction between trials and item status whereby we ask whether the effect of time differs based on item status. The effect of item status on the rate of change. It has a value of ($\gamma 11 = 0.021$). Based on item status, an interaction happens. The purpose is to determine whether the effect of trials' progression differs if stimuli are old or new. Random intercept: $u0i = 0.53$

- Random slope: $u1i = 0.02$
- Residual: $eit = 1.43$

We estimated the model that includes 'item_status' as a Level 2 predictor as follows:

- Level 1: $Respit = \beta 0i + \beta 1itrialit + eit$
- Level 2: $\beta 0i = \gamma 00 + \gamma 01item\_statusi + u0i$
- $\beta 1i = \gamma 10 + \gamma 11item\_statusi + u1i$

That is, we have individual differences in intercept u0i, individual differences in the rate of change u1i, and the residual □it. Also, we have a fixed intercept $\gamma 00$, an effect of item status when time = 0 ($\gamma 01$), an effect of trials $\gamma 10$, and an interaction between item status and trials $\gamma 11$. Similarly, the Japanese sample was modeled, and the pertaining values are as follows:

- The fixed intercept ($\gamma 00=4.57$)
- An effect of item status when time = 0 ($\gamma 01 = -1.70$)
- Fixed slope ($\gamma 10 = -0.023$) which is the average rate of change of responses across trials
- An interaction between trials and item status of ($\gamma 11 = 0.018$)- that is the average rate of change for new stimuli.
- Random intercept: $u0i = 0.5$
- Random slope: $u1i = 0.02$
- Residual: $eit = 1.43$
- Response for old stimuli started at 4.57 at the first trial.

- Response for new stimuli started at 2.87 at the first trial.

Model 3's Plot:

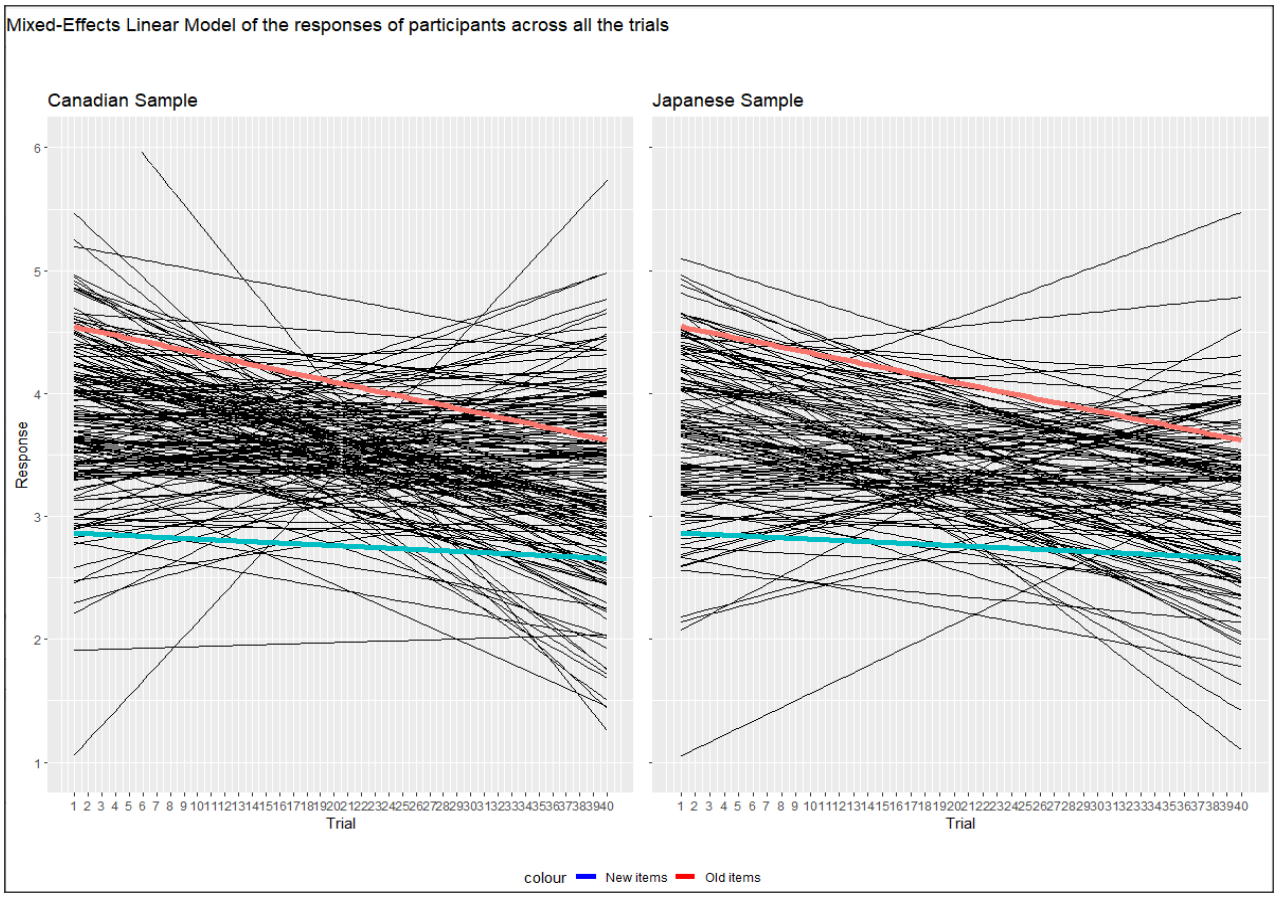


Figure 29: reveals the difference between the estimated confidence ratings for old (red) compared to new (blue) stimuli.

The focus of this analysis is to explore response ratings' tendency across trials, and understand within or between-subject differences, if any. Generally, the responses plotted have a decreasing slope. Regarding the Canadian sample, for old, studied stimuli, responses started at 4.87 at the first trial. This means that when subjects saw an old stimulus that they had studied before, on average their response started at 4.87 for the first trial. On the other hand, responses for new items started at 2.81 on average for the first items. This means that when subjects did not study the item, their response on age started at the level of 2.81 for the first trial. This shows that there is a significant difference in responses (2.06, $p< 0.005$) between recognizing old and new stimuli. At the beginning of the test, the mean estimated rating (collapsing across old and new items) was about 3.84, slightly above the midpoint of 3.5, suggesting slightly liberal bias

Towards the end of the test, the estimated average response rating across old and new items in the Canadian sample had declined highlighting the tendency of responses in becoming less liberal.

Concerning the rate of change in old/new ratings across the test, it seemed to be clearly changing for both stimuli, old and new, which was expected. Yet, the rate of change significantly differed depending on the item status. For old items, the rate of change was 0.024 units. However, for new stimuli, the rate of change was 0.021 units. The interaction showed that the difference in the rate of change across trials is statistically significant between old and new stimuli ($p < 0.005$). With this model, we were able to understand that the calculated responses were expected to differ significantly depending on the item status, which is consistent with prior findings that highlighted a difference in responses over trials depending on item type (Fallow and Lindsay, 2022). Compared to the old items, ratings for new items on the old/new scale changed at a slower rate. Old stimuli varied in the response rate by 0.024 units while new stimuli changed at a value of 0.021.

Regarding the Japanese sample, for old, studied stimuli, responses started on average at 4.57 at the first trial. This means that when subjects saw an old stimulus that they had studied before, their average response started at 4.57 for the first trial. On the other hand, responses for new items started at 2.87 for the first items. This means that when subjects did not study the item, on average their response started at the level of 2.87 for the first trial. This shows that there is a significant difference in responses (1.70, $p< 0.005$) between recognizing old and new stimuli. At the beginning of the test, the mean estimated rating (collapsing across old and new items) was about 3.72, somewhat above the midpoint, suggesting slightly liberal bias. Towards the end of the test, the estimated rating across old and new items changed indicating a tendency towards less liberal

responses.

Concerning the rate of change, it seemed to be clearly varying for both stimuli, old and new, which was expected. Yet, the rate of change significantly differed depending on the item status. For old items, the rate of change was 0.023. However, for new stimuli, the rate of change was changing by 0.018. The interaction showed that the difference in the rate of change across trials is statistically significant between old and new stimuli ($p<$ 0.005). With this model, we were able to understand that the calculated responses were expected to differ significantly depending on the item status. Compared to the old items, response for new items was changing in response at a slower rate. Old stimuli changed in the response rate by 0.023 units while new stimuli changed at 0.018.

The table below summarizes the differences between the samples' statistical output

Intercept	The average response for the first trial of an old item (item_status = 0) is higher in the Canadian sample (4.87) compared to the Japanese sample (4.57).
Trial Number	The response decrease per trial is very similar between the two groups, -0.025 in the Canadian sample and -0.024 in the Japanese sample, but slightly higher in the Canadian

	sample.
Item Status	The decrease in response when an item is new (item_status = 1) is more pronounced in the Canadian sample (-2.06) compared to the Japanese sample (-1.70). This suggests that Canadian participants had a larger drop in responses for new items compared to Japanese participants.
Interaction Effect	The interaction effect of 'trial_no' and 'item_status' is larger in the Canadian sample (0.022) compared to the Japanese sample (0.018). This indicates that the rate at which response varies over trials is less steep for new items compared to old items, and this effect is more pronounced in the Canadian sample
Intercept Variance (subj)	The variance of the random intercepts for 'subj' is slightly higher in the Canadian sample (0.53)

	compared to the Japanese sample (0.50). This suggests a bit more variability in the average responses of different subjects in the Canadian sample.
Slope Variance (trial_no)	The variance of the random slopes for 'trial_no' within 'subj' is higher in the Canadian sample (0.022) compared to the Japanese sample (0.015). This indicates more variability in the change in responses over trials between subjects in the Canadian sample.
Residual variance	The residual variance is very similar between the two groups, 1.43 in the Canadian sample and 1.43 in the Japanese sample. This suggests that the unexplained variability in the responses is similar in both groups.
Correlation between intercept and slope	The correlation between the random intercepts and slopes is higher in the

	Canadian sample (-0.78) compared to the Japanese sample (-0.65). This suggests that in the Canadian sample, those with higher average responses show a steeper decrease in responses over trials compared to the Japanese sample.

The standard deviations of these random effects indicate the extent of variability among subjects in their baseline responses and rates of change in responses. Looking at the intercept and slope variance values, we can see that there is slightly more variability in the Canadian sample than the Japanese sample both in terms of their baseline responses and how their responses change across trials. In terms of the correlation between the intercept and slope, subjects who have higher initial responses tend to have a steeper decline in responses across trials in both samples. However, this correlation is somewhat stronger in the Canadian sample. This could suggest that there is greater individual variability in the recognition memory response patterns among the Canadian sample compared to the Japanese sample. One possible interpretation is that the Canadian sample shows slightly more individual variability in their response patterns than the Japanese sample, at least in terms of changes over trials. This might suggest greater individual differences in processing the stimuli, perhaps reflecting a more individualistic culture. The Japanese sample, showing slightly less variability, might reflect more shared processing mechanisms or responses within the group, possibly consistent with a more collectivist

culture. All in all, the difference in the response over trials is somewhat less steep when the item is new compared to when it is old. However, this effect is somewhat more pronounced in the Canadian sample. It seems that the greater discrimination between old and new items for Canadians reflect better memory performance compared to Japanese. In the next section, we aim to test this statistically.

A highlight of the results was constructing estimates of responses for old and new items, the amount of separation between red and blue (which indicate old/new discrimination). For both samples, early in the test most subjects were liberal, while by the end of the test most subjects were less liberal. It is crucial to understand that a difference over the course of the test in ratings of oldness to old test items does not say anything about a change in response bias. Such a decline could be due partly or entirely to forgetting. These multilevel models allowed for more flexibility in modeling error structures, using time as continuous variable, including random effects, providing detailed temporal information and no averaging of data. This aids in detection of learning effects and more flexible modeling of time.

Trial-by-trial analysis can provide a clearer view of how an effect develops over time compared to quartile analysis. Compared to quartile analysis, which involves categorizing the data into larger bins, trial-by-trial analysis offers a more thorough and precise picture of how answers evolve over time. Additionally, because it more effectively accounts for individual differences, it calls for advanced statistical models. I believe we can infer better insights from observing Figure 29 in this document which shows the changes in responses at each trial. Perhaps using a set of stimuli that are of the same kind does not show dramatic changes in responses on a trial-by-trial basis, but if the same method is used over a range of stimuli that are intermixed in the same experiment, such as

line drawings, diffeomorphs, and paintings, such a method would be accurate in describing differences in responses. Furthermore, a disadvantage of quartile analysis is that it allows the researcher to look at only a small number of points. With an MLM, researchers can look at bias without having to rely on c or other measures of response bias (recognition memory data is known to violate the equal variance assumption necessary for using c) and that's the advantage of the MLM model.

In brief, the results of both samples showed that responses decreased across trials, and they decreased more slowly for new, non-studied than for old, studied stimuli. Another important realization is that, in addition to the liberal response bias detected in the initial analysis, responses were more liberal (closer to 6) for old compared to new (neutral at 3) stimuli. This analysis led to the understanding that analyzing data in novel ways is eye-opening for understanding datasets in new ways that could be insightful.

D- Multi-level modeling proposed methods to measure sensitivity and bias

Drs Masson and Krawitz have provided valuable suggestions based on the following observations from the given data:

1- The absolute level of both the red (slope for estimated response for old stimuli) and blue (slope for estimated response for new stimuli) lines represent response bias. Any change in slope for old or new items reflects a change in response bias.

2- The degree of separation between the red and blue lines reflects accuracy/discrimination/sensitivity.

Taking these observations into account, Dr Masson recommended the following adjustments:

1- Develop an analogy to describe the separation between the red and blue lines- thus explaining accuracy.

2- Create an analogy to represent the average value of the red and blue lines- thus explaining bias.

These suggestions aim to enhance the understanding and interpretation of the results. By following the committee's recommendations, I hope my suggested methods provide a better understanding of my multi-level model results contributing to an enhanced understanding of accuracy and response bias.

Note: the disadvantage of quartile analysis is that it allows you to look at only a small number of points. With an MLM, we can look at bias without having to rely on c or other measures of response bias and that's the advantage of the MLM model.

Kindly note: These suggestions were made based on the following figure which demonstrates a multi-level model that has been visualized across 40 trials. The figure is based on the raw confidence weighted scale responses:

1- Definitely not studied	2 - Probably not studied	3- Maybe not studied	4- Maybe studied	5- Probably studied	6- Definitely studied

Title: Mixed Effects Model for both Canadian and Japanese Sample

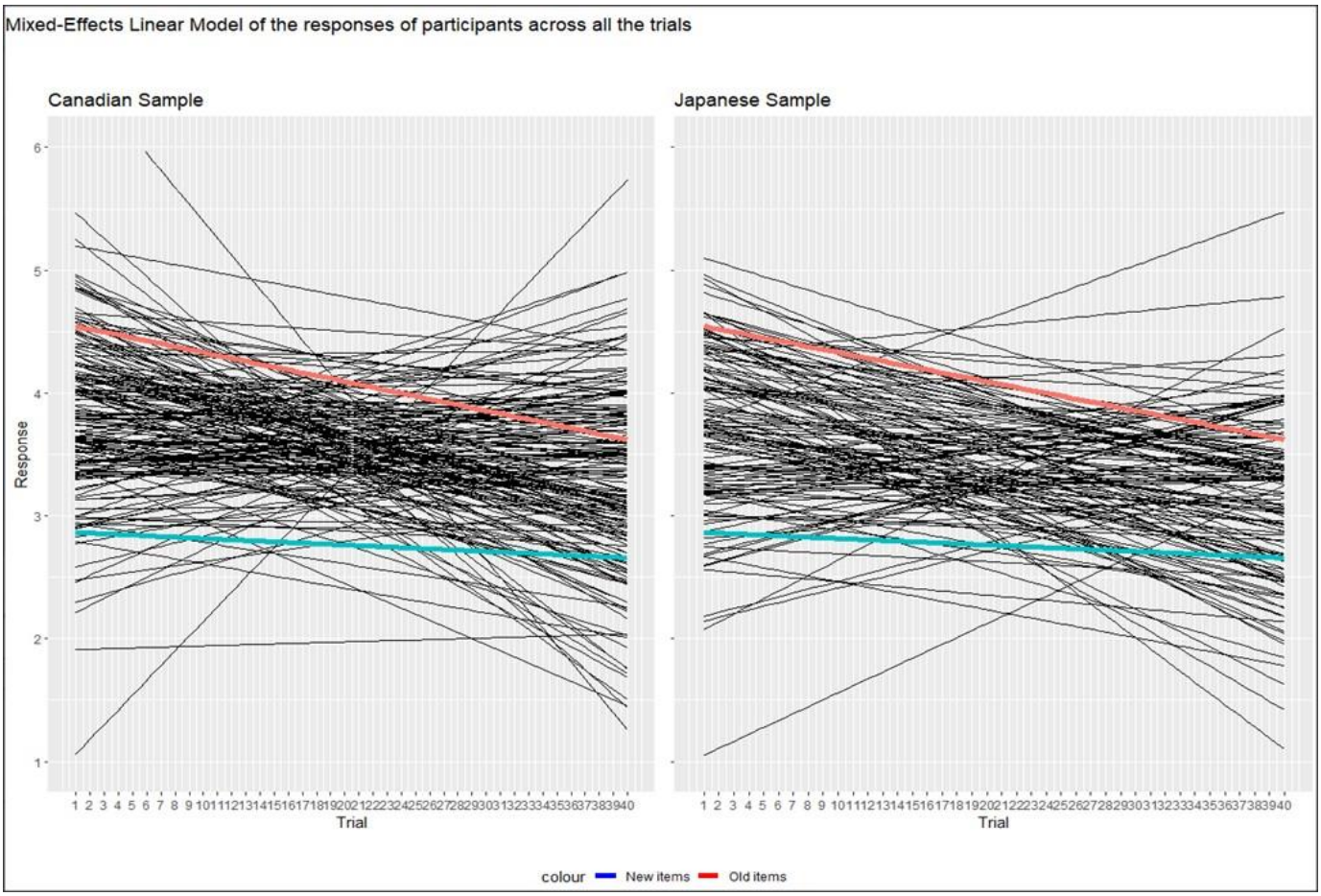


Figure 30: This figure illustrates the visualized absolute value demonstrating the separation between red and blue lines across trials.

Objective 1: To find an analogy to the separation of these 2 lines, I propose:

The first method:

Another way I measured the separation between the two curves was by calculating the mean absolute deviation (MAD). The MAD represents the average absolute difference between the two functions.

Here's how I calculated the MAD in R:

1- I defined the range of x-values over which the functions were defined. In this case, the range was from 1 to 40.

2- I created a vector of x-values within this range, for example, using the seq() function.

3- I calculated the difference between the two functions at each x-value in the range.

4- I calculated the absolute value of the differences. 5

5- I calculated the mean of the absolute differences.

Here's the code in R that I used:

```
# Define the range of x-values x_range <- seq(1, 40, by = 1)

 # Calculate the differences between the two functions at each x-value differences <-
fun_m3_newitem(x_range) - fun_m3_olditem(x_range)

# Calculate the absolute value of the differences absolute_differences <- abs(differences)

# Calculate the mean of the absolute differences mad <- mean(absolute_differences)

# Display the MAD print(mad)
```

Output: 1.321038

The mad value represents the average absolute difference between the two functions, with a lower value indicating a smaller separation between the curves.

Now, if I wanted to calculate the absolute value of the separation at each point (x-value) along the x-axis, I could simply use the absolute_differences vector calculated above. This vector contains the absolute value of the separation between the red and blue lines at each trial (x-value).

I plotted the absolute_differences vector to visualize the absolute value of the separation at each point:

```
# Display the absolute differences print(absolute_differences) :

[1] 1.6798400 1.6614399 1.6430399 1.6246398 1.6062397 1.5878396 1.5694395 1.5510395 1.5326394 1.5142393
1.4958392 1.4774392 1.4590391

[14] 1.4406390 1.4222389 1.4038388 1.3854388 1.3670387 1.3486386 1.3302385 1.3118384 1.2934384
1.2750383 1.2566382 1.2382381 1.2198381

[27] 1.2014380 1.1830379 1.1646378 1.1462377 1.1278377 1.1094376 1.0910375 1.0726374 1.0542373
1.0358373 1.0174372 0.9990371 0.9806370

[40] 0.9622370

# Plot the absolute differences plot(x_range, absolute_differences, type = "b", main =
"Absolute Differences", xlab = "Trial", ylab = "Absolute Difference")
```

This will give the absolute value of the separation between the red and blue lines at each trial.

Title: Absolute differences across trials

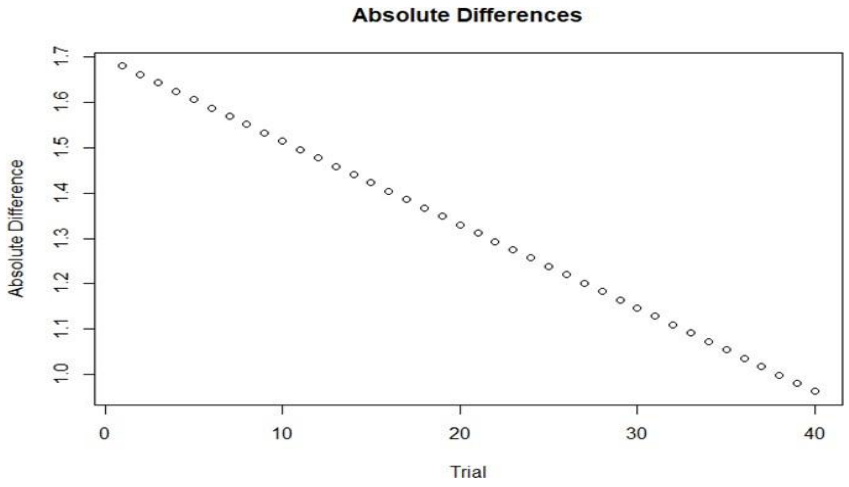


Figure 31: This figure illustrates the visualized absolute value demonstrating the separation between red and blue lines across trials.

The second method:

For my second method, I calculated the difference between the two functions (fun_m3_newitem and fun_m3_olditem) at each point along the x-axis (trial_no) as a direct way to measure accuracy.

Here's how I calculated the accuracy differences at each trial:

1- I defined the range of x-values (trials) over which the functions were defined. In this case, the range was from 1 to 40.

2- I created a vector of x-values (trials) within this range, for example, using the seq() function.

3- I calculated the difference between the two functions at each x-value (trial) in the range.

Here's a code example in R:

Define the range of x-values (trials) x_range <- seq(1, 40, by = 1)

Calculate the differences between the two functions at each x-value (trial) differences <- fun_m3_newitem(x_range) - fun_m3_olditem(x_range)

Display the differences print(differences)

[1] -1.6798400 -1.6614399 -1.6430399 -1.6246398 -1.6062397 -1.5878396 -1.5694395 -1.5510395 -1.5326394 -1.5142393 -1.4958392

[12] -1.4774392 -1.4590391 -1.4406390 -1.4222389 -1.4038388 -1.3854388 -1.3670387 -1.3486386 -1.3302385 -1.3118384 -1.2934384

[23] -1.2750383 -1.2566382 -1.2382381 -1.2198381 -1.2014380 -1.1830379 -1.1646378 -1.1462377 -1.1278377 -1.1094376 -1.0910375

[34] -1.0726374 -1.0542373 -1.0358373 -1.0174372 -0.9990371 -0.9806370 -0.9622370

The differences vector represents the accuracy differences between the red and blue lines at each trial.

Objective 2: To find an analogy to the absolute value of the red and blue lines:

(Continuation of the second method used to achieve objective 1)

To calculate the absolute level of either the red or blue line at a specific trial, I can use their respective functions (fun_m3_newitem for the blue line and fun_m3_olditem for the red line) and apply the abs() function to obtain the absolute value of the function's output.

For example, let's say I wanted to calculate the absolute level of the blue line at trial 5 (new items):

Calculate the value of the blue line (fun_m3_newitem) at trial 5 blue_value <- fun_m3_newitem(5)

Calculate the absolute value of the blue line at trial 5 abs_blue_value <- abs(blue_value)

Display the result print(abs_blue_value)

Output: (Intercept)

2.842949

Similarly, I can calculate the absolute level of the red line at trial 5:

R code

Calculate the value of the red line (fun_m3_olditem) at trial 5 red_value <- fun_m3_olditem(5)

Calculate the absolute value of the red line at trial 5 abs_red_value <- abs(red_value)

Display the result print(abs_red_value)

Output: (Intercept)

4.449189

An overview of what I did:

Regarding the first objective:

The of two separate functions, fun m3 newitem and fun m3 olditem, were represented by the red and blue lines, and I investigated various approaches to measure the spacing between them. Using numerical integration methods in R, I estimated the difference between the two estimated lines. The two suggested methods yielded identical results.

Regarding the second objective:

Using the linear mixed-effects model that I've used earlier to fulfill objective 1, I was able to estimate the absolute values at trials of both red and blue lines.

Final notes:

I genuinely hope that the work that I've presented in this document addresses the key points that we have come across during my thesis proposal in a way that takes me closer towards finalizing my thesis analysis and contributes to the knowledge on response bias and sensitivity.

E- Different ways of visualizing the Multilevel models

The first plot titled "Group-Level Collapsed Lines" would display group-level collapsed trajectories of responses over time. This plot does not differentiate lines by individual subjects (subj). Instead, it shows a single smoothed line in black, representing the overall trend of responses across all subjects. The line is created using the lm method and represents the collapsed

trajectory of the entire group.

Additionally, the plot includes two more lines represented by the functions fun_m3_newitem and fun_m3_olditem. These lines are colored blue and red, respectively. Similarly, these red and blue lines represent the functions of the estimated responses for old (red) and new (blue) lines modeled across trials or test positions.

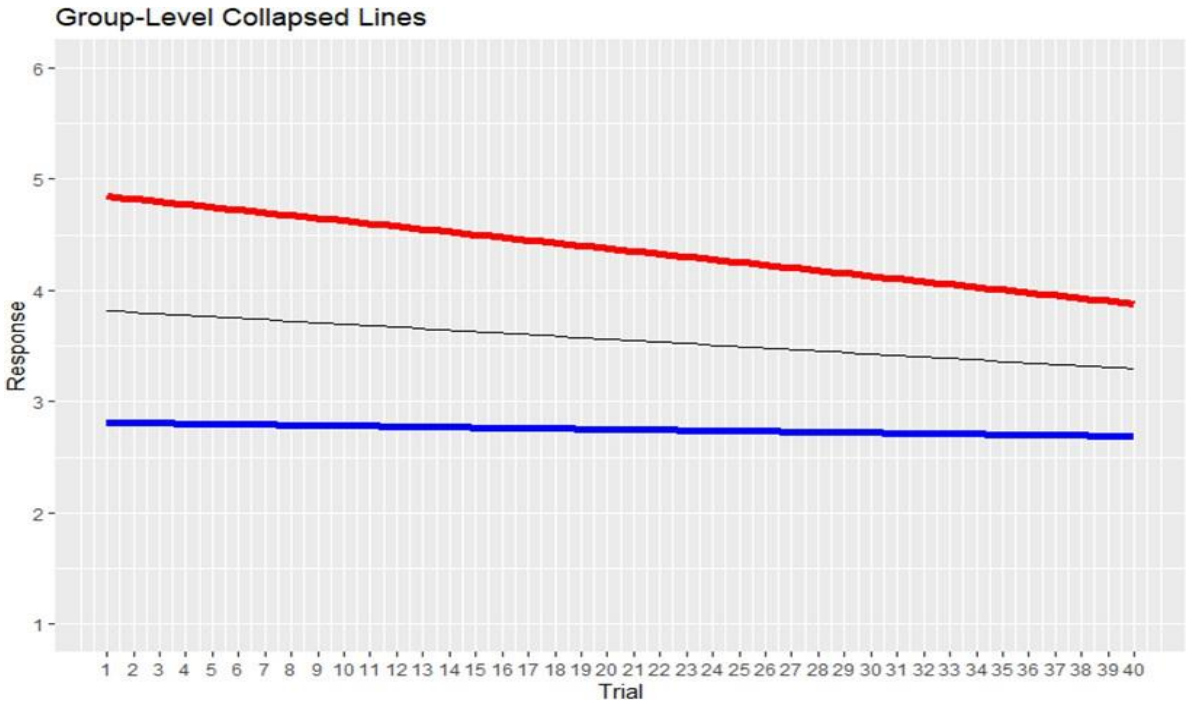


Figure 32: Figure showing the group-level collapsed line and the individual-level collapsed lines, which would highlight the overall trend in responding across trials.

2- Plot of group-level sensitivity and bias across trial: The provided plot shows the mean values of variables mean_c and mean_dprime as a function of trials, along with 95% confidence interval (CI) error bars. Note that c is calculated at each trial (1-40) as well as dprime. Thus mean_c is the mean of c across all trials. The same applies for mean_dprime.

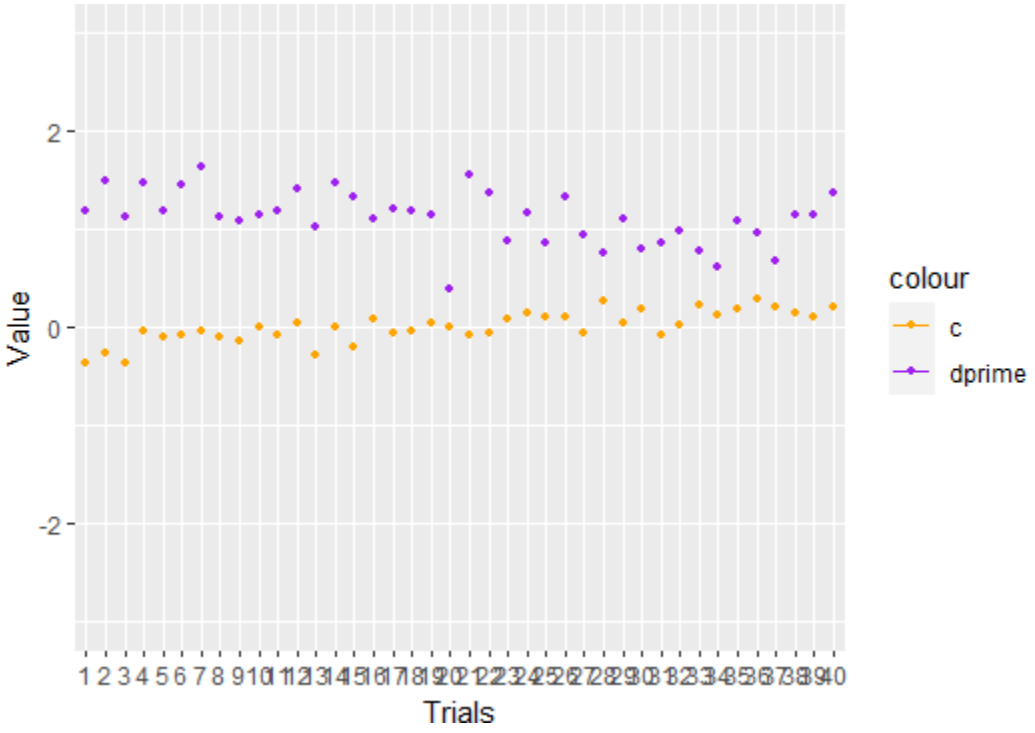


Figure 33: ***Plot of group-level***