Contents

Chapter 1

Some example of some normal text. Some example of some normal text. Some example of some normal text. Some example of some normal text. Some example of some normal text. Some example of some normal text. Some example of some normal text. Some example of some normal text. Some example of some normal text.

Chapter 2

Some example of some normal text. Some example of some normal text. Some example of some normal text. Some example of some normal text. Some example of some normal text. Some example of some normal text. Some example of some normal text. Some example of some normal text. Some example of some normal text.

Chapter 1

Lit Review

Hamilton and Geraci (2006)

IMPLICIT MEMORY: PSE results from conceptual processing of a picture's distinctive features (rather than semantic information). General semantic task: "What is a used car sometimes called?" No PSE. Distinctive conceptual information task: "What fruit is egg shaped?" PSE.

EXPLICIT RECOGNITION: PSE always evident?

aMCI Show larger PSE effects than controls.

Impaired REC, so this PSE must rely on FAM?

Mixed findings whether fam is intact in aMCI. Intact - generally use picture stim. Impaired - generally use verbal stim.

Is PSE in aMCI driven by intact FAM for pictures, but impaired FAM for words? Yes (Embree, Budson, & Ally, 2012): aMCI - Picture FAM - same as healthy OAs aMCI - Word FAM - impaired compared to healthy OAs

Ally, McKeever, 2009: Examined early frontal old/new effect (FAM) in aMCI: Intact for pictures. Impaired for words. BUT, P did not provide subjective Rec/Fam reports.

Embree, Budson, & Ally, 2012: Deep encoding (verbal like/dislike response). Modified Old/New (6-point rating scale): 6. Certain the item is old - to - 1. Certain the item is new.

Both used the same picture stim - colour photos.

###———————————————-

Chapter 2

Dual-process theories of recognition memory suggest that two independent processes - recollection and familiarity - are implicated in the successful recognition of previously encountered material (Paivio, 1971, 1972). Recollection typically refers to the conscious recall of encoded information, whereby contextual details (usually obtained by mentally re-experiencing a previous encounter with the stimulus) facilitate successful recognition. Familiarity, on the other hand, describes the unsubstantiated *feeling* of having encountered the stimulus before, and despite the inability to retrieve any associated diagnostic information, is still able to produce accurate recognition (Schoemaker, Gauthier, & Pruessner, 2014). While single-process accounts of recognition memory have been proposed, with the view that such experiences can be understood simply as varying levels of memory strength (Dunn, 2008; Squire, Wixted, & Clark, 2007), the majority of memory researchers agree that multiple processes are necessary to account for a range of dissociable experimental findings (Yonelinas, 2002). Evidence from studies utilising event related potentials (ERPs; Curran & Doyle, 2011), functional magnetic resonance imaging (fMRI; Scalici, Caltagirone, & Carlesimo, 2017) and comparisons between healthy and clinical subject groups (e.g. Mild Cognitive Impairment; Belleville, Ménard, & Lepage, 2011) all implicate the existence of two functionally distinct processes. Despite this consensus, disagreement persists in the literature regarding the extent to which recollection and familiarity are independent, and the methods that should be used to measure them most effectively (Schoemaker et al., 2014; Yonelinas, 2002).

Experiments into recognition memory often focus on obtaining separate estimates of recollection and familiarity using process-estimation methods (Yonelinas, 2002). The most commonly used process-estimation method is the Remember/Know (RK) paradigm (Tulving, 1985) - a task endorsed by a wide body of literature (Gardiner, 2000; Jacoby, 1991; Jacoby, Yonelinas, & Jennings, 1997; Yonelinas & Jacoby, 1995). In a typical RK procedure, participants are generally tasked with making ‘old’ vs. ‘new’ recognition decisions toward a randomised list of items, many

of which were presented during an earlier encoding phase (targets) amongst novel items with highly similar characteristics (lures). When a subject recognises an item, and thus selects *Old*, a follow-up judgement probes how they arrived at this decision (*Remember* or *Know*). If the subject was able to recognise the item based on recollection (i.e. conscious recall of some diagnostic information: "I remember seeing this item earlier"), they should classify their recognition as *Remember*. If the subject arrived at their recognition decision due to familiarity (i.e. a feeling of certainty that the item was studied in the encoding phase, but unable to recall and details: "I know I saw this item earlier, but cannot determine why"), they should classify their recognition as *Know*. In addition to the literature endorsing the task in healthy samples, a large body of research also reports that the RK procedure produces reliable estimations of recollection and familiarity in clinical populations (Lombardi, Perri, Fadda, Caltagirone, & Carlesimo, 2016); for example, those with Mild Cognitive Impairment (MCI) typically produce results to suggest recollection impairments but intact familiarity compared to healthy older adults (Belleville et al., 2011; Hudon, Belleville, & Gauthier, 2009; Lombardi et al., 2016; Serra et al., 2010; Wang et al., 2013).

The RK procedure has been modified in a number of ways since its conception, and continues to adapt as understandings of recollection and familiarity processes evolve. An early development was the "independence correction" - a formula devised to 'correct' the inherent underestimation of familiarity processes within the mutually exclusive paradigm (Yonelinas & Jacoby, 1995). Participants are generally only instructed to select *Know* (a reflection of familiarity) when there is an absence of recollection, however, this approach does not allow for the possibility of recollection and familiarity co-occurring. Proportions of *Know* responses will likely always be lower than *Remember* if subjects do indeed perceive to experience both processes simultaneously, since the presence of recollection necessitates that they select the *Remember* option among the two choices. When the Yonelinas & Jacoby (1995) independence correction is applied, estimates of familiarity are determined by also taking into account the number of times *Remember* was selected when calculating the proportion of *Know* responses (Schoemaker et al., 2014). An alternative to this correction is to modify the response options available to subjects, so they are

able to individually determine the relative contributions of each process. Higham & Vokey (2004) proposed an independent ratings methodology whereby, instead of the binary *Remember*/*Know* options, subjects are provided with one rating scale to report the contribution of recollection and another to report the contribution of familiarity (RF-Ratings). Participants rate their recognition experience for each process accordingly: 1 = *definitely no*, 2 = *probably no*, 3 = *probably yes*, 4 = *definitely yes*. Such options allow for great variability in the way participants are able to respond, and for the possibility of both processes occurring conjointly: i) Recollection without Familiarity (high rating on R, low rating on F); ii) Familiarity without Recollection (high rating on F, low rating on R); iii) both Recollection *and* Familiarity (high rating on R and F); iv) neither R or F, i.e. a guess (*1* rating on R and F). The methodology of Higham & Vokey (2004) has been used in numerous studies (Brown & Bodner, 2011; Kurilla & Westerman, 2008; Tousignant & Bodner, 2012), however, it could be argued that this rating task is somewhat removed from the original *judgement* task, and the extent to which the increased task complexity affects reports of recognition is unknown (Tousignant, Bodner, & Arnold, 2015).

Further modifications retain the original two binary response options, but avoid the mutual exclusivity issue by simply including a *Both* option (Tousignant et al., 2015). When calculating proportions of recollection and familiarity, the total proportion of *Both* responses can then be separately added to the totals for each process. Recent adaptations of the RK paradigm have also begun to include a *Guess* response option, allowing participants to report uncertainty in their recognition decision (Belleville et al., 2011; Eldridge, Sarfatti, & Knowlton, 2002; Larsson, Öberg, & Bäckman, 2006; Tunney & Fernie, 2007; Williams, 2019). Previous studies have found that subjects may falsely assign guesses to the *Know* option when there is no explicit *Guess* option available (Gardiner, Java, & Richardson-Klavehn, 1996; Gardiner & Ramponi, 1998; Gardiner, Ramponi, & Richardson-Klavehn, 2002), on the assumption that this option more closely resembles their state of low confidence (Tunney & Fernie, 2007). Responding in this manner may artificially inflate obtained estimates of familiarity (Tunney & Fernie, 2007). By including *Guess*, the likelihood of obtaining false *Know* responses (i.e. those that do not reflect underly-

ing familiarity processes) is reduced (Migo, Mayes, & Montaldi, 2012).

Despite its widespread use, the RK procedure has been criticized for its reliance on participants' subjective understanding of the provided instructions (Schoemaker et al., 2014), and the introspective nature of recognition judgements make it difficult to confirm whether all participants have understood the definitions (and thus responded) similarly (Lombardi et al., 2016). It is also difficult to determine whether subjects interpret the *Remember* and *Know* labels in the same way that researchers intend (Umanath & Coane, 2020), especially as there is evidence to suggest participants struggle to understand the distinction between the terms (Geraci, McCabe, & Guillory, 2009; Rubin & Umanath, 2015; Williams & Moulin, 2014). Williams (2019) assessed the ways in which non-recollective subjective experiences were defined to participants, and found a great deal of inconsistency across a range of RK experiments. Some studies even changed the *Remember* and *Know* labels altogether; many exchanged the *Know* label with *Familiar* in an effort to reduce subjects defaulting to colloquial understandings of the word "know" which typically indicate high certainty; e.g. "I **know** I saw this item in the study phase" (Bastin, Van der Linden, Michel, & Friedman, 2004; Dobbins, KroU, & Liu, 1998; Donaldson, Mackenzie, & Underhill, 1996; Ingram, Mickes, & Wixted, 2012). Others also substitute *Remember* for *Recollection* (Harlow, MacKenzie, & Donaldson, 2010). Labels that accurately match the processes they intend to measure - *Recollection* and *Familiarity* - have been proposed in an effort to reduce the potentially misleading effects of the more colloquial *Remember* and *Know*, and thus make it easier for participants to 'map on' the definitions provided by researchers (Harlow et al., 2010; Mayes, Montaldi, & Migo, 2007).

In addition to the availability of different response options, and the labels used to describe the underlying processes, there is evidence to suggest that the format of to-be-remembered stimuli also plays a role in obtained estimates of recollection and familiarity. The Picture Superiority Effect (PSE) refers to a robust phenomenon whereby stimuli presented as pictures are markedly

better remembered on tests of recall or recognition than stimuli presented as words (Shepard, 1967). There is general agreement that, in recognition memory paradigms, picture superiority manifests as enhanced recollection rather than familiarity (Curran & Doyle, 2011; Rajaram, 1996a). Word stimuli, on the other hand, appear to produce increased familiarity ratings at test (Ally & Budson, 2007). Understanding this phenomenon could help to conceptualise how memory breaks down in healthy ageing, and in the earliest stages of amnestic Mild Cognitive Impairment (aMCI). For example, Ally et al. (2008) demonstrated that, despite similar levels of overall performance on a recognition task, healthy older adults showed greater picture superiority effects than younger adults. The memorial benefit of pictures was indeed evident in both the young and older groups, but the magnitude of this effect was greater in older adults, who only showed worse performance when responding to word stimuli. Interestingly, picture superiority also allows those with aMCI to show performance that is comparable to healthy older adult controls; despite exhibiting impaired performance overall, those with aMCI often show intact familiarity processes when pictures are utilised in recognition memory paradigms (and impaired familiarity when word stimuli are utilised; Embree, Budson, & Ally (2012); Ally et al. (2009a); Ally et al. (2009b); Wolk, Signoff, & DeKosky (2008); Algarabel et al. (2009); Anderson et al. (2008); Hudon et al. (2009); O'Connor & Ally (2010); Serra et al. (2010); Westerberg et al. (2006)].

The objective of the current programme of research is to better understand how different methodologies inform understandings about the underlying processes of recollection and familiarity. Across a number of experiments, the distinctiveness of to-be-remembered stimuli will be systematically examined to determine the level at which successful recognition is impacted, and which process(es) are most susceptible. The aim of the first experiment, outlined below, is to establish baseline PSE response patterns in a novel, modified RK paradigm. In a 2x3 mixed factorial design, a within-subjects variable of stimulus type (words / simple pictures) will be used to determine whether the magnitude of picture superiority effects (PSEs) is mediated by the particular response options available at test (between-subjects variable of response option: RFG, RFBG, RF-Ratings). In each condition, the labels *Recollection*/*Familiarity* will be used in place

of the standard *Remember*/*Know*, in an effort to reduce the impact of colloquial understandings on the current experimental definitions. To avoid guesses biasing estimations of familiarity (Belleville et al., 2011; Eldridge et al., 2002; Larsson et al., 2006; Tunney & Fernie, 2007; Williams, 2019), participants in all response-option conditions are also given the option to report that they are merely *Guessing* that an item is old. At test, subjects will be presented with either i) three response options (RFG); ii) four response options (RFBG; where a *Both* response option allows subjects to report the co-occurrence of R and F; iii) separate 0-5 rating scales for R and F (where subjects could report either process occurring alone, both processes occurring conjointly, or that they are guessing by providing a '0' rating on both scales). To establish whether a PSE is evident in the current paradigm, *d'* (d-prime) scores will be calculated for each participant. *d'* is a signal detection statistic, calculated by taking the standardised difference between the signal (i.e. correct hits) and signal+noise (i.e. false alarms); in other words, *d'* offers a representation of global recognition performance and participants' ability to distinguish target items from lures (Wixted, 2014). Higher *d'* scores demonstrate better overall performance on the memory task. Based on the discussed research, the following results are hypothesised:

1. **Overall PSE**: a PSE will be evident within the current paradigm, manifesting as:
    - i) higher overall *d'* scores for pictures compared to words;
    - ii) higher proportion of correct hits;
    - ii) lower proportion of false alarms;
    - iv) better overall recognition.
2. **PSE in rates of Recollection and Familiarity**:
    - i) pictures will produce a higher proportion of R hits and a lower proportion of R FAs than words.
    - ii) words will produce a higher proportion of F hits and a higher proportion of F FAs than pictures.
3. **PSE and the availability of different response options**:

- i) comparable PSEs will be evident in each of the response option conditions (RFG, RFBG, RF-Ratings).
- ii) the availability of different response options will affect whether a PSE manifests as increased recollection, increased familiarity, or both (RFBG, RF-Ratings).

Experiment 1: Establishing PSEs in novel Remember/Know paradigm

Method

Participants A total of 186 subjects completed the online experiment (M = 26.7 years (SD = 10.36 years; see Table 1 for a comprehensive breakdown of the sample). The current sample was primarily comprised of participants sourced from voluntary participation websites such as Prolific Academic (52.15%) (where they received payment at the rate of £5/hr) and via the in-school research participation system (where they received course participation credits; 41.4%). A small number of participants were also recruited from social media and other online sources (Facebook: 3.76%; Call For Participants: 1.61%; Reddit: 0.54%; unspecified: 0.54%). To meet our YA requirements, all participants were required to be between 18-59 years of age (actual range: 18-59). As our experiment involved English word stimuli, we also asked subjects whether English was their first language; the vast majority (93.01%) reported that English was indeed their first language.

Table 1: Gender and age (*SD*) of the current sample.

Gender	N	Age	
Female	*122*	26.02	*(10.04)*
Male	*60*	28.10	*(10.98)*
Non-binary	*2*	19.50	*(2.12)*
Unspecified	*2*	39.00	*(0)*
Total	***186***	**26.70**	***(10.36)***

Materials Pictures of innocuous, everyday objects (e.g. clock, rabbit, shoe) and their written-word names were sourced from Rossion & Pourtois (2004). The picture stimuli consisted of greyscale line-drawn illustrations (containing shaded surface details), while word stimuli were

simply the written-word names of each object presented in a clear Sans-serif typeface. A total of 136 unique items were randomly selected for use in the current experiment, from a pool consisting of: i) items with a written name between 4 and 7 letters; ii) items that would conjure the same intended concept in our UK-based sample (e.g. "ladder" should be universally understood across English-speaking cultures, whereas "wagon" or "pants" can be interpreted differently); iii) items that were not unknown, or uncommon, for our sample (e.g. Americanisms such as "wrench"); and iv) non-specific concepts such as "bird" (since the pool of items already contained specific exemplars of birds, such as "peacock" and "penguin"). As the current experiment involved memorising word stimuli, a single item ("glass") was also removed as it shared too many letters with another item ("glasses"). Selected items were split into four separate lists for counterbalancing purposes; using the normative data provided by Rossion & Pourtois (2004), each list was balanced based on the length of the written name, as well as scores of naming accuracy, familiarity, visual complexity, and mental imagery agreement. A series of independent samples t-tests confirmed that no list was significantly different from another on any of the aforementioned criteria.

The picture stimuli utilised in the current study were created in Photoshop CC (20.0.04 Release), by importing the greyscale, surface-shaded, line-drawings onto a plain 250x250px white canvas. Written word stimuli were created using the Calibri sans-serif typeface on the same size canvas (see Figure 1 for example stimuli). All items were exported as .pngs files for presentation by the online survey platform.

bottle	ladder	orange	shirt
			

Figure 1: Example word and picture stimuli from the current study.

Design The current study utilised a mixed design, with a 2-level within-subjects factor of stimuli format (words, drawings), and a 3-level between-subjects factor of response option (RFG, RFBG, RF-Ratings). Subjects completed two study blocks - one consisting only of word stimuli, the other consisting only of picture stimuli - before completing a single mixed format recognition test, where previously studied word and picture items were randomly shown among new, unseen items. Subjects passed through 2 levels of blocked randomization during the experiment (equally sized, predetermined blocks). First, subjects were randomly allocated into one of two study block orders, which determined the order in which they were presented with the picture and word blocks at study. Second, subjects were assigned into one of three possible recognition tests (identical aside from the response options available when categorising recognition experiences): 1) RFG: "Recollection", "Familiarity","Guessing"; 2) RFBG: "Recollection", "Familiarity", "Guessing", "Both", or 3) RF-Ratings: two independent 0-5 rating scales to separately report the contribution of Recollection and Familiarity. These randomisation processes were completed automatically by the experiment software using balanced methods.

Procedure Data collection was conducted via the online survey platform Qualtrics. Subjects initially completed an encoding block, where target words and pictures were randomly presented one-at-a-time on-screen. To ensure attention was directed to the presented stimuli, participants were required to respond to a simple encoding question toward each item at study: "Is this a picture or a word?". This question allowed for the assessment of performance during the study block (to determine whether participants were concentrating at study), whilst also avoiding potential levels-of-processing effects that can accompany deeper encoding judgements (e.g. pleasantness ratings). The encoding phase was followed by a short distractor task comprised of 20 multiplication sums. Finally, subjects completed the recognition task, where they were again randomly presented with word and picture items one-at-a-time on-screen, and were required to respond *Old*/*New* depending on whether they recognised the item or not. *Old* responses were

succeeded by a follow-up screen whereby participants were asked to report their recognition experience for the current item; the response options available during this follow-up response page differed between participants, with random allocation into either the RFG, RFBG, or RF-Ratings response option conditions. Recollection and Familiarity were defined identically across conditions, and the only deviations in instructions were: i) to define the additional "Both" response option in the RFBG condition; and ii) explain how certain responses should be reported in the RF-Ratings condition (i.e. subjects could still report a "Guess" in this condition by providing a 0-rating on both of the scales).

Data processing Measured variables included the total number of hits and FAs, and the total number of hits and FAs assigned to each of the available response options (RFG, RFBG, and RF Ratings). In order to create a common dependant variable, proportions were calculated from these variables in slightly different ways depending on the response option group. In the RFG-judgement group, simple proportions were created from the total number of R responses and the total number of F responses. In the RFBG condition, similar proportions were calculated by separately adding the proportion of Both responses to the proportion of R and proportion of F responses. In the RF-Ratings group, proportions of R and F were calculated based on the number of responses scoring +>3; a response was classified R when subjects rated between 3-5 on the "Recollection" scale (regardless of the Familiarity rating), and a response was classified F when subjects rated between 3-5 on the "Familiarity" scale (regardless of the Recollection rating). The scales therefore allowed for pure R responses (R=3-5 + F=0-2), pure F responses (F=3-5 + R=0-2), both responses (R=3-5 + F=3-5) and Guessing responses (R=0 + F=0). Additional DVs included: i) d' (d-prime, a signal detection measure of sensitivity); ii) c-value (a measure of response bias); iii) overall accuracy (hits / (hits + FAs)); iv) reaction times for all responses.

All analyses were conducted using R (R Core Team, 2020). *d'* (sensitivity) and *c* (bias) scores were calculated using the 'psycho' package (v0.5.0; Makowski, 2018). *d'* scores were calculated via: z-scores for correct hits minus z-scores for false alarms (Hautus, 1995 adjustments for extreme values were applied). *c* scores were calculated

A series of exclusion criteria were defined before analysis. First, subjects were to be excluded from analysis if they showed poor performance during the encoding task; the relative ease of reporting whether each item was shown as a word or picture prompted a performance cut off of 90% accuracy. This would allow for some accidental clicks, though subjects scoring less than 90% were to be excluded on the assumption they did not dedicate their full attention to the task. Second, subjects would be considered outliers (and thus excluded from analysis) if they presented extreme z-scores of +/- 3 for total hits, total FAs, or overall recognition (hits minus FAs). However, no subjects were found to meet any of these criteria.

Results

Picture superiority To establish baseline picture superiority effects in the current paradigm, and assess whether there were any interactions with the availability of different response options at test, a series of 2 (stimuli format: words, pictures) x 3 (response option condition: RFG-judgements, RFBG-judgements, RF-ratings) mixed ANOVAs were conducted on a number of outcome variables. Namely, the signal detection measures of *d'* (sensitivity) and *c* (decision criterion), as well as the proportion of overall hits, false alarms (FAs), and overall recognition (hits - FAs) [see Table 2]. Significant main effects and interaction effects were followed-up with Bonferroni-adjusted pairwise comparisons.

Table 2: Mean *d'* (sensitivity), *c* (decision criterion), proportion of hits, FAs, and overall recognition by stimuli-format and response-option condition.

	d'	c	Hits	FAs	Overall recognition
Stimuli-format					
Words	0.86	0.53	0.47	0.21	0.27
Pictures	1.62	0.48	0.62	0.12	0.50
Response-option					
RFG	2.71	0.67	0.62	0.19	0.43
RFBG	2.41	1.00	0.54	0.16	0.38
RF Ratings	2.32	1.33	0.48	0.14	0.34

The ANOVA on *d'* scores demonstrated a significant main effect of stimuli-format, $F(1, 183) = 295.80$, *MSE* $= 0.18$, $p < .001$, $\hat{\eta}^2_G = .223$; a PSE was evident, with pictures ($M = 1.62$) producing significantly better discrimination between hits and FAs than words ($M = 0.86$), $t(183) = -17.20$, $p < .001$. The ANOVAs on the proportion of hits, FAs, and overall recognition also produced findings consistent with a PSE. For hits, there was a significant main effect of stimuli-format, $F(1, 183) = 131.77$, *MSE* $= 0.01$, $p < .001$, $\hat{\eta}^2_G = .092$, with pictures ($M = 0.62$) showing a higher number of overall hits compared to words ($M = 0.47$), $t(183) = -11.48$, $p < .001$. Similarly, the ANOVA on the proportion of FAs showed a significant main effect of stimuli-format $F(1, 183) = 61.18$, *MSE* $= 0.01$, $p < .001$, $\hat{\eta}^2_G = .084$, with words ($M = 0.21$) producing more FAs than pictures ($M = 0.12$), $t(183) = 7.82$, $p < .001$. Overall recognition performance (a measure that takes into account both hits and FAs) offered further support for a PSE in the current paradigm; a significant main effect of stimuli-format $F(1, 183) = 409.20$, *MSE* $= 0.01$, $p < .001$, $\hat{\eta}^2_G = .236$ showed pictures ($M = 0.50$) produced better overall recognition on the task compared to words ($M = 0.27$), $t(183) = -20.23$, $p < .001$. Finally, *c* scores showed no significant main effect of stimuli-format, $F(1, 183) = 2.31$, *MSE* $= 0.11$, $p = .130$, $\hat{\eta}^2_G = .002$, suggesting response biases were similarly conservative between pictures and words. No interaction effects were found between stimuli format and response option for any of the variables.

Taken together, the findings demonstrate a replication of the PSE in the current memory paradigm, and suggest stimuli format plays a key role in memorability that is independent from the particular response options available to participants. The current findings support the hypotheses of a PSE manifesting as i) higher overall d' scores for pictures compared to words, ii) a higher proportion of correct hits, iii) lower proportion of false alarms, and iv) better overall recognition.

PSE in rates of Recollection and Familiarity: To determine the impact of stimuli format on the rates of R and F, additional 2 (words, pictures) x 3 (RFG-judgements, RFBG-judgements, RF-ratings) mixed ANOVAs were conducted on the mean proportion of hits (see Figure 2) and FAs (see Figure 3) assigned R, F, and G.

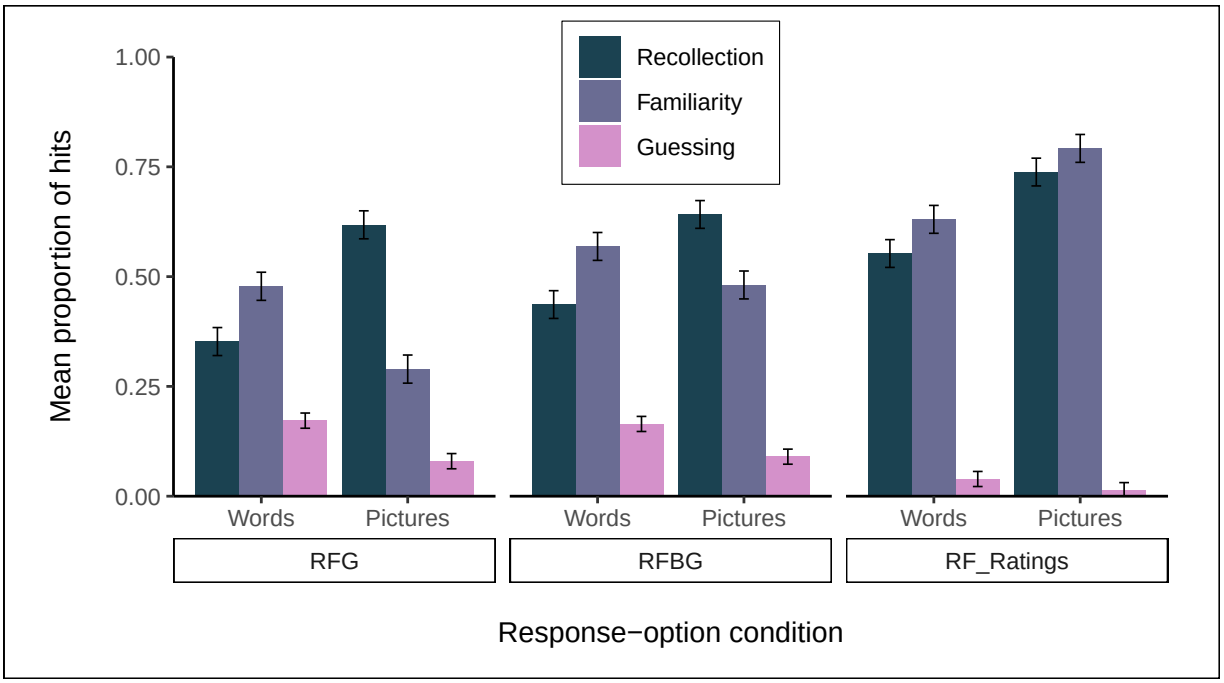


Figure 2: Proportion of hits assigned *Recollection*, *Familiarity*, and *Guessing*, by stimuli-format and response-option condition.

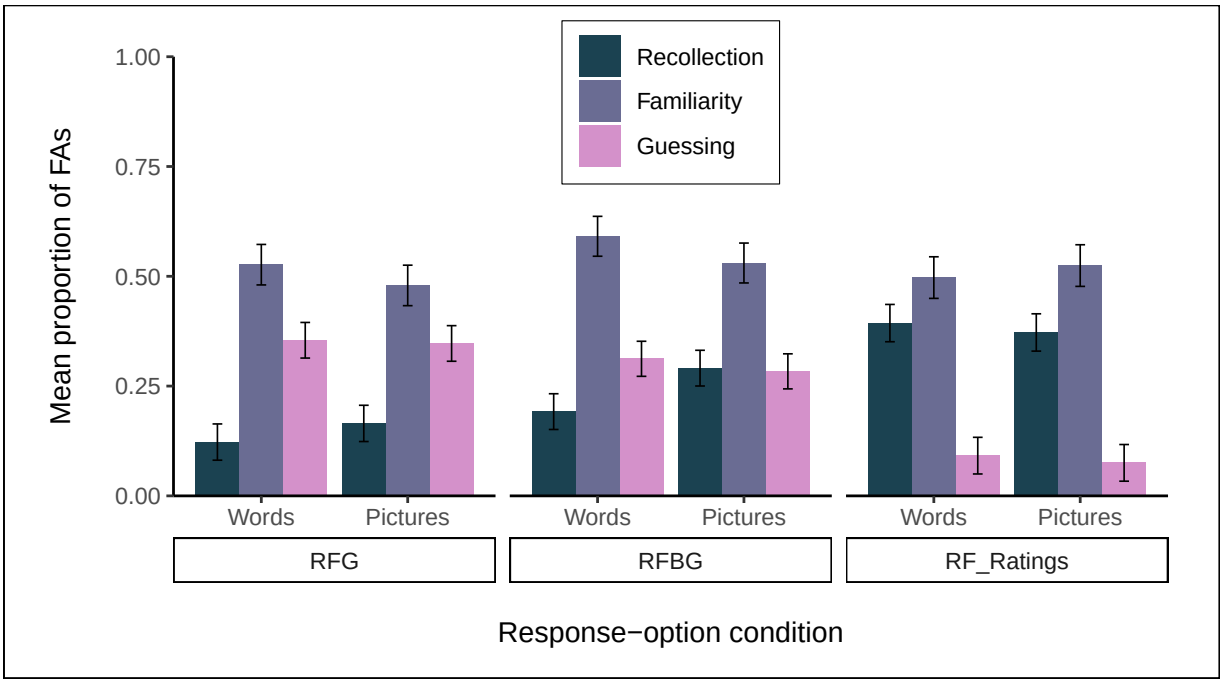


Figure 3: Proportion of FAs assigned *Recollection*, *Familiarity*, and *Guessing*, by stimuli-format and response-option condition.

Recollection: For R hits, there was a significant main effect of stimuli-format, $F(1, 178) = 158.42$, *MSE* $= 0.03$, $p < .001$, $\hat{\eta}^2_G = .167$; pictures ($M = 0.67$) showed a higher proportion of Recollected hits than words ($M = 0.45$), . For R FAs, there was no significant main effect of stimuli-format, $F(1, 136) = 2.78$, *MSE* $= 0.04$, $p = .098$, $\hat{\eta}^2_G = .005$. There were no significant interaction effects between stimuli format and response option condition for R hits or FAs.

Familiarity: For F hits, there was a significant interaction between stimuli format and response option, $F(2, 178) = 34.42$, *MSE* $= 0.03$, $p < .001$, $\hat{\eta}^2_G = .083$ (see Figure 4). Words resulted in more Familiarity hits than pictures in both the RFG group (words: $M = 0.48$; pictures: $M = 0.29$, $t(178) = 6.07$, $p < .001$) and RFBG group (words: $M = 0.57$; pictures: $M = 0.48$, $t(178) = 2.87$, $p = .005$). Conversely, pictures ($M = 0.79$) resulted in more Familiarity hits than words ($M = 0.63$) in the RF-Ratings group, $t(178) = -5.29$, $p < .001$. The ANOVA on Familiarity FAs did not yield any significant results; with no significant main effect of stimuli-format, $F(1, 136) = 1.12$, *MSE* $= 0.04$, $p = .292$, $\hat{\eta}^2_G = .002$ or significant interaction effects, $F(2, 136) = 1.12$, *MSE* $= 0.04$, $p = .331$, $\hat{\eta}^2_G = .004$.

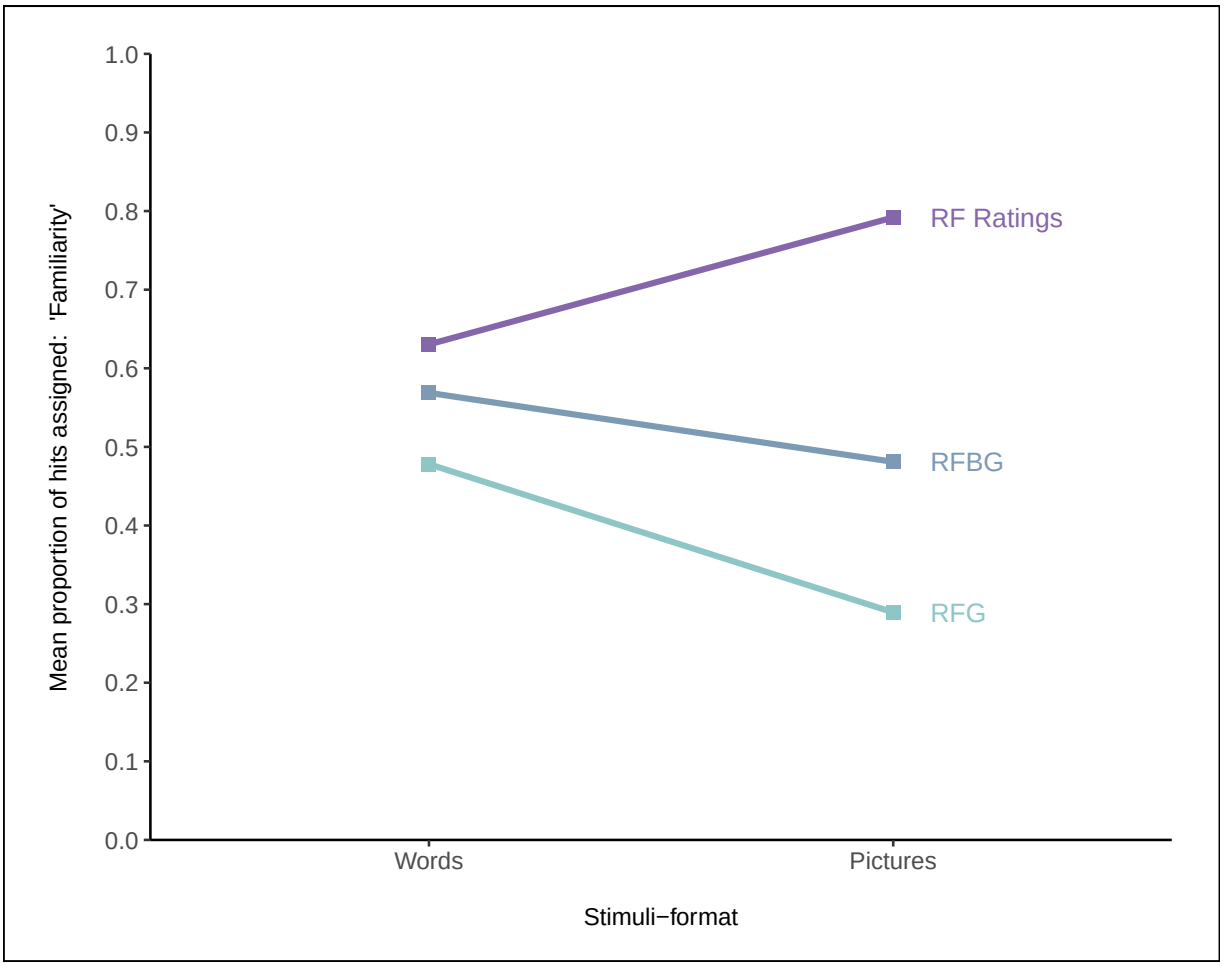


Figure 4: Interaction plot between stimuli-format and response-option condition for the mean proportion of hits assigned *Familiarity*.

Guessing: The ANOVA on Guessing hits also showed a significant interaction between stimuli format and response option,$F(2, 178) = 4.17$, *MSE* $= 0.01$, $p = .017$, $\hat{\eta}^2_G = .011$, (see Figure 5). Words resulted in more Guessing hits than pictures in both the RFG group (words: $M = 0.17$; pictures: $M = 0.08$), $t(178) = 5.38$, $p < .001$; and the RFBG group (words: $M = 0.16$; pictures: $M = 0.09$), $t(178) = 4.42$, $p < .001$. There was no difference in the number of Guessing hits between words $M = 0.04$ and pictures $M = 0.01$ in the RF-Ratings group, $t(178) = 1.51$, $p = .133$. For Guessing FAs, there was no significant main effect of stimuli-format or significant interaction effects.

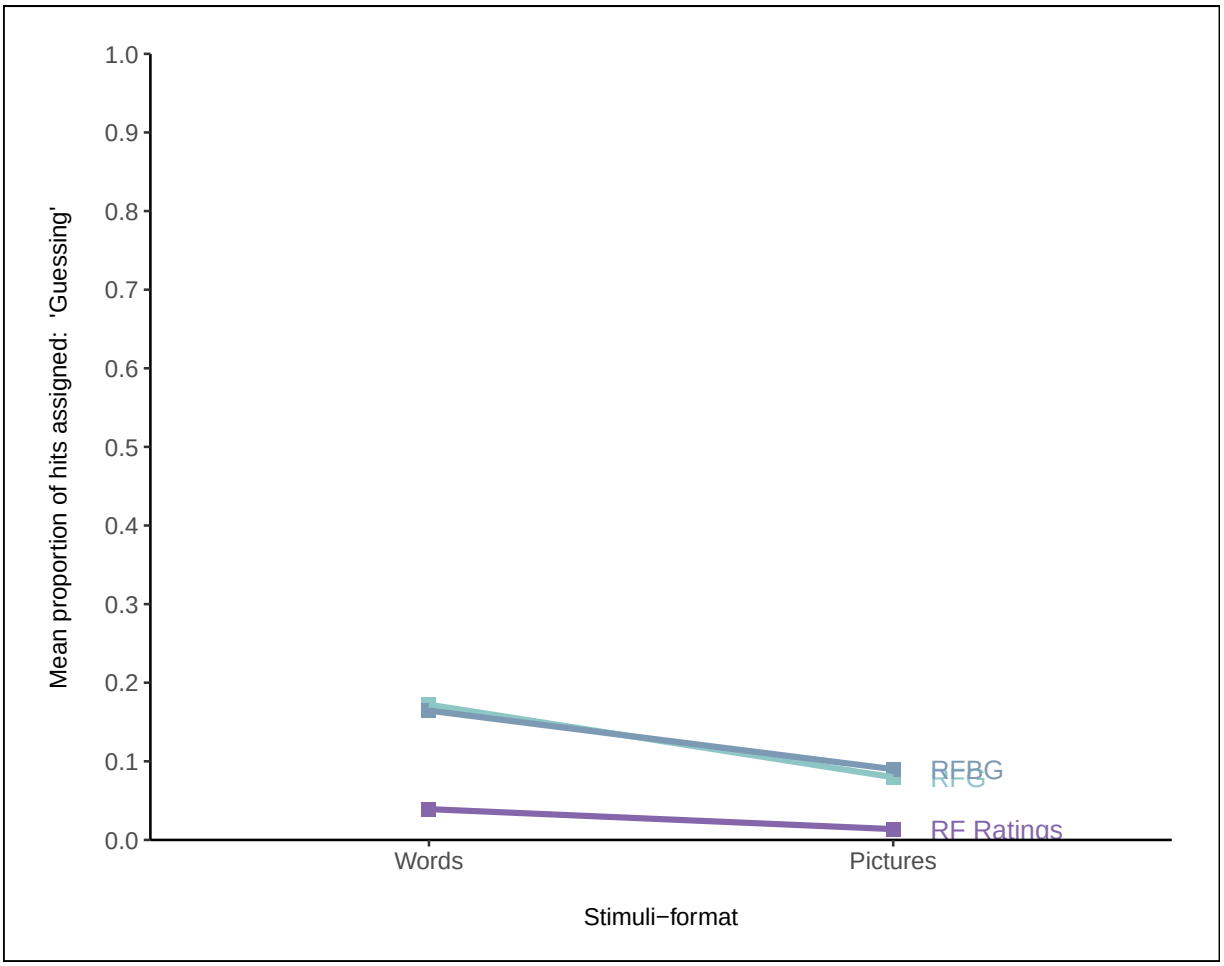


Figure 5: Interaction plot between stimuli-format and response-option condition for the mean proportion of hits assigned 'Guessing'.

Such findings mostly support the proposed hypotheses. Pictures indeed produced a higher proportion of R hits in comparison to words, though no picture superiority was evident in the number of R FAs. This suggests that, despite words showing a decreased level of memorability compared to pictures, they do not elicit high certainty false recognition at any higher rate. Words also produced a higher proportion of F hits compared to pictures, as predicted, but only in the RFG and RFBG conditions. It is unclear why the same pattern was not evident in the RF-Ratings group, aside from the possibility that participants avoided the more complex ratings screen (and instead more often chose *New*, inaccurately), unless they had a high certainty of their recognition (as evidenced for R hits). Again, there was no evidence of picture superiority in regard to the

number of F FAs. The hypotheses put forward for G responses were again mostly supported; there were more guesses made toward words than pictures, however, this again only applied to the RFG and RFBG conditions. Such findings align with the possible explanation outlined above, whereby participants avoided having to provide two separate ratings unless they were very certain they recognised the item.

PSE and the availability of different response options: To determine whether the availability of different options had an impact on picture superiority within the current paradigm, the main effects of response option were also examined in the aforementioned ANOVAs. Discriminability (*d'*) between hits and FAs showed no significant main effect of response option, supporting the hypothesis that a PSE would be evident in each of the response option conditions. Response bias (*c*) scores did show a main effect of response option though, $F(2, 183) = 6.44$, *MSE* $= 0.51$, $p = .002$, $\hat{\eta}^2_G = .054$; those in the RF-Ratings condition ($M = 0.67$) showed higher c-scores (and thus a more conservative response bias) than those in the RFG condition ($M = 0.34$), $t(183) = -3.59$, $p = .001$. This indicates subjects were less likely to respond *Old* when they were required to provide more detailed follow-up recognition judgements (i.e., using separate 0-5 scales for R and F), compared to simply selecting one of three options (R,F, or G). This again supports the notion of avoidance from participants when they were required to provide two separate ratings. The ANOVAs on the proportion of hits, FAs, and overall recognition also mostly supported our hypothesis that a PSE would be evident in each of the response option conditions. While the ANOVAs on the proportion of FAs and overall recognition showed no significant main effects of response option, there was a significant main effect for the proportion of hits, $F(2, 183) = 6.46$, *MSE* $= 0.09$, $p = .002$, $\hat{\eta}^2_G = .057$, with the RFG group ($M = 0.62$) showing more hits than the RF-Ratings group ($M = 0.48$), $t(183) = 3.60$, $p = .001$. This unexpected result may also reflect the conservative response bias of those in the RF-Ratings group, whereby fewer total *Old* responses necessitates fewer hits as a result. However, a lack of differences between groups for FAs and overall recognition indicates a comparable PSE across each of the response option groups.

With regard to rates of R and F, there was a significant main effect of response-option for R hits, $F(2, 178) = 8.55$, *MSE* $= 0.09$, $p < .001$, $\hat{\eta}^2_G = .069$; the RF-Ratings group ($M = 0.65$) showed significantly more R hits compared to both the RFG-group ($M = 0.49$), , and the RFBG-group ($M = 0.54$), . R FA's also exhibited an identical pattern; a significant main effect of response-option, $F(2, 136) = 10.70$, *MSE* $= 0.12$, $p < .001$, $\hat{\eta}^2_G = .106$, showed that those in the RF-Ratings group ($M = 0.38$) produced more Recollection FAs than both the RFG-group ($M = 0.14$), , and RFBG-group ($M = 0.24$),). These findings align with those above to suggest a more conservative response bias in the RF-Ratings group; participants were more likely to respond *Old* in the RF-Ratings condition when they experienced high certainty in their recognition - regardless of whether or not that recognition was accurate or false.

The significant interaction for F hits showed that word stimuli produced significantly more F hits in the RF-Ratings group ($M = 0.63$) compared to the RFG group ($M = 0.48$), $t(276.78) = -3.37$, $p = .002$. An identical pattern was also observed for pictures between the RF-Ratings group ($M = 0.79$) and RFG group ($M = 0.29$), $t(276.78) = -11.13$, $p < .001$, however, F hits in the RF-Ratings group were also higher than the RFBG group ($M = 0.48$) when the stimuli were pictures, $t(276.78) = -6.94$, $p < .001$. The RFBG group ($M = 0.48$) also showed a significantly higher number of F hits compared to the RFG group ($M = 0.29$), $t(276.78) = -4.24$, $p < .001$. There was no significant main effect of response-option for F FAs. Such findings suggest that the proportion of F responses increases when participants are given the option to report *Both* - either explicitly, or from rating high on both rating scales.

Finally, the significant interaction for G hits demonstrated that word stimuli produced significantly fewer G hits in the RF-Ratings group ($M = 0.04$) compared to both the RFG group ($M = 0.17$), $t(281.42) = 5.44$, $p < .001$, and the RFBG group ($M = 0.16$), $t(281.42) = 5.18$, $p < .001$. Likewise, pictures also produced significantly fewer G hits in the RF-Ratings group ($M = 0.01$) compared to both the RFG group ($M = 0.08$), $t(281.42) = 2.70$, $p = .020$

and RFBG group ($M = 0.09$), $t(281.42) = 3.15, p = .005$. For G FAs, a significant main effect of response-option, $F(2, 136) = 15.69$, *MSE* $= 0.11, p < .001, \hat{\eta}^2_G = .144$ revealed that the RF-ratings group ($M = 0.08$) again showed significantly fewer G FAs than both the RFG ($M = 0.35$), , and RFBG groups ($M = 0.30$), . These findings again align with the previous results that suggest those in the RF-Ratings group responded conservatively overall, and were less likely to respond *Old* than the other groups unless they felt a high level of certainty.

Discussion

The aim of the current study was to establish baseline PSE response patterns in a novel, modified RK paradigm. Substituting the classic *Remember* / *Know* labels for *Recollection* / *Familiarity*, recognition for words and pictures was tested across three separate response option conditions (RFG, RFBG, RF-Ratings). Analysis of the behavioural data demonstrated a clear Picture Superiority Effect (PSE) in the current paradigm, with picture stimuli showing better discrimination, a higher number of overall hits, lower number of FAs, and better overall recognition performance than words. Taken together, these findings are consistent with the notion that pictures offer an enhanced memorability in comparison to words. When word stimuli were correctly identified, they were not recognised in the the same context-rich nature as pictures, evidenced by a higher proportion of F responses. The current findings also align with those from previous studies, with pictures showing enhanced recollection (Curran & Doyle, 2011; Rajaram, 1996a) and words showing enhanced familiarity (Ally & Budson, 2007). While most of the proposed hypotheses were supported, there were some unexpected results. Stimuli format had no effect on the obtained proportions of FAs; regardless of whether FAs were assigned R or F, there was no evidence of picture superiority. This finding does not refute the notion of a PSE in the current paradigm - the memorial advantage of pictures over words is evident, but it instead indicates that stimuli without this advantage (i.e. words) may produce more misses, but not increased levels of false recognition.

Many of the unexpected results are centred around the RF-Ratings response option condition. Word stimuli were hypothesised to produce more Familiarity and Guessing hits than pictures,

since it was expected that they would not be recognised in the the same context-rich nature as pictures. This result was indeed obtained in both the RFG and RFBG response-option conditions, however, the RF-Ratings did not produce the same finding (no difference between stimuli formats was observed). Similarly, while the RFG and RFBG conditions showed comparable proportions of hits and levels of response bias (*c* scores), the RF-Ratings group again produced different findings, showing significantly fewer hits and significantly higher *c* scores (and thus a more conservative response bias) compared to the RFG group. As the proportion of hits and mean *c* scores were not significantly different between the RF-Ratings and RFBG response option groups, it indicates that these results may be attributable to participants having the ability to report that they experience *Both* recollection and familiarity processes conjointly. However, as performance differences were most notable in the RF-Ratings condition, it suggests these findings are attributable to the increased task complexity from RFG to RFBG, and RFBG to RF-Ratings. The option to report *Both* may be confusing to participants, especially to those who struggle to understand the distinction between recollection and familiarity to begin with (Geraci et al., 2009; Rubin & Umanath, 2015; Williams & Moulin, 2014). Providing the *Both* option in the form of two scales may exacerbate this confusion further, and thus lead to results that are significantly different from the condition with the least complexity (RFG). Such a hypothesis is supported by a number of other findings. First, the more conservative response bias exhibited by those in the RF-Ratings group demonstrates how subjects were less likely to respond *Old* when they were required to provide more detailed follow-up recognition judgements (i.e., using separate 0-5 scales for R and F), compared to simply selecting one of three options (R,F, or G). Second, despite *Guessing* responses being permissible in any of the response-option groups, participants were significantly less likely to report a guess when two independent ratings were required - a finding evident from the reduced number of *Guessing* hits and FAs compared to the other response option conditions. Third, the RF-Ratings group showed significantly more R hits and R FAs compared to both the RFG-group and the RFBG-groups, indicating participants were more likely to respond *Old* in the RF-Ratings condition when they experienced high certainty in their recognition - regardless of whether or not that recognition was accurate or false. Taken

together, these findings all support the notion of a certain level of avoidance from participants when they were required to provide more detailed reports of their recognition.

Establishing baseline PSEs in the current paradigm is important for allowing further experimental manipulations in the experiments that follow. While the independent ratings paradigm proposed by Higham & Vokey (2004) is undoubtedly useful in the discussion around the most effective methods of measuring recollection and familiarity, it does not suit the needs of current programme of research going forward, where comparisons between different stimuli formats is the primary concern. In the current experiment, picture stimuli consisted of simple greyscale illustrations (Rossion & Pourtois, 2004), though the extent to which stimuli of increasing levels of detail impacts recognition is unclear. The distinctiveness of to-be-remembered stimuli will be systematically compared across a number of experiments, following the conception of a new set of detailed realistic photograph stimuli. Following the unique results obtained from the RF-Ratings group in the current experiment, it is likely that this condition would exhibit further differences if included in the proposed experiments, which may become difficult to interpret when further stimuli formats are introduced. The current findings to suggest avoidant response patterns in the RF-Ratings group further highlight the unique results this condition might produce. Therefore, only the RFG and RFBG response option conditions will be taken forward into the proposed recognition experiments that focus on comparisons of stimuli distinctiveness.

######———————————————-

Chapter 3

The Picture Superiority Effect (PSE) is a highly robust and replicable phenomenon. In recognition memory paradigms, the PSE has been shown to manifest as both increased recollection and familiarity (Dewhurst & Conway, 1994; Rajaram, 1993, 1996b; Wagner, Gabrieli, & Verfaellie, 1997; Yonelinas, 2002). The effect is present in children, adolescents and healthy older adults (Whitehouse, Maybery, & Durkin, 2006), though perhaps more striking is the fact that patients with Alzheimer's disease or those presenting early isolated memory impairments, known as amnestic mild cognitive impairment (aMCI), also show memorial benefits toward pictures (Ally, 2012). This is supported by ERP studies demonstrating comparable enhancements to recollection-based ERP components between healthy older and aMCI groups when pictures, rather than words, are utilised (Ally et al., 2009a). There is debate within the literature attempting to characterise the nature of memory deficits in aMCI, whereby despite general agreement that recollection processes are impaired in such individuals, findings show great inconsistency with regard to familiarity (Algarabel et al., 2012; Belleville et al., 2011; Pitarque, 2016; Wolk, Dunfee, Dickerson, Aizenstein, & DeKosky, 2011; Wolk, Mancuso, Kliot, Arnold, & Dickerson, 2013). The PSE may have been largely overlooked as an area for further research in an effort to help settle this debate, despite recent reviews highlighting methodological differences across studies as the potential source of inconsistent findings (Koen & Yonelinas, 2014; Migo et al., 2012; Schoemaker et al., 2014). The level at which stimuli distinctiveness impacts successful recognition is currently unclear, and there is little consistency across studies with regard to what is considered a 'picture'.

Many experiments utilise illustrations for their picture stimuli (van der Meulen et al., 2012; Westerberg et al., 2013; Wolk et al., 2011), with a standardised set of items published by Snodgrass & Vanderwart (1980) among the most-used illustrated picture stimuli within the domain of memory research (Bermúdez-Margaretto, Beltrán, Cuetos, & Domínguez, 2018; Deason, Hussey, Flannery, & Ally, 2015; Hockley, 2008; Martins & Lloyd-Jones, 2006; McBride & Anne Dosher, 2002; Meade, Ahmad, & Fernandes, 2019; Schmitter-Edgecombe, Woo, & Greeley, 2009; van der Meulen et al., 2012; Wagner et al., 1997; Wammes, Meade, & Fernandes, 2016; Weldon, Iii, &

Challis, 1989; Weldon & Roediger, 1987; Whitehouse et al., 2006). The set consists of 260 line drawings of common, everyday objects (in black ink), along with their written word counterpart (e.g. “shoe”). Items were selected on the basis of exemplifying a number of semantic categories, including animals, furniture, fruit, etc., and a range of normative data was collected for each item; indices of naming agreement, mental imagery agreement, visual complexity, and familiarity were all recorded for each drawing. The normative data for the Snodgrass & Vanderwart (1980) items has been continually revisited, with a number of studies gathering culturally-appropriate norms (e.g. in Spanish (Sanfeliu & Fernandez, 1996), Chinese (Yoon et al., 2004), and Russian (Tsaparina, Bonin, & Méot, 2011), and additional testing of the relationship between reaction time and naming agreement (Székely et al., 2003). There are multiple theories of object recognition; the recognition-by-components theory proposed by Biederman (1987) identifies shape as the most crucial factor for successful recognition, in which case, the object outlines found in the set by Snodgrass & Vanderwart (1980) should be more than sufficient for experimental cognitive research. Other theories, however, posit that surface details such as colour and texture are just as crucial in forming object representations (Tanaka, Weiskopf, & Williams, 2001; Tarr & Bülthoff, 1998). The wide-ranging applicability of the Snodgrass & Vanderwart (1980) items throughout a number of cognitive disciplines has led to a more recent revision of the items by Rossion & Pourtois (2004). This revision consists of the exact same objects, digitally re-drawn to include surface textures and shading. Additionally, this set provides greyscale and colour versions for all items, as opposed to the greyscale-only items found in the Snodgrass & Vanderwart (1980) set (see Figure 6 for example items contained in the Snodgrass & Vanderwart (1980) and Rossion & Pourtois (2004) stimuli sets). The Rossion & Pourtois (2004) revision now appears to be favoured over the original Snodgrass & Vanderwart (1980) set among many cognitive researchers (Rollins & Riggins, 2018, p. @ensor2019b; Stenberg, 2006; Wolk et al., 2008), almost certainly attributable to the increased detail and ability to choose whether colour is a necessary condition.

Despite their widespread use, line drawings have been criticised for their relative simplicity and lack of realism (Viggiano, Vannucci, & Righi (2004)), with many researchers favouring the use of

photographs as experimental stimuli (Embree et al., 2012; Pitarque, 2016; Troyer et al., 2012; Troyer, Vandermorris, & Murphy, 2016; P. Wang et al., 2013). Photographs of faces are especially useful in research examining emotion and face recognition (Barba, 1997; Bowen, Fields, & Kensinger, 2019; Cui et al., 2016; Herzmann, Minor, & Curran, 2018), though a number of common-object photograph sets have also emerged as ecological alternatives to line-drawn items (Adlington, Laws, & Gale, 2009; Moreno-Martínez & Montoro, 2012; Viggiano et al., 2004). While the published sets of photographs are undoubtedly useful in a range of cognitive domains, they do not allow us to specifically examine stimuli format as a factor on its own, as the concepts depicted are unique to the set they derive from. In order to make such comparisons, and ensure any differences in performance (e.g. recognition memory ability) are indeed attributable to stimuli format, the objects depicted must be consistent across stimuli formats. The current study presents a new set of photographic stimuli that extend the set of words and drawings provided by Rossion & Pourtois (2004), wherein each of the concepts depicted has been carefully matched across formats. These new stimuli will be utilised throughout a number of planned recognition experiments that aim to systematically compare measures of recognition against different 'levels' of stimuli. The curation of a new set of photographs - carefully matched to other formats - allows investigation into whether picture superiority magnitudes are mediated by the format pictures are presented in. The inconsistent use of different formats across studies has previously made it difficult to reconcile effects obtained in response to drawings with those obtained in response to photographs - an inherent problem when concepts are not matched across format. Normative data for the new set of photographs is also presented, allowing others who also wish to use our photograph stimuli to filter items by measures of naming agreement, mental imagery agreement, familiarity, visual complexity, and colour diagnosticity.

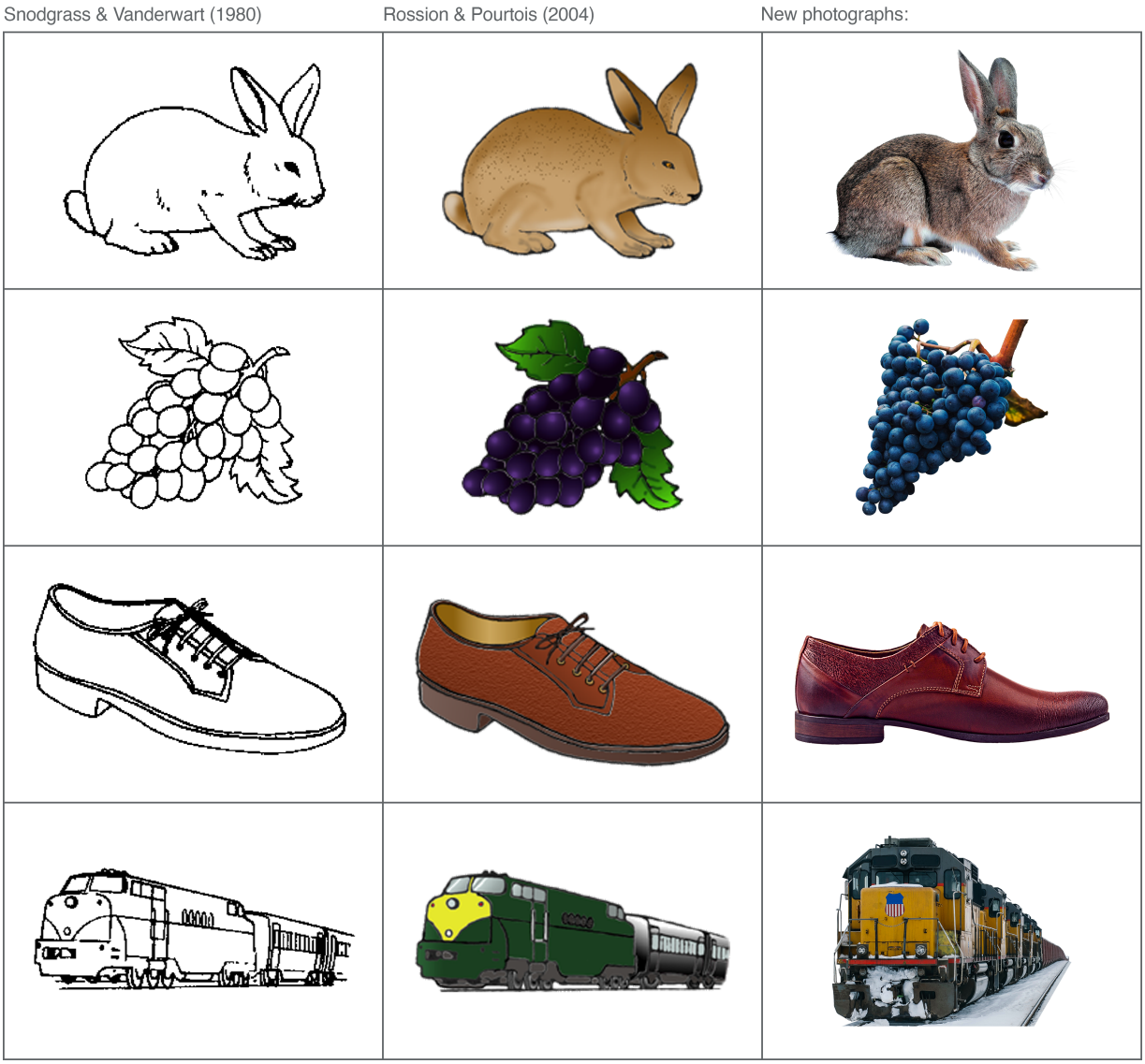


Figure 7: Examples of matching pictures across Snodgrass & Vanderwart (1980), Rossion & Pourtois (2004), and photographs from the current study. Greyscale versions of the drawings and photographs are not presented in this example.

Experiment: Development of a new set of standardised photographic stimuli

Method

Participants A total of 377 subjects completed the online experiment (see Table 3 for a breakdown of the gender and age of the sample). This sample size provided 20 data points for each of the five response types, while also ensuring the experiment did not last too long for participants (approx 25-mins). Subjects were recruited from both voluntary participation websites such as Prolific Academic (where they received payment at the rate of £5/hr), and via the in-school research participation system (where they received course participation credits).

Table 3: Gender and age (*SD*) of the current sample.

Gender	N	Age	
Female	*196*	33.22	*(11.28)*
Male	*171*	33.15	*(10.3)*
Non-binary	*2*	23.50	*(-)*
Unspecified	*5*	29.40	*(6.11)*
Total	***377***	**NA**	***NA***

To meet our YA requirements, all participants were required to be aged between 18-59 years (actual obtained range: 18-59 years). As our experiment involved typing the English labels for a range of image stimuli, subjects were also asked whether English was their first language; all but one participant indicated that English was indeed their first language (99.2%).

Materials A pool of 136 line drawings (Rossion & Pourtois, 2004) - depicting common, everyday objects - were brought forward from the previous experiment. These items (along with their written-word labels) would form two of the unique stimuli formats that would be used in future recognition experiments (words and drawings). In this study, the drawings from Rossion & Pourtois (2004) were simply used as a reference in the photograph matching process. Corresponding photographs were obtained online with the aim of depicting the everyday objects in a similar manner to the drawings. The inherent subjectivity involved in this process may have led to images that were not a reliable 'match' to the concepts they were selected to depict (for example, the photograph chosen to depict the concept "bottle" may inadvertently provoke the

majority of participants to give the label “wine”, thus indicating that this particular photograph fails to accurately depict the intended concept). To address this issue, and ensure all photographs more objectively depict the same concepts as the line drawings, three different photograph variations were found for each everyday object, with the aim of taking the best ‘match’ forward. An emphasis was placed on variety across these variations, with the aim of obtaining at least one photograph that very closely resembled the line-drawn depiction, and another offering a more modern depiction. Some items were substituted due to unique restrictions that meant they could not easily be translated into photographic format (for example, the shapes “arrow” and “star” can not be represented similarly as photographs). Photo stimuli were obtained by searching open-source, copyright-free image websites (e.g. Unsplash; Pexels) for photographs that depicted the same everyday objects as the line drawings (see Appendix B for the full list of image references).

The matching process produced a total of 408 unique photographs. All were imported into Adobe Photoshop (20.0.04 Release), where the background was removed to isolate the object of interest from other potentially distracting visual details. This was completed manually using the magnetic lasso and polygonal lasso tools (edges were either feathered by 1px or left un-feathered). The orientation of isolated objects was adjusted to ensure they matched as closely as possible with their line-drawn counterpart (e.g. all photograph variations of the item ‘boot’ were adjusted so the toe was facing left and the heel facing right, as in the line drawing); this was often achieved by flipping or mirroring the object to ‘correct’ the direction.

Despite isolating objects from their background, a small number of photographs still contained irrelevant and potentially distracting details. For example, in one photograph variation of the item ‘piano’, there was a sign on the object that may have impacted how the item was named or rated. Such details were removed as best as possible using the clone stamp and content-aware fill tools. Any obvious text (e.g. brand names) and numbers were also removed from photographs using the same method (see Figure 8). The primary aim of the current study was to obtain photographs that could be clearly distinguished as a unique stimuli format among words and line drawings; it is conceivable that combining these formats (i.e. inadvertently including photographs that also contain written words) might affect recognition performance in ways that are not directly

comparable to items defined only by a single category. Any text in our photographs was therefore removed, apart from a couple of exceptions whereby such details happened to be integral to the depiction of the object (e.g. the numbers found on a ruler or clock).

All photographs were exported from Photoshop in ".png" format in both their original colour and in greyscale (by setting saturation levels to 0). Final edits were completed in Adobe Lightroom (Classic, 8.2 Release): exposure (brightness) adjustments were made on images that appeared too light or too dark; highlights were decreased if some areas were too bright compared to the rest of the photograph; shadows were raised if some areas were too dark compared to the rest of the photograph; noise reduction was applied to some items after isolating the subject had inadvertently made unwanted noise/grain more visible. The changes made to each image were systematically applied to both the colour and greyscale versions (e.g. if one variation of "shoe" had an exposure increase of .010 for the colour version, the greyscale version also received an exposure increase of .010). Some colour-specific adjustments were made to the colour photographs only, however; common photo artefacts such as chromatic aberration (purple fringing) were corrected, along with white balance normalisation. Finally, all photographs were placed on a 600x600 pixel white background, and made to fill this frame as much as possible (i.e. some items were restrained by height, whilst others were restrained by width).

Original Manipulated

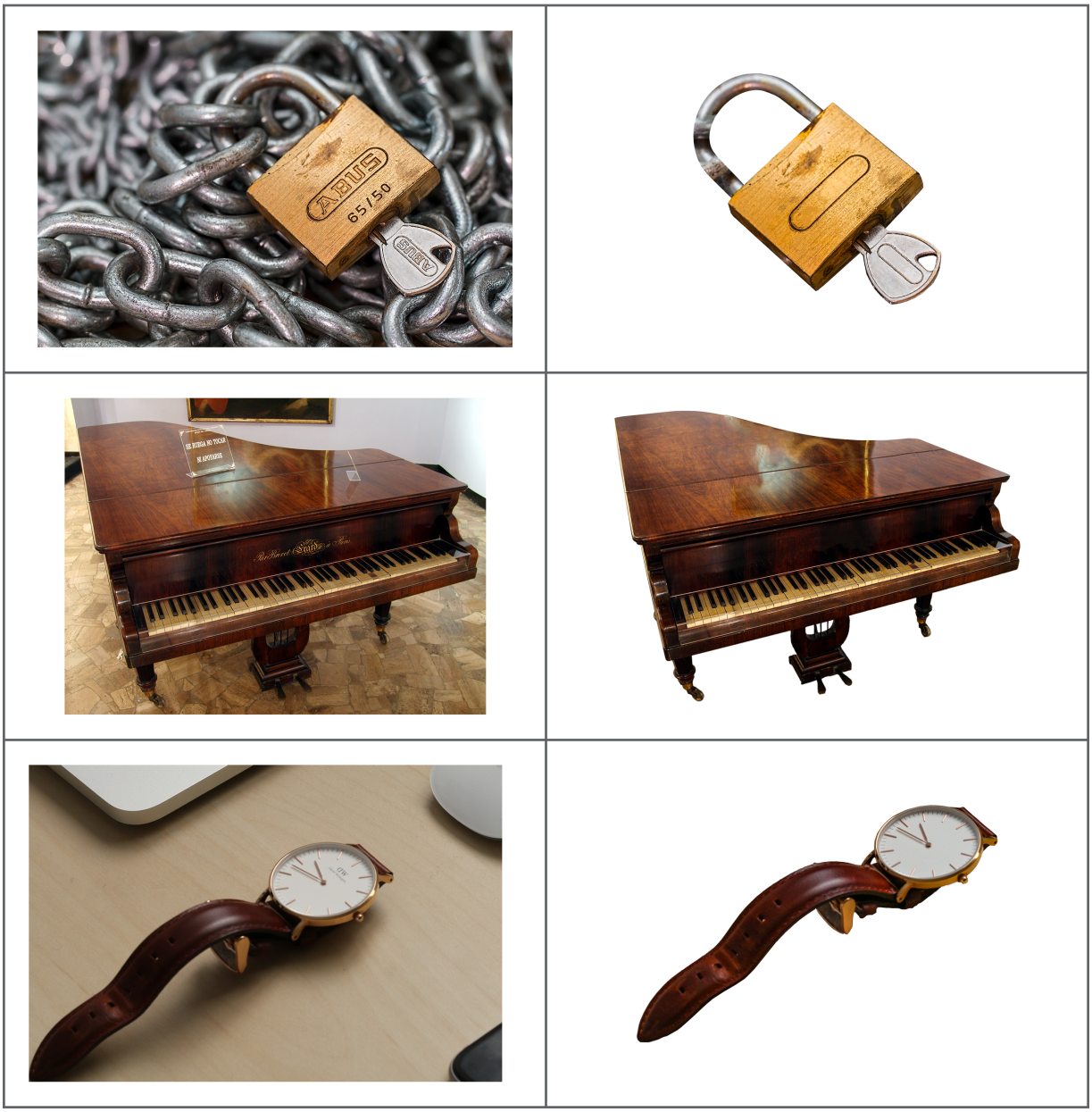


Figure 8: Examples of background and text removal in photograph items.

Design This was a descriptive study; a mix of qualitative and quantitative data were gathered. Across three blocks, all participants provided five types of response toward photograph stimuli: i) Naming; ii) Familiarity; iii) Visual Complexity, iv) Colour Diagnosticity; and v) Mental Imagery Agreement. Excluding the Naming task (consisting of a typed single-word answer), all responses

were provided on a 5-point ordinal scale. Within participants, the maximum number of response type provided for any one item was two; Naming and Familiarity responses were paired in one block, Visual Complexity and Colour Diagnosticity responses were paired in another, and Mental Imagery Agreement responses were always presented in a separate block. The order of these three blocks was counterbalanced across participants. Toward each individual photograph, participants made only one or two types of response before moving on to the next item, and the same items were not repeated to participants. For each photograph, the five types of required data were obtained by counterbalancing between participants (e.g. for the first variation of the "cat" photograph, the Naming and Familiarity data was obtained from one participant, the Visual Complexity and Colour Diagnosticity data was obtained from another, and the Mental Imagery Agreement data was obtained from another).

Procedure Data collection was conducted via two online platforms; i) Qualtrics - a survey platform that allowed for straightforward collection of consent, demographics, and computer compatibility data, and ii) Pavlovia - an open-source experiment hosting platform for studies programmed in Javascript (Peirce et al., 2019).

In the Naming and Familiarity block, participants were first asked "What is the name of the item depicted?". Subjects were instructed to name each photograph as briefly and unambiguously as possible, with one name only, and respond by typing their answer into the response box. If they did not know the name of an item, or had a tip-of-the-tongue experience, participants were instructed to type "no" for their answer (the term "don't know" was avoided so as not to encourage subjects to deviate from single-word responses, as instructed).Following the naming judgement, with the same photograph still present on-screen, participants were next asked "How familiar is the item depicted?". Subjects were instructed to judge each photo according to how usual or unusual the item was in their realm of experience; specifically, familiarity was defined as "the degree to which you come in contact with, or think about, the concept", and encouraged participants to rate the concept itself rather than the particular way it was currently shown. Participants selected one value from the 5-point scale, ranging from very unfamiliar (1) to very familiar (5),

and were encouraged to use the full range of the scale throughout the set of photographs.

In the Visual Complexity and Colour Diagnosticity block, participants were first instructed to respond to the question “How visually complex is this picture?” using a 5-point scale that ranged from “very simple” (1) to “very complex” (5). Complexity was defined to subjects as “the amount of detail in the picture”; in contrast to the familiarity ratings, participants were encouraged here to rate the complexity of the picture itself, rather than the real-life item. If the photograph shown was greyscale, subjects would simply move on to the next item. If the item shown was in colour, however, participants were also required to make a colour diagnosticity judgement. This concept was defined as “how typical / normal the colour of the item is”, instructing subjects to rate on a 5-point scale ranging from “Not at all diagnostic (i.e. this item could be in any other colour equally well)” (1) to ”Highly diagnostic (i.e. this item appears only in this colour in real life). Participants were instructed to utilise the full range of options on the scale when making visual complexity and colour diagnosticity judgements. After making these ratings, a fixation cross was presented during a 1s interstimulus interval.

Due to the slight change in procedure and increased task complexity, Mental Imagery Agreement ratings were always acquired in an individual block (i.e. not alongside any other response types). First, participants were presented with a written label for 3s (e.g. “cat”) and told to focus their attention on the word. Once the written word disappeared, a beep tone was played alongside the instruction “close your eyes and imagine this item” (subjects were encouraged to close their eyes and begin imagining the item as soon as they heard the tone, but the written instruction were included as a further prompt). After 3s a second beep tone sounded to alert subjects to open their eyes, where they were presented with a photograph of the item they had been instructed to imagine. On a 5-point scale, participants were asked to “rate the agreement between your mental image and the picture”, from “low agreement” (1) to “high agreement” (5). The degree of agreement was defined as “how similar your mental image of the item is to the picture shown”. A fixation cross was displayed for 1s before the next word item was shown.

All responses were self-paced; the timing was only controlled during the study/imagine section of the Mental Imagery Agreement block.

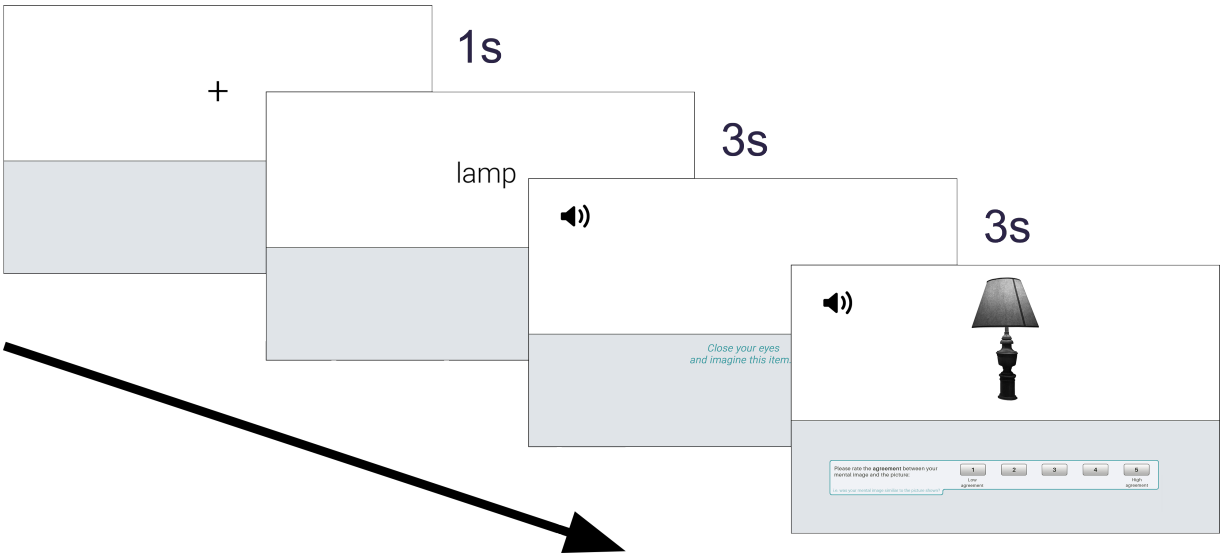


Figure 9: Data collection procedure for Mental Imagery Agreement responses.

Data processing The naming responses for each photograph item were manually assessed for spelling and typing errors. Automatic spell checking software was avoided in an effort to avoid inadvertently introducing unique names that were not actually given by participants. The vast majority of errors were unambiguous and easy to correct (e.g. “anker” = “anchor”, “peguin” = “penguin”, “ssnowman” = “snowman”), or consisted of transforming plural words to singular (or vice versa, depending on the form of the intended label - e.g. “sock” to “socks”). Some responses were a little more ambiguous, and necessitated comparison to the photographs they were in response to for additional clarity (e.g. a photograph depicting a plug that would fit into North American electrical sockets was labelled as “usplug” - given the nature of our UK-based sample, it’s likely the subject was responding: “U.S. (i.e. United States) plug”.

There were instances where subjects provided a sensible and correctly spelled English word, but that were clearly typos when examined against the photograph they were in response to (e.g. “dock” for a photograph depicting a duck, “frock” for a frog, and “beer” for a “bear”, etc). The most ambiguous spelling error to correct was “bittle”, which was provided by more than one participant and to more than one item; separate inspections of the photographs participants were responding to made this easy to correct though, with one participant clearly meaning to respond

"bottle", whilst the other meant to respond "beetle". Though participants were instructed to only give a single label for each item, some multiple word responses were found (without spaces) during the spell checking process. On such occasions, a judgement was made regarding whether multiple words were retained, or whether the response could be shortened into a single word. A general rule was applied whereby if the other words provided additional information, they were retained (e.g."maledear" - presumably "male deer" - was kept as a two-word answer). Multiple word responses were generally shortened into a single word when the intended label for the item was clearly present, and no information was lost in the process (e.g. "haircomb" was shortened to the intended answer "comb"). It is noted that there was some inherent subjectivity in this process, though as such items were not common among straightforward responses, their overall effects are estimated to be negligible.

Finally, there were some responses that were changed to "no" as they were clearly intended to signify that the responder did not know the name of the item shown; the experiment instructed participants to type "no" in these instances, though the labels "none" and "idk" (common abbreviation for "I don't know) were provided instead. There was also a single response that was manually changed to"no", as the provided label was a single letter and thus entirely unclear what the intended answer should be (see Appendix A for full list of manipulations to naming responses). This process yielded data that could be used to determine which photograph variation best matched the intended concepts (e.g. 100% of participants labelled the object "bottle", indicating a perfect match), and which did not (e.g. only 50% of participants labelled the item "bottle", whilst the other 50% gave the label "wine", indicating a poor match). Photographs showing poor agreement across participant-generated labels, or those where the majority of labels differed from the intended concept, could be replaced with the variation demonstrating the most accurate depiction.

Analysis preperation A number of variables were calculated prior to analysis. For familiarity, visual complexity, colour diagnosticity, and mental imagery agreement, mean ratings were calculated for each (see Appendix B). Mean reaction times (RTs) were also calculated for each

photograph / response variable, including naming responses. For naming responses, accuracy was defined as the proportion of subjects reporting the correct/intended label for any given item (e.g. 80% of subjects correctly labelled a photograph of the moon as "moon"). Percentage agreement was also calculated (i.e. the proportion of subjects providing the most frequent name, regardless of whether it matched the correct/intended label) in order to compute *H* values for each item. The *H* statistic also reflects naming agreement, but it takes into account the total number of unique labels given for an item. This is especially useful for comparing similar items, as it captures information not provided by simple agreement proportions. For instance, if the first variation of the photo moon ('moon-1') demonstrated 90% naming agreement among subjects, and the second variation ('moon-2') also demonstrated 90% naming agreement, it would appear as if both versions offer the same level of agreement among participants. However, 'moon-1' may have received a total of 2 unique names (e.g. moon, planet), while 'moon-2' received a total of 4 unique names (e.g. moon, planet, earth, comet). *H* values utilise this useful information to determine which item shows the best naming agreement (in other words, the item with the least number of unique names). The original formula by Snodgrass & Vanderwart (1980) was used to calculate *H* values:

$$H = \sum_{i=1}^{k} p_i \log_2 \frac{1}{p_i},$$

A *H* value of 0 indicates perfect naming agreement (all subjects responded with the same label for that item). Items showing a *H* value of 1 signify two unique names were provided, with identical proportions (e.g. 10 subjects responded "moon" and 10 subjects responded "planet"). As the *H* value increases, overall naming agreement decreases.

Results

Summary statistics (mean and *SD*) for each of the measured variables are shown in Table 4. Data for the grey and colour photographs are presented alongside previously obtained normative values for a number of other stimuli formats (all obtained from Rossion & Pourtois (2004),

who published revised norms for Snodgrass & Vanderwart (1980)'s (S&V) original line drawings, as well as their own re-drawn versions that contained shading and texture detail). The data from previous studies were not used in any statistical analyses. To examine whether the grey and colour photographs from the current study demonstrated any differences, a series of independent samples t-tests were run on each variable, as well as their corresponding reaction times (excluding scores of colour diagnosticity, which were obtained only in response to the colour items and thus cannot be compared). Mean (and *SD*) values for all x816 unique photograph items are presented in Appendix B.

Naming Naming accuracy was very high for all photographs (*M* = 0.95), indicating that overall, the selected items closely depicted the intended concepts. Compared with the other stimuli formats, there appears to be a steady increase in accuracy as items become more distinctive (see Table 4). Accuracy rates did not differ between the grey (*M* = 0.94) and colour (*M* = 0.95) versions of the photographs [$t(745.64) = -0.56, p = .576$].

H values were also low across all items (*M* = 0.23), showing that subjects generally agreed on how the items should be named. Similar to naming accuracy, naming agreement also appears to steadily increase as items become more distinctive (as indicated by decreasing H values - see Table 4. While Rossion & Pourtois (2004) observed significantly better naming agreement for their colour - rather than greyscale - items, this pattern did not reach significance with the current set of photographs; *H* values did not differ between the grey (*M* = 0.24) and colour (*M* = 0.22) photographs [$t(743.66) = 0.62, p = .537$].

A mean reaction time (RT) of (3.9s) was observed for naming responses. While this was of little interest on its own, and could not be compared to those obtained in response to the other stimuli formats as our methodology was slightly different (RTs were only recorded when subjects had typed their response *and* clicked the mouse to signify they had finished), they were useful for marking comparisons between the grey and colour items (though no difference was observed [*M* grey = 4s, *M* colour = 3.8s, $t(651.86) = 1.57, p = .117$]). Overall, these analyses suggest that the current photographs closely resemble the drawings they were designed to match, with

high levels of naming accuracy and agreement among subjects. The absence of any colour differences indicates there were no naming advantages when photographs were made even more distinctive through the addition of colour.

Table 4: Summary statistics for each of the measured variables. Mean values are presented in bold (SDs are shown in parentheses).

	Rossion & Pourtois (2004)			Current study	
	S&V lines	Grey shaded	Colour shaded	Grey photos	Colour photos
Naming accuracy	**88.2** *(17.1)*	**89.2** *(17.2)*	**90.3** *(16.9)*	**0.94** *(0.08)*	**0.95** *(0.08)*
Naming agreement (H)	**0.44** *(0.56)*	**0.38** *(0.52)*	**0.32** *(0.46)*	**0.24** *(0.33)*	**0.22** *(0.31)*
Mental imagery agreement	**3.73** *(0.48)*	**3.76** *(0.55)*	**3.74** *(0.63)*	**3.46** *(0.56)*	**3.74** *(0.65)*
Familiarity	**3.59** *(0.94)*	**3.52** *(1.01)*	**3.44** *(1.01)*	**4.13** *(0.56)*	**4.19** *(0.54)*
Visual complexity	**2.76** *(1.03)*	**2.88** *(1.03)*	**2.7** *(0.94)*	**2.87** *(0.62)*	**3.16** *(0.63)*
	-	-	-	-	-
Colour diagnosticity	**-**	**-**	**-**	**-**	**3.22**
	-	-	-	-	*(0.84)*

Mental imagery agreement Scores of mental imagery agreement were moderate across all items (*M* = 3.6). While no colour differences were previously observed between stimuli formats, the grey (*M* = 3.46) photographs in the current study showed significantly lower mental imagery agreement scores than the colour (*M* = 3.74) items [$t(800.06) = -6.54, p < .001$]. Comparisons with previous normative data also highlight how the grey photographs exhibited uniquely poorer mental imagery agreement scores than any of the other stimuli formats (see Table 4). RTs between the grey (*M* = 3.04) and colour (*M* = 2.81) items did not significantly differ [$t(571.37) = 2.14, p = .033$].

Familiarity Familiarity scores were high overall (*M* = 4.16), and like previous findings, there was no difference between the grey (*M* = 4.13) and colour (*M* = 4.19) items [$t(813.19) = -1.63, p = .103$]. However, familiarity scores for the current set of photographs were higher than those obtained for any of the other stimuli formats, and while there previously appeared to be a decline in familiarity as stimuli become more distinctive (from line drawings, to grey shaded, to colour shaded), such a pattern was not evident with the current photographs (see Table 4). RTs between the grey (*M* = 0.97) and colour (*M* = 0.98) items did not significantly differ [$t(783.66) = -0.30, p = .762$].

Visual complexity Visual complexity ratings were moderate across all of the items (*M* = 3.3). Colour (*M* = 3.16) photographs showed significantly higher scores of visual complexity than grey (*M* = 2.87) photographs [$t(813.51) = -6.65, p < .001$]. This finding is further demonstrated when compared to the scores from the other stimuli formats (see Table 4); where grey photographs show comparable levels of visual complexity, the colour photographs show higher scores than all of the other formats. There was no significant difference between the RTs of grey (*M* = 3.26) and colour (*M* = 3.35) items [$t(754.08) = -1.21, p = .228$].

Selection of final items For each concept represented in the photographs, one variation (e.g. shoe-1, shoe-2, or shoe-3) was selected for inclusion in a final list of stimuli that would be taken forward into subsequent recognition experiments. The normative naming data was assessed to establish which version best matched the existing line-drawn depictions of the concepts (Rossion & Pourtois, 2004). Naming was favoured over all of the other variables as, if an item was found to primarily convey a different concept than was intended during the naming task (e.g. if a photograph of the fruit 'orange' was labelled 'grapefruit' by the majority of subjects), then it could not be sufficiently compared to its line-drawn (and written-word) counterpart during recognition studies.

At least 20 unique naming responses were collected for each of the 816 photographs (408 grey items and 408 colour items). The proportion of 'correct' responses (i.e. names that were con-

gruent with the intended concept) and the proportion of 'don't know' responses were calculated for each item. Photographs were excluded if they:

1. received a high proportion of "don't know" responses (20%; all of the photographs depicted common, everyday objects, and so if a number of subjects were unable to name the item, that particular photograph was considered to be a poor representation of the item);
2. were incorrectly named by the majority of subjects (i.e. if the proportion of correct responses equalled ≤ 50%, since it was essential for the photographs to depict the same concepts as those found in the line drawings and word stimuli);
3. had particularly poor naming agreement (≤ 20% subjects named the object similarly). Items may not have been flagged by the second criteria (e.g. if it received 4 different names, each with a 25% ratio), but could still be considered poor representations of the intended concepts.

54 photographs were found to meet at least one of the above criteria, and therefore excluded. Regardless of whether these items were grey or colour, it was also necessary to remove its grey or colour partner (since both versions were needed to make comparisons across recognition experiments). Thus, a total of 64 items (32 grey / 32 colour) were excluded at this stage (many items already had both grey and colour versions flagged by the original criteria).

Next, the proportion of correct responses were compared between grey and colour photographs in order to identify items showing the lowest difference. In order to manipulate colour in later recognition experiments, it was important to select items where naming was congruent across colour/grey items; in other words, it would be difficult to attribute particular recognition response patterns to the addition of colour (if a difference were found) when the grey version could not be identified (or encoded) similarly. Variations exhibiting the least difference between colour and grey items (for the proportion of correct responses) were taken forward, while the rest were excluded. In a number of instances, multiple variations for the same object had the same 'difference' score. For example, all three variations of the item "balloon" exhibited perfect naming agreement, irrespective of whether they were presented in colour or grey (and thus "balloon1",

"balloon2", and "balloon3" had a difference score of 0). For items where more than 1 variation remained, manual rankings were obtained from two of the researchers to determine which variation best depicted the intended concept. For each item, the researchers independently studied the remaining variations and provided a rank of which they thought was best (1) to worst (2 or 3, depending on the number of variations that remained). The ratings from both researchers were collated; items where there was agreement as to which variation best depicted the intended concept were selected for inclusion in the final stimuli list. For all the items where there was disagreement between the researchers rankings, one of the variations was simply selected at random.

Discussion

The role of colour For naming responses (accuracy, agreement [*H*], and RTs), no differences were observed between the grey and colour photographs. Such a result was expected for accuracy and agreement scores; the addition/absence of colour should not alter how participants identify (and thus label) items, except in rare instances whereby a lack of colour may lead to the misidentification of an object (e.g. incorrectly labelling a greyscale photograph of an orange as 'grapefruit'). The data indicates, however, that this was not common, with the grey set of photographs exhibiting equally high levels of naming accuracy as the colour photographs. The absence of RT differences between the colour and greyscale sets was not expected for naming responses. It is reasonable to assume that colour photographs - with an additional layer of contextual information compared to grey items - would be identified (and therefore named) quicker than grey photographs (e.g. a colour photograph of an orange should avoid the potential ambiguity that might accompany a greyscale depiction, which could initially be confused for another type of fruit). Indeed, Rossion & Pourtois (2004) demonstrated RTs consistent with this hypotheses, with colour drawings showing significantly quicker RTs than grey items. The lack of difference in the current data could be attributable to ceiling effects, whereby all photographs were sufficiently unambiguous, and were quickly identified irrespective of whether they were presented in greyscale or colour. Examination of the other naming data, showing similarly high

levels of accuracy and agreement across grey and colour, supports this notion.

Scores of mental imagery agreement produced particularly interesting results between the grey and colour items. Grey photographs exhibited a significantly poorer match with subjects imagined presentation of the objects than the colour items. Colour differences were not observed previously between drawings (Rossion & Pourtois, 2004), and comparing the current data with that obtained in other studies (see Table 4) demonstrates how the greyscale photographs show uniquely lower mental imagery agreement scores compared with any of the other stimuli formats. To imagine the objects, it seems likely that subjects would conjure an image of how they naturally see the item in their everyday lives - which for the majority of subjects, would presumably be a colour representation. Therefore, when presented with greyscale depictions, subjects may have been more inclined to report that that item did not align quite as well as those presented in colour. However, it is unclear why a similar pattern is not also evident when comparing grey and colour drawings (Rossion & Pourtois, 2004). It may be that photographs promote stricter internal criteria when subjects must decide whether an item is a good match to their mental image. With line-drawn / illustrated items, subjects may simply accept that the items are baseline depictions, and that they will only able to match their real-world mental images to a certain degree - thus leading to a generally more liberal response bias throughout. The addition of colour may therefore do very little to further reconcile the match between the drawing and real-world mental representation. When subjects are responding only to photographs, the ecological nature of the items may facilitate deeper critical evaluation of whether they offer a good match to mental images, and thus promote a more conservative response bias. Colour may therefore be a far more important factor in photographs than it is in line drawings for allowing participants to decide whether an item matches well with their mental image.

There were no colour differences in familiarity scores. This result was expected - participants were asked to rate the degree to which they came in contact with, or think about, the concept itself rather than the particular depiction shown, and there is no apparent reason why colour should influence such ratings. Visual complexity, on the other hand, where participants were required to directly rate the amount of detail in the picture, did show an expected difference. Colour

photographs were rated as significantly more visually complex than grey items, presumably due to their additional layer of contextual information. When compared to the previous data obtained for drawings, the greyscale photographs showed comparable levels of visual complexity, while the colour photographs showed higher levels than any of the other formats. It is unclear why the photographs of the current study showed colour differences, when grey and colour drawings did not differ, though it may tie in with the hypotheses proposed to explain the mental imagery agreement data. Subjects may apply stricter internal criteria when rating stimuli that are perceived as being closer to how they would be experienced in real life - when viewing a colour photograph of a rabbit, it is difficult to see how we could make the item any more visually complex than it already is (at least in a 2D medium). It's probable that subjects notice the absence of colour when viewing the greyscale items, since they depict the items in a way that they are not usually seen, and thus determine that these items could be made more complex if they were shown in colour (and so give lower visual complexity ratings as a result).

Establishing a new set of stimuli The objective of the current study was to establish a new set of ecological photograph stimuli to be taken forward into subsequent recognition memory experiments. Matching items with previously established drawings (and words) would allow for the effects of stimuli-format on recognition response patterns to be directly examined. A range of normative data was collected for 816 unique photograph items. These items may prove useful for a range of cognitive researchers that wish to utilise a set of high quality and realistic object stimuli, especially given the flexibility of items that can be filtered based on colour, naming agreement, familiarity, etc. For the needs of the current body of research, the naming data was used to determine which photographs best matched the intended concepts among a number of possible variations. This allowed for the systematic comparison of recognition memory performance toward three distinct stimuli formats (words, drawings, and photographs) in the following study, in an effort to establish how stimuli of varying perceptual distinctiveness may affect recognition response patterns. Such comparisons might help to reconcile the inconsistencies present across recognition memory research, such as those attempted to determine whether familiarity

processes are preserved in those with amnestic Mild Cognitive Impairment (aMCI).

Experiment: Effect of stimuli format (greyscale) and response option on recognition memory judgements.

For the recognition memory experiment, everyday objects were presented in three stimuli formats: i) words (written in simple, black ink); ii) drawings (shaded line-drawn illustrations); and iii) photographs (detail rich exemplars of the real world object). Rossion & Pourtois (2004) demonstrated that naming agreement could be improved by adding surface texture and shading to the original Snodgrass & Vanderwart (1980) items; however, it is unclear how manipulations to distinctiveness actually impact performance in recognition memory paradigms. As well as general inconsistencies regarding the type of stimuli used in recognition memory experiments, there is also much variability in the response options available to participants when reporting their recognition, for example: Remember/Know (Lombardi et al. (2016)), Recollection/Familiarity (???), or Low/Med/High confidence (???). In the current experiment, the availability of different response options when reporting recognition will also be examined by randomly assigning participants into a paradigm with three response options (Recollection / Familiarity / Guessing) or four response options (RFG + Both).

Based on the results of Experiment 1, which compared recognition to for words and drawings only, a number of hypotheses are proposed as to the potential effects of adding a third stimuli format (highly distinctive photograph stimuli). As stimuli become increasingly distinctive (from words, to drawings, to photographs), it seems likely that the number of hits (correctly recognised items) will increase, and the number of false alarms (FAs) will decrease. RFG responses are expected to show a similar pattern, with the most detailed stimuli showing the highest number of hits assigned “Recollection”, while the less detailed formats show increasing levels of “Familiarity” and “Guessing” hits. Whilst we expect the overall number of FAs to increase as stimuli become less distinctive (i.e. words will show the highest rate of FAs), there is no reason to believe that these FAs will be biased toward any particular RFG judgement across formats. It is also hypothesised that the rates of reported Recollection and Familiarity will differ across response

option conditions (RFG / RFBG), though the direction of this difference is currently unclear.

Method

Participants A total of 169 subjects completed the online experiment (see Table 5 for a breakdown of the gender and age of the sample). To meet our YA requirements, all participants were required to be between 18-59 years of age (actual range: 18-58). As our experiment involved English word stimuli, we also asked subjects whether English was their first language; the vast majority (95.86%) reported that English was indeed their first language. Subjects were recruited from voluntary participation websites such as Prolific Academic (73.37%), where payment at the rate of £5/hr was given, and via the in-school research participation system (15.38%), where they received course participation credits. A small number of participants were also recruited from Psychological Research on the Net (11.24%). In order to detect a medium effect size of Cohen's f = 0.25 with 80% power (α = .05, two-tailed), G*Power indicated that we would need 79 participants per group (*N* = 158) in a 3x2 mixed ANOVA.

Table 5: Gender and age (*SD*) of the current sample.

Gender	N	Age	
Female	*102*	29.64	*(10.22)*
Male	*63*	30.98	*(10.97)*
Questioning	*1*	21.00	*(0)*
Unspecified	*3*	50.33	*(4.93)*
Total	***169***	**30.46**	***(10.75)***

Materials A total of 126 innocuous, everyday objects (e.g. clock, rabbit, shoe) were presented across three individual stimuli formats: written words, line drawings, and photographs. The line drawings were obtained from Rossion & Pourtois (2004), and consisted of greyscale shaded illustrations that contained some surface details. The word stimuli were simply the written word names of the line-drawn objects, presented in a clear Sans-serif typeface. The photograph stimuli were curated in the previous study; high quality photographs were sourced to similarly depict the same everyday objects as the line drawings. All objects in the photographs were isolated

from their original background, converted to greyscale, and rotated to match the orientations shown in the line-drawn items.

Design The current study utilised a mixed design, with a 3-level within-subjects factor of stimuli format (words, drawings, photographs), and a 2-level between-subjects factor of response option (RFG, RFBG). Subjects passed through 2 levels of blocked randomization (equally sized, predetermined blocks); first, participants were randomly assigned one of six possible study lists (of equal length, and containing an even number of word, drawing, and photograph items) for counterbalancing purposes. Subjects were then either assigned into a recognition test with three possible response options (RFG: “Recollection”, “Familiarity”,“Guessing”), or four possible response options (RFBG: “Recollection”, “Familiarity”, “Guessing”, “Both”). These randomisation processes were completed automatically by the experiment software using balanced methods.

Words:	Drawings:	Photographs:
guitar		
mouse		
pumpkin		

Figure 10: Examples of the word, greyscale drawing, and greyscale photograph stimuli utilised in the current experiment.

Procedure Data was collected online using Gorilla - a platform for the building and hosting of online experiments. The experiment consisted of three self-paced phases: i) study phase, ii) distractor task, and iii) recognition test. In the study phase, an even mix of word, drawing, and photograph stimuli were presented one-at-a-time on the computer screen. Subjects were instructed to learn the items in preparation for a later memory test. To ensure attention was directed to the presented stimuli, subjects were required to report whether each item was shown as a word, drawing, or photograph using the computer mouse. Following the study phase, participants completed some simple multiple choice mathematical questions (e.g. 6 x 4 = ?) as a distractor. Finally, participants memory of the previously studied items was testing in the recognition task. An even mix of word, drawing, and photograph stimuli were again presented one-at-a-time on the screen; half of the test items had been shown previously in the study phase, while the other half were new (and were not on the study list). For each item, subjects were instructed to press *Old* if they believed it was an item they had studied earlier, and *New* if they had not. *Old* responses led to a follow-up judgement, where participants reported whether they had experienced recognition through “Recollection”, “Familiarity”, or were simply taking an uninformed “Guess”. Participants that had been randomised into the RFBG test condition had a fourth option here, whereby they could report that they had experienced Recollection and Familiarity simultaneously (“Both”). Stimuli format was congruent across the study and test blocks (e.g. items presented as photos at study were also presented as photos at test). For each concept depicted across the three stimuli formats, subjects were only presented with one variation (in other words, if a subject saw a photograph for the item “shoe”, they did not see the word or line-drawn version of “shoe”).

Data processing Measured variables included the total number of hits and FAs, and the total number of hits and FAs assigned to each of the available response options (R/F/G and R/F/B/G).

In order to create a common dependant variable, proportions were calculated from these variables slightly differently depending on the response option group. In the RFG-judgement group, simple proportions were created from the total number of R responses and the total number of F responses. In the RFBG condition, however, the proportion of Both responses was separately added to R proportions and F proportions. Additional DVs included: i) d' (d-prime, a signal detection measure of sensitivity); ii) c-value (a measure of response bias); iii) overall accuracy (hits / (hits + FAs)); iv) reaction times for all responses.

Participants were excluded from analysis if they showed poor performance during the encoding task; the relative ease of reporting whether each item was shown as a word, drawing, or photograph prompted a performance cut off of 90% accuracy. This allowed for some accidental clicks / incorrect responses toward potentially ambiguous items, though subjects scoring less than 90% were excluded on the assumption they did not dedicate their full attention to the task. Subjects with extreme z-scores were also excluded from analysis; those presenting z-scores of +/- 3 (for total hits, total FAs, or overall recognition [hits minus FAs]) were considered outliers. These criteria resulted in the exclusion of 8 datasets, leaving a total of 161 in analyses.

Results

A series of 3x2 mixed ANOVAs were conducted on each of the DVs, using a within-subjects factor of stimuli format (words / grey drawings / grey photos) and a between-subjects factor of response option (RFG / RFBG); the Greenhouse-Geisser correction was applied when data was found to violate the assumption of sphericity (assessed using Mauchlys test statistic). Significant main effects of stimuli format (and interaction effects) were assessed using Bonferroni-adjusted pairwise comparisons. When no interaction effects were present, significant main effects of response option were assessed via standard two sample t-tests (when group variances were equal) or Welch two sample t-test (when variances were not equal); variance was assessed using Levene's test.

Stimuli distinctiveness Table 6: Mean proportion of hits, FAs, and mean *d'* scores, by stimuli format and response option condition.

	Hits	FAs	d'
Stimuli format			
Words	0.54	0.21	1.15
Grey drawings	0.76	0.09	2.38
Grey photographs	0.85	0.05	3.08
Response option			
RFG	0.74	0.13	2.25
RFBG	0.69	0.11	2.16

The mean proportion of hits and FAs, and mean *d'* scores are presented in Table 6. ANOVA results demonstrated a significant main effect of stimuli format for the mean proportion of hits, $F(1.76, 280.25) = 225.67$, $p < .001$, $\eta_p^2 = .59$. Paired samples t-tests showed that grey photographs ($M = 0.85$) produced a significantly higher proportion of hits than both words ($M = 0.54$), $t(160) = 18.56$, $p < .001$; $d = 1.46$, 95% CI [1.27, 1.71], and grey drawings ($M = 0.76$), $t(160) = -8.04$, $p < .001$; $d = -0.63$, 95% CI [-0.83, -0.46]. A significantly higher proportion of hits was also evident for grey drawings ($M = 0.76$) compared to words ($M = 0.54$), $t(160) = 13.6$, $p < .001$; $d = 1.07$, 95% CI [0.87, 1.3]). There were no significant interaction effects between stimuli format and response option condition, $F(1.76, 280.25) = 0.58$, $p = .540$, $\eta_p^2 < .01$.

The ANOVA on the mean proportion of FAs also demonstrated a significant main effect of stimuli format $F(1.43, 226.74) = 90.19$, $p < .001$, $\eta_p^2 = .36$. Grey photographs ($M = 0.05$) produced significantly fewer FAs in comparison to both words ($M = 0.21$), $t(160) = -11.84$, $p < .001$; $d = -0.93$, 95% CI [-1.07, -0.79], and grey drawings ($M = 0.09$), $t(160) = 5.59$, $p < .001$; $d = 0.44$, 95% CI [0.31, 0.58]). The grey drawings ($M = 0.09$) also showed a significantly lower proportion of FAs compared to words ($M = 0.21$), $t(160) = -7.98$, $p < .001$; $d = -0.63$, 95% CI [-0.81, -0.46]. There were no significant interaction effects between stimuli format and response option condition, $F(1.43, 226.74) = 1.17$, $p = .299$, $\eta_p^2 < .01$.

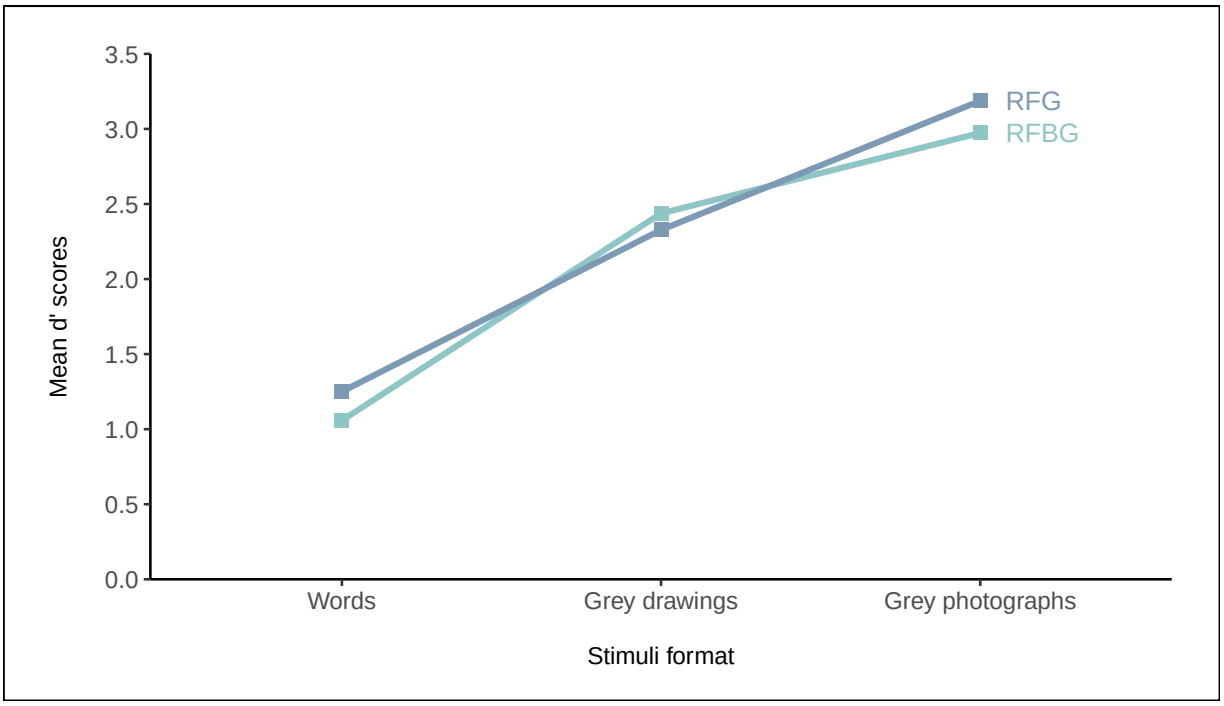


Figure 11: Interaction plot between stimuli format and response option for d' scores.

Results from the ANOVA on mean *d'* scores showed a significant interaction between stimuli format and response option condition, $F(2, 318) = 3.34$, $p = .037$, $\eta_p^2 = .02$ (see Figure 11). While *d'* scores were numerically higher for words and grey photographs in the RFG group (words $M = 1.25$; grey photographs $M = 3.19$) compared to the RFBG group (words $M = 1.06$; grey photographs $M = 2.97$), neither were significantly different from one another (words: $t(320.69) = -1.38$, $p > .999$; grey photographs: $t(320.69) = -1.54$, $p > .999$. For grey drawings, however, this pattern was reversed; *d'* scores were numerically higher in the RFBG condition ($M = 2.44$) rather than the RFG condition ($M = 2.33$); though again, these means were not significantly different from one another (grey drawings: $t(320.69) = 0.79$, $p > .999$).

Recollection and Familiarity To determine the effects of stimuli format and response option on the classification of recognition memory judgements, separate 3 (stimuli format: words, drawings, photographs) x 2 (response option condition: RFG-judgements, RFBG-judgements) mixed ANOVAs were conducted on the mean proportion of hits assigned Recollection, Familiarity, and Guessing (see Figure 12).

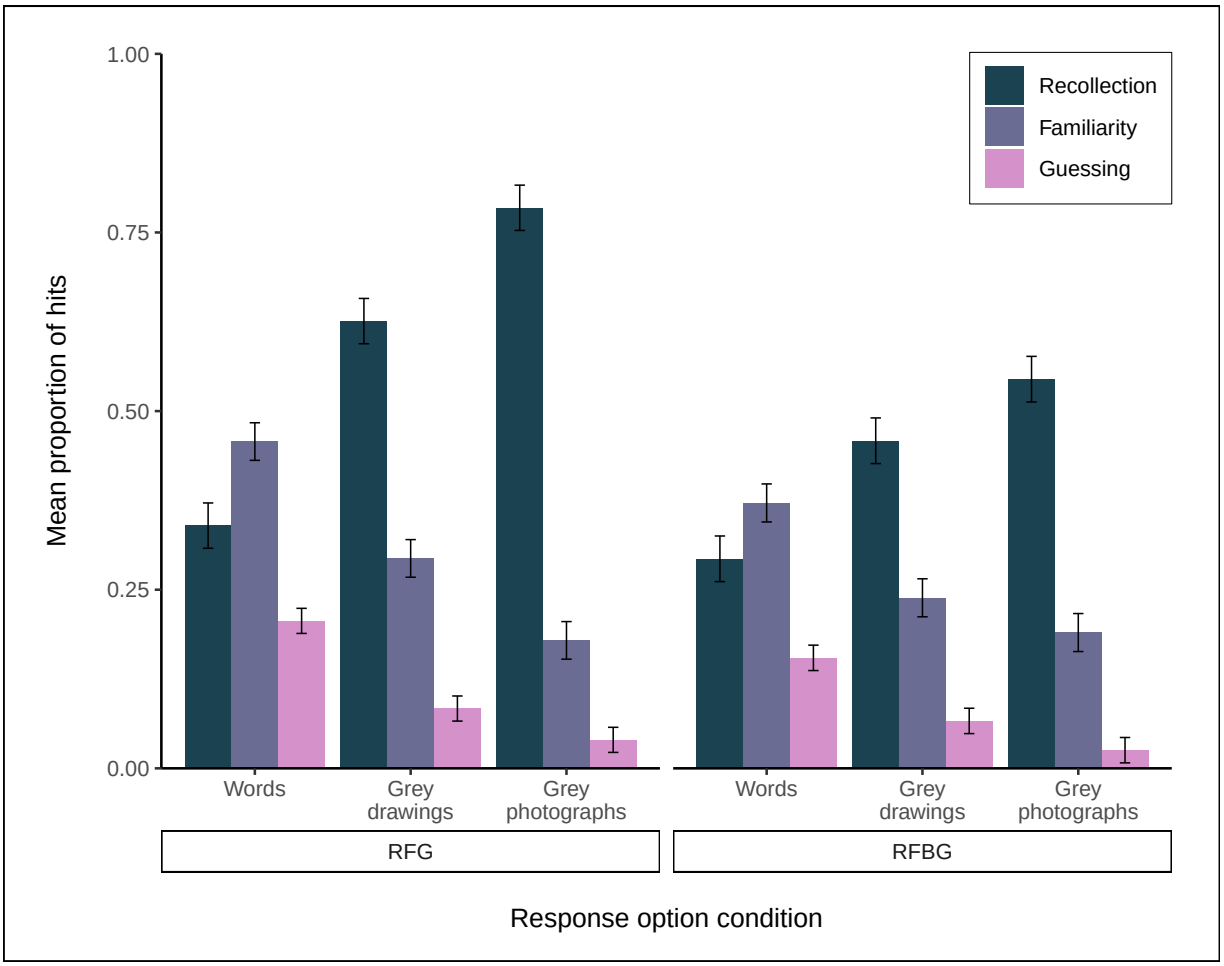


Figure 12: Proportion of hits assigned Recollection, Familiarity, and Guessing, by stimuli format and response option condition.

Recollection (hits)

Results from the ANOVA on the mean proportion of hits assigned Recollection showed a significant interaction between stimuli format and response option condition, $F(1.74, 276.60) = 12.67$, $p < .001$, $\eta_p^2 = .07$ (see Figure 13).

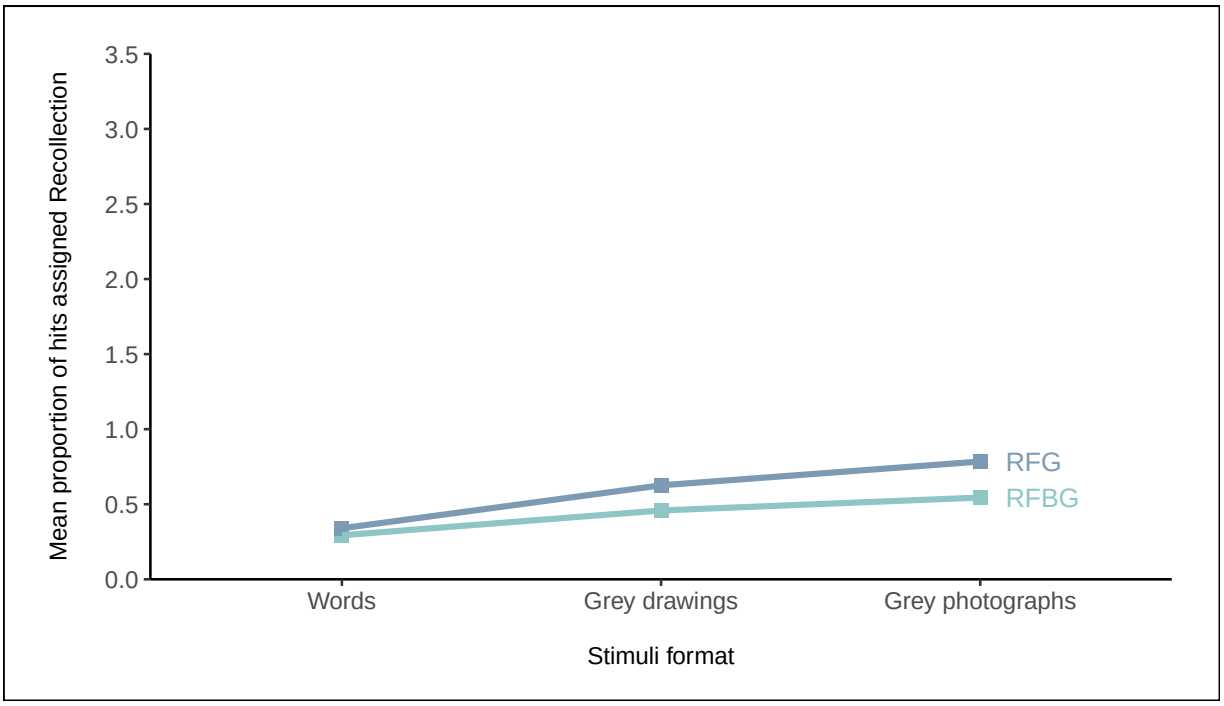


Figure 13: Interaction plot between stimuli format and response option for the mean proportion of hits assigned Recollection.

Comparisons across stimuli formats showed an expected pattern. Grey photographs produced a significantly higher proportion of R hits than words and grey drawings in both the RFG group (grey photographs [*M*= 0.78] vs. words [*M*=], $t(318) = -16.46$, $p < .001$; grey photographs [*M*= 0.78] vs. grey drawings [*M*= 0.63], $t(318) = -5.87$, $p < .001$) and the RFBG group (grey photographs [*M*= 0.54] vs. words [*M*=], $t(318) = -9.02$, $p < .001$; grey photographs [*M*= 0.54] vs. grey drawings [*M*= 0.46], $t(318) = -3.09$, $p = .032$). Likewise, grey drawings produced a significantly higher proportion of R hits in comparison to words in both the RFG (grey drawings [*M*= 0.63] vs. words [*M*=], $t(318) = -10.59$, $p < .001$) and RFBG conditions (grey drawings [*M*= 0.46] vs. words [*M*=], $t(318) = -5.93$, $p < .001$).

The interaction is evident following comparisons of the same stimuli format across response option conditions. The RFG group produced a significantly higher proportion of R hits than the RFBG group for grey photographs (RFG [$M = 0.78$] vs. RFBG [$M = 0.54$], $t(266.67) = -5.33$, $p < .001$) and for grey drawings (RFG [$M = 0.63$] vs. RFBG [$M = 0.46$], $t(266.67) = -3.72$, $p = .004$). However, this was not the case for words, where there was no difference in the proportion of R

hits between the RFG (*M* =) and RFBG groups (*M* = ; *t*(266.67) = -1.03, *p* > .999).

Familiarity (hits)

Results from the ANOVA on the mean proportion of hits assigned Familiarity showed a significant interaction between stimuli format and response option condition, *F*(1.61, 256.13) = 3.52, *p* = .041, η_p^2 = .02 (see Figure 14).

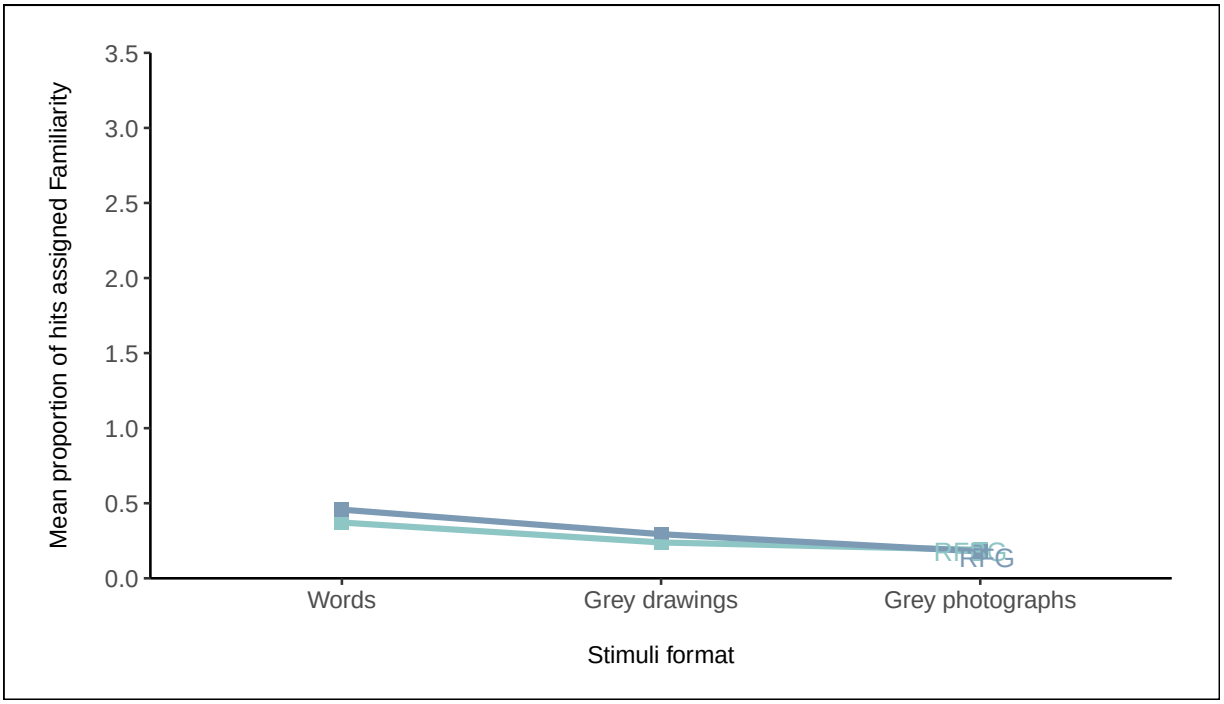


Figure 14: Interaction plot between stimuli format and response option for the mean proportion of hits assigned Familiarity

In the RFG group, grey photographs (*M*=) produced a significantly lower proportion of F hits than both words (*M*= ; *t*(318) = 10.72, *p* < .001) and grey drawings (*M*= ; *t*(318) = 4.42, *p* < .001). Grey drawings (*M*=) in the RFG group similarly produced significantly fewer F hits compared to words (*M*= ; *t*(318) = 6.30, *p* < .001). In the RFBG group, however, while grey photographs (*M*=) again produced a significantly lower proportion of F hits compared to words (*M*= ; *t*(318) = 6.77, *p* < .001), the difference in comparison to grey drawings (*M*=) was no longer evident, *t*(318) = 1.82, *p* > .999). Grey drawings (*M*=) in the RFBG group did continue to produce significantly fewer F hits compared to words (*M*= ; *t*(318) = 4.96, *p* < .001).

Response option condition had no effect on the proportion of F hits obtained, for either grey photographs (RFG [*M* =] vs. RFBG [*M* =], *t*(316.57) = 0.29, *p* > .999), grey drawings (RFG [*M* =] vs. RFBG [*M* =], *t*(316.57) = -1.47, *p* > .999), or words (RFG [*M* =] vs. RFBG (*M* = , *t*(316.57) = -2.30, *p* = .336).

Guessing (hits)

The ANOVA on the mean proportion of hits assigned Guessing demonstrated a significant main effect of stimuli format *F*(1.33, 211.61) = 69.27, *p* < .001, η_p^2 = .30. Grey photographs (*M*= 0.03) produced significantly fewer G hits in comparison to both words (*M*= 0.18; *t*(160) = -9.35, *p* < .001; *d* = -0.74, 95% CI [-0.87, -0.63]) and grey drawings (*M*= 0.08; *t*(160) = 5.85, *p* < .001; *d* = 0.46, 95% CI [0.34, 0.59]). The grey drawings (*M*= 0.08) also showed a significantly lower proportion of G hits compared to words (*M*= 0.18; *t*(160) = -7.54, *p* < .001; *d* = -0.59, 95% CI [-0.72, -0.5]. There were no significant interaction effects between stimuli format and response option condition *F*(1.33, 211.61) = 1.28, *p* = .269, η_p^2 < .01.

To determine the effects of stimuli format and response option on the classification of recognition memory judgements, separate 3 (stimuli format: words, drawings, photographs) x 2 (response option condition: RFG-judgements, RFBG-judgements) mixed ANOVAs were conducted on the mean proportion of FAs assigned Recollection, Familiarity, and Guessing (see Figure 15).

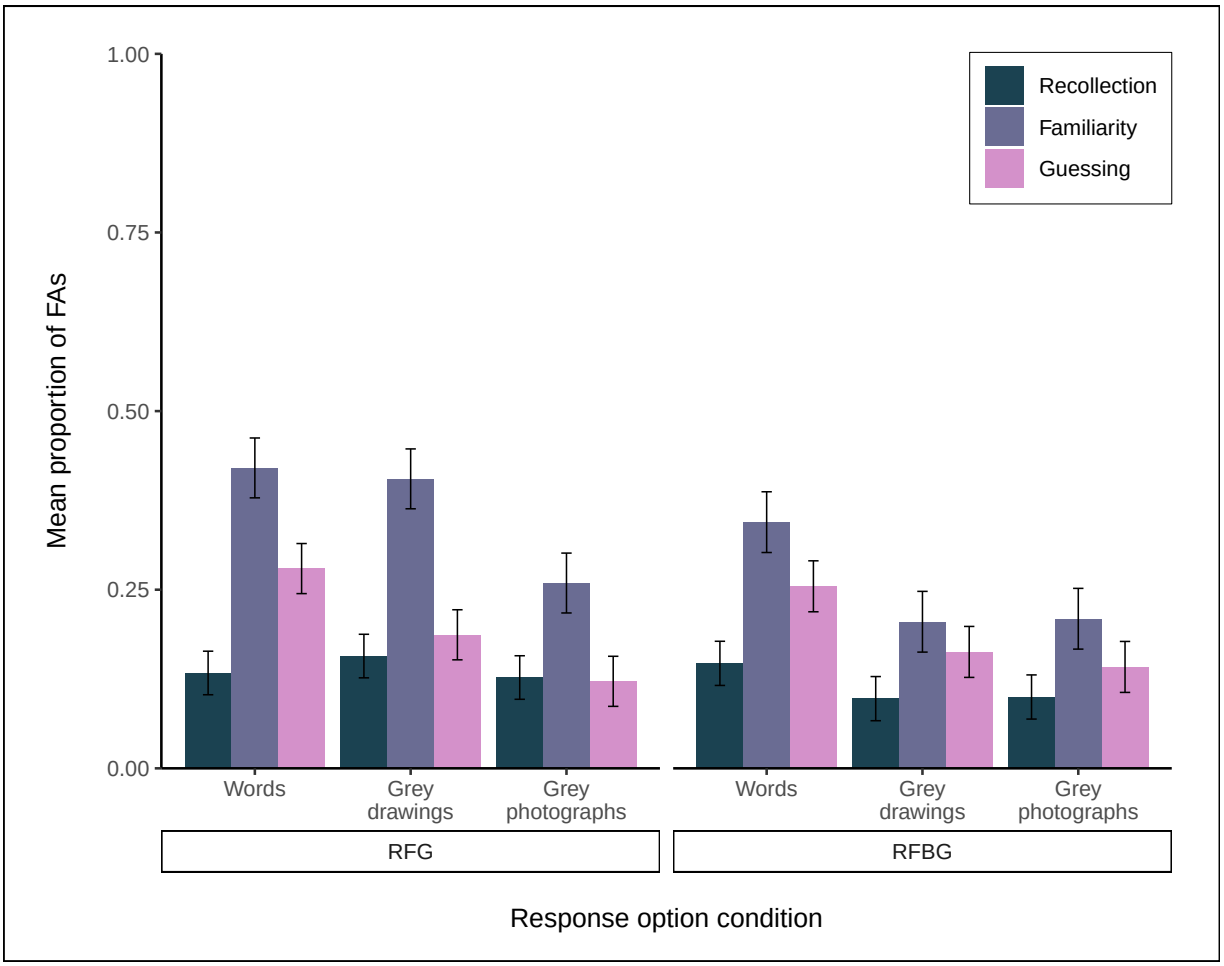


Figure 15: Proportion of FAs assigned Recollection, Familiarity, and Guessing, by stimuli format and response option condition.

Recollection (FAs) For FAs assigned *Recollection*, there was no significant main effect of stimuli format [$F(1.92, 304.64) = 0.56$, $p = .567$, $\eta_p^2 < .01$ or interaction [$F(1.92, 304.64) = 1.04$, $p = .352$, $\eta_p^2 < .01$].

Familiarity (FAs)

The ANOVA for FAs assigned *Familiarity* showed a significant main effect of stimuli format, $F(2, 318) = 8.66$, $p < .001$, $\eta_p^2 = .05$. Grey photographs ($M = 0.23$) produced significantly fewer F FAs than words ($M = 0.38$), $t(160) = -4.26$, $p < .001$; $d = -0.34$, 95% CI [-0.5, -0.19]. Likewise, grey drawings ($M = 0.31$) also showed a significantly lower proportion of FAs compared to words ($M = 0.38$), $t(160) = -1.97$, $p = 0.15$; $d = -0.16$, 95% CI [-0.32, -0.0026]. However, there was no sig-

nificant difference in the proportion of FAs assigned Familiarity between grey photographs (*M*= 0.23) and grey drawings (*M*= 0.31), *t*(160) = 2.15, *p* = 0.1; *d* = 0.17, 95% CI [0.02, 0.33]. There were no significant interaction effects between stimuli format and response option condition, *F*(2, 318) = 2.53, *p* = .081, η_p^2 = .02.

Guessing (FAs)

The ANOVA on the mean proportion of FAs assigned *Guessing* demonstrated a significant main effect of stimuli format *F*(1.92, 305.67) = 8.95, *p* < .001, η_p^2 = .05. Grey photographs (*M*= 0.13) produced significantly fewer G FAs in comparison to words (*M*= 0.27; *t*(160) = -3.98, *p* < .001; *d* = -0.31, 95% CI [-0.48, -0.16]). Likewise, grey drawings (*M*= 0.17) also showed a significantly lower proportion of FAs compared to words (*M*= 0.27; *t*(160) = -2.7, *p* = 0.02; *d* = -0.21, 95% CI [-0.37, -0.05]. However, there was no significant difference in the proportion of FAs assigned Guessing between grey photographs (*M*= 0.13) and grey drawings (*M*= 0.17; *t*(160) = 1.5, *p* = 0.41; *d* = 0.12, 95% CI [-0.04, 0.27]). There were also no significant interaction effects between stimuli format and response option condition *F*(1.92, 305.67) = 0.31, *p* = .725, η_p^2 < .01.

Response option availability In the previously discussed ANOVAs, significant main effects were also observed for response option condition (RFG / RFBG) for the mean proportion of hits (*F*(1, 159) = 4.04, *p* = .046, η_p^2 = .02) and the mean proportion of FAs assigned *Familiarity* (*F*(1, 159) = 6.30, *p* = .013, η_p^2 = .04). For all other variables, response option was found to either significantly interact with stimuli format (discussed previously), or had no significant main effect (see Table 7 for all response option ANOVA results).

For the mean proportion of hits, follow up t-tests showed a higher proportion of hits in the RFG group (*M*= 0.74) compared to the RFBG group (*M*= 0.69), *t*(481) = -2.37, *p* = .018, *d* = 0.22. For the mean proportion of FAs assigned *Familiarity*, those in the RFG group (*M*= 0.36) showed a significantly higher proportion than those in the RFBG group (*M*= 0.25), *t*(480.99) = -3.12, *p* = .002, *d* = 0.28.

Table 7: Main effects of response option condition across all variables of interest. Signif.

codes: ***$p < .001$; **$p < .01$; *$p < .05$; + involved in significant interaction [see previous section for interpretation]).

Variable	Main effect of response option	Signif.
Mean proportion: ***Hits***	$F(1, 159) = 4.04$, $p = .046$, $\eta_p^2 = .02$	*
Mean proportion: ***FAs***	$F(1, 159) = 1.03$, $p = .312$, $\eta_p^2 < .01$	
Mean scores: ***d'***	$F(1, 159) = 0.76$, $p = .385$, $\eta_p^2 < .01$	+
Mean proportion: ***Recollection hits***	$F(1, 159) = 15.02$, $p < .001$, $\eta_p^2 = .09$	+
Mean proportion: ***Familiarity hits***	$F(1, 159) = 2.01$, $p = .159$, $\eta_p^2 = .01$	+
Mean proportion: ***Guessing hits***	$F(1, 159) = 1.93$, $p = .166$, $\eta_p^2 = .01$	
Mean proportion: ***Recollection FAs***	$F(1, 159) = 0.58$, $p = .446$, $\eta_p^2 < .01$	
Mean proportion: ***Familiarity hits***	$F(1, 159) = 6.30$, $p = .013$, $\eta_p^2 = .04$	*
Mean proportion: ***Guessing hits***	$F(1, 159) = 0.08$, $p = .772$, $\eta_p^2 < .01$	

Discussion

Across a range of performance variables, the results show a clear effect of stimuli distinctiveness. As distinctiveness increased (from words, to drawings, to photographs), this produced more hits, less FAs, better overall recognition, and better discrimination between hits / FAs. The absence of any interaction effects across these variables demonstrates that the availability of different response options (i.e. the addition of a Both option) had little impact on overall performance. RF(B)G responses for accurate recognition displayed a similar pattern; as distinctiveness increased, the number of Recollected hits also increased, while the number of Familiarity and Guessing hits decreased. The rate of both Familiarity FAs and Guessing FAs was also highest for the least distinctive stimuli (words).

######———————————————-

Chapter 4 - The role of colour

Background

In *Experiment 3*, this was demonstrated by comparing recognition toward stimuli of increasing levels of distinctiveness: words, line-drawn illustrated pictures, and detailed real-world photographs.

Another inconsistency across recognition studies relates to whether picture stimuli are presented in greyscale or colour. The extent to which colour alters the distinctiveness of stimuli is unclear, though there is evidence to suggest it may play an important role in the memorability of stimuli. In a two-alternative forced choice recognition paradigm, Suzuki & Takahashi (1997) demonstrated that black & white images may not facilitate successful recognition in the same way as colour images. In an initial study phase, subjects were instructed to passively memorise a number of picture stimuli, consisting of detailed real-world photographs of scenes (e.g. train stations, city streets etc.). At test, participants were presented with two images side-by-side (one target + one similar lure) and asked to select the item shown during the study phase. Manipulations were also made with regard to the congruency; photos were either presented in the same colour modality across study and test (1. b&w + b&w; 2. colour + colour) or different (3. b&w + colour; 4. colour + b&w). Participants were informed that old photos may be presented in a different colour format (e.g. a colour item shown at study might not be shown in b&w), and were also asked to make a source judgement about whether the colour format for each item was the same or different. Results showed that, of the four possible congruency conditions, recognition performance was best when items were congruently shown in colour at both study and test. Such findings indicate colour information provided a recognition benefit at encoding and retrieval, though interestingly, this benefit was *only* evident in the congruent colour condition; if colour images were shown only at study, or only at test, there were no recognition benefits. Performance on the source judgement question - regarding whether the the colour of the item had changed since the study phase - was similar across all four conditions, but performance was particularly poor for

items that had been presented in colour. This indicates recognition benefits were not a result of accessing memory for the colour information itself; Suzuki & Takahashi (1997) instead hypothesised that colour indirectly highlighted certain features within the photographs that were not otherwise noticed as prominently in b&w, and as a result, the colour photographs were overall more distinctive.

(colour = no effect) Stróżak

Indeed, colour images have previously been linked to enhanced memory performance in recognition tasks (Ally et al., 2009a), though it was difficult to disentangle the specific effects of colour from the complexity of detail featured.

The current experiment aims to establish whether colour information facilitates recognition

irrespective of the level of detail present in an image.

The methodology of *Experiment 3* will be repeated utilising colour (rather than greyscale) versions of both types of picture: i) a set of simple, line-drawn object illustrations sourced from Rossion & Pourtois (2004); ii) a set of detailed real-world object photographs established in *Experiment 2*.

A secondary aim is to determine whether

these response patterns are differentially affected by the availability of different response options at test (*Recollection*/*Familiarity*/*Guessing* or *Recollection*/*Familiarity*/*Both*/*Guessing*).

It is hypothesised that:

1. A photograph superiority effect (similar to that observed in *Experiment 3*) will again be evident, whereby photograph items produce better recognition performance compared to drawings. Based on the results of the previous study, this performance benefit is expected to manifest as:

    - i) a higher proportion of correct hits;

- ii) a lower proportion of false alarms (FAs);
- iii) higher overall *d'* scores.

2. Colour information will enhance the relative distinctiveness of picture stimuli, resulting in enhanced recognition performance compared to previously utilised greyscale items. Exploratory analyses comparing the results of the current study with those of *Experiment 3* are expected to reveal numerical differences between the colour and greyscale findings, such that the colour items exhibit performance enhancements in the same direction as those outlined in Hypothesis 1 when compared to the greyscale items.

3. Any effects associated with manipulating the availability of response options at test (RFG/RFBG) will remain unaffected by the addition of colour information to picture stimuli. Based on the findings of *Experiment 3*, it is expected that:

    - i) The RFG group will produce a significantly higher proportion of overall hits compared to the RFBG group.
    - ii) The RFG group will produce a significantly higher proportion of FAs assigned *Familiarity* compared to the RFBG group.
    - iii) Significant interaction effects between response option and stimuli format will be evident in the analyses of mean *d'* scores, mean proportion of hits assigned *Recollection*, and mean proportion of hits assigned *Familiarity*.

Experiment: Effect of stimuli format (colour) and response option on recognition memory judgements.

Method

Participants

164 participants completed the experiment online (see Table 8 for a breakdown of the age/gender

of the current sample). All participants were required to be between the age of 18-59 years in order to meet our YA criteria (actual range: 18-57). As our experiment involved written words as to-be-remembered stimuli, we also asked that subjects first language be English; the vast majority (96.95%) reported that English was indeed their first language. Subjects were recruited from the voluntary participation website Prolific Academic (85.98%), where payment at the rate of £5/hr was given, and via the in-school research participation system (14.02%), where they received course participation credits. *G*Power* software was used to calculate an appropriate sample size; to detect a medium effect size of Cohen's f = 0.25 with 80% power (α = .05, two-tailed), 79 subjects per group would be necessary (N = 158) in a 3x2 mixed ANOVA.

Table 8: Gender and age (*SD*) of the current sample.

Gender	N	Age	
Female	*99*	31.72	*(11.16)*
Male	*61*	31.87	*(10.25)*
Non-binary	*1*	19.00	*(0)*
Transgender	*1*	32.00	*(0)*
Unspecified	***2***	**38.50**	***(3.54)***
Total	*164*	31.78	*(10.73)*

Materials

Stimuli were the same as those utilised in *Experiment 3*, except the greyscale drawings and photographs were substituted for their colour versions. Items consisted of 126 innocuous, everyday objects (e.g. clock, rabbit, shoe), presented across three individual stimuli formats: written words, line drawings, and photographs. Words and line drawings were sourced from Rossion & Pourtois (2004); the drawings consisted of shaded, colour illustrations, and the words were simply the written names of the depicted objects (these were presented in a clear Sans-serif typeface in the current experiment). A matching set of photograph stimuli were curated in *Experiment 2*; high quality photographs were sourced to similarly depict the same everyday objects as those found in the Rossion & Pourtois (2004) line drawings. In each photograph, the object

of interest was isolated from its original background and rotated to match the orientations shown in the line-drawn items. See Figure 16 for examples of each stimuli format.

Design

A mixed 3x2 design was utilised, consisting of a within-subjects factor of stimuli format (words, drawings, photographs) and a between-subjects factor of response option (RFG, RFBG). Counterbalancing was achieved via blocked randomisation, whereby participants were presented with: 1) one of six possible study lists (equal length, with the same number of words, drawings, and photographs); 2) one of two possible recognition tests (either RFG: “Recollection”, “Familiarity”, “Guessing”), or RFBG: “Recollection”, “Familiarity”, “Guessing”, “Both”). All counterbalancing routes were of equal length, and subjects were randomly assigned into blocks via balanced methods.

Procedure

The procedure was identical to that of *Experiment 3*; data collection was conducted online using the experiment platform Gorilla. All subjects completed three self-paced phases: i) study phase, ii) distractor task, and iii) recognition test. At study, subjects were instructed to learn each of the word, drawing, and photograph items (shown at random, one-at-a-time) in preparation for a later memory test. For each item, participants were required to report whether the current format was a word, drawing, or photograph - an encoding judgement that ensured attention was directed toward the to-be-remembered stimuli. Next, subjects completed some simple multiple choice mathematical questions (e.g. 6 x 4 = ?) as a distractor task. Finally, participants were presented with the recognition test; word, drawing, and photograph items were once again shown one-at-a-time at random. Half of the test items had been shown previously in the study phase, while the other half were new (not shown at study). Subjects were first required to make an *Old*/*New* judgement, based on whether they believed they had studied the item earlier or not. While *New* judgements simply led to the next item, *Old* judgements led to a follow-up screen where participants were asked whether they had recognised the item via *Recollection*, *Familiarity*, or were

simply *Guessing* that it was old. Those in the RFBG response option condition had an additional *Both* option at this stage, where they could report that they had experienced recollection and familiarity simultaneously. Stimuli format stayed the same across study and test (e.g. if the item "penguin" was shown in word format at study, it was also shown as a word at test), and the same concepts were not repeated across the other formats within-subjects (e.g. if the item "penguin" was shown as a word, that subject would not view the drawing or photo version).

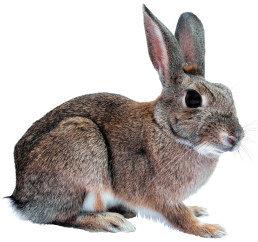
Words:	Drawings:	Photographs:
cloud		
lock		
penguin		

Figure 16: Examples of the word, colour drawing, and colour photograph stimuli utilised in the current experiment.

Data processing

The primary DVs of interest consisted of the mean proportion of hits and false alarms (FAs), mean *d'* scores (d-prime, a signal detection measure of sensitivity), and the total number of hits

and FAs assigned to each of the available response options (R/F/G or R/F/B/G). Proportions of Recollection and Familiarity were calculated slightly differently depending on the response option condition; in the RFG-judgement group, simple proportions were created from the total number of R responses and the total number of F responses. In the RFBG condition, the proportion of Both responses were added separately to R proportions and F proportions. All analyses were conducted using R (R Core Team, 2020) using the 'rstatix' package (v0.6.0; Kassambara, 2020).

Subjects were excluded from analyses on the basis of two key criteria; 1) less than 90% accuracy during the encoding task ("Is this a word, drawing, or photograph?"); 2) extreme z-scores (those presenting z-scores of +/- 3 for total hits, total FAs, or overall recognition [hits minus FAs]). A total of 3 participants were found to meet (at least) one of these criteria, and were thus considered outliers and excluded from analysis, leaving a total of 161 datasets.

Results

A series of 3x2 mixed ANOVAs were conducted on each of the DVs, with a within-subjects factor of stimuli format (words / colour drawings / colour photos) and a between-subjects factor of response option (RFG / RFBG). Significant main effects and interaction effects were followed-up with Bonferroni-adjusted pairwise comparisons. Greenhouse-Geisser corrections were applied when ANOVA data was found to violate the assumption of sphericity (assessed according to Mauchly's test statistic).

Stimuli distinctiveness

Mean proportions of hits and FAs, and mean *d'* scores are presented for Experiments 3 and 4 in Table 9. Visual inspection of the data shows some expected patterns with regard to stimuli distinctiveness. As the intended distinctiveness increases (from words, to drawings, to photographs), the i) mean proportion of hits increase; ii) mean proportion of FAs decrease; iii) mean *d'* scores increase.

Table 9: Data from Experiment 3 (using greyscale stimuli) shown alongside that of the current experiment (using colour stimuli), showing the mean proportion of hits, FAs, and mean *d'* scores, by stimuli format and response option condition. Signif. codes: ***$p < .001$; **$p < .01$; *$p < .05$; + involved in significant interaction.

Experiment 3: Grey		Hits	FAs	d'	**Experiment 4: Colour**		Hits	FAs	d'
Stimuli format:		***	***	+	**Stimuli format:**		***	***	***
	Words	0.54	0.21	1.15		Words	0.55	0.23	1.11
	Drawings	0.76	0.09	2.38		Drawings	0.73	0.08	2.39
	Photographs	0.85	0.05	3.08		Photographs	0.87	0.04	3.25
Response option:		*		+	**Response option:**			*	
	RFG	0.74	0.13	2.25		RFG	0.74	0.13	2.25
	RFBG	0.69	0.11	2.16		RFBG	0.7	0.1	2.25

Results from the ANOVAs demonstrated a significant main effect of stimuli format for each of the key variables of interest; hits ($F(1.70, 271.03) = 187.25$, $p < .001$, $\eta_p^2 = .54$), FAs ($F(1.26, 200.79) = 123.14$, $p < .001$, $\eta_p^2 = .44$), and *d'* scores ($F(2, 318) = 465.93$, $p < .001$, $\eta_p^2 = .75$) - though no interaction effects were evident between stimuli format and response option; hits ($F(1.70, 271.03) = 1.22$, $p = .291$, $\eta_p^2 < .01$), FAs($F(1.26, 200.79) = 2.72$, $p = .092$, $\eta_p^2 = .02$), or *d'* scores ($F(2, 318) = 0.20$, $p = .817$, $\eta_p^2 < .01$).

To determine whether photo superiority effects were exhibited in the current set of colour stimuli - comparable to those previously observed using greyscale items (Experiment 3) - pairwise t-tests were performed between the colour photos and drawings. For the mean proportion of hits, colour photographs (M= 0.87) exhibited a significantly higher proportion than colour drawings (M= 0.73), $t(160) = -11.04$, $p < .001$; $d = -0.87$, 95% CI [-1.04, -0.74]. The photographs (M= 0.04) also produced significantly fewer FAs compared to drawings (M= 0.08), $t(160) = 6.36$, $p < .001$; $d = 0.5$, 95% CI [0.38, 0.64]). Mean *d'* scores were also significantly higher for the colour photographs (M= 3.25) compared to the colour drawings (M= 2.39), $t(160) = -13.56$, $p < .001$; $d = -1.07$, 95% CI [-1.27, -0.91]. These findings replicate those found previously using greyscale items, whereby photographs offer a number of recognition benefits when compared to less detailed line-drawn illustrations.

Visual inspection of the data (and significant results) reveals only one difference with regard to response option when compared to that obtained in *Experiment 3*: the ANOVA on *d'* scores failed to produce a significant interaction between stimuli format and response option in the current experiment, as was previously demonstrated. Previous follow-up comparisons revealed this interaction was driven by numerically higher *d'* scores for drawings in the RFBG group compared to the RFG group - a deviation from words and photographs, whereby *d'* scores were both higher in the RFG group compared to the RFBG group. The difference between *d'* scores for drawings in the RFG and RFBG groups did not reach significance though, and this negligible difference may explain why such an interaction was absent in the current study.

Recollection and Familiarity

To determine the effects of stimuli format and response option on the classification of recognition memory judgements, separate 3 (stimuli format: words, drawings, photographs) x 2 (response option condition: RFG-judgements, RFBG-judgements) mixed ANOVAs were conducted on the mean proportion of hits assigned Recollection, Familiarity, and Guessing (see Figure 17), and the mean proportion of FAs assigned RFG (see Figure 18)

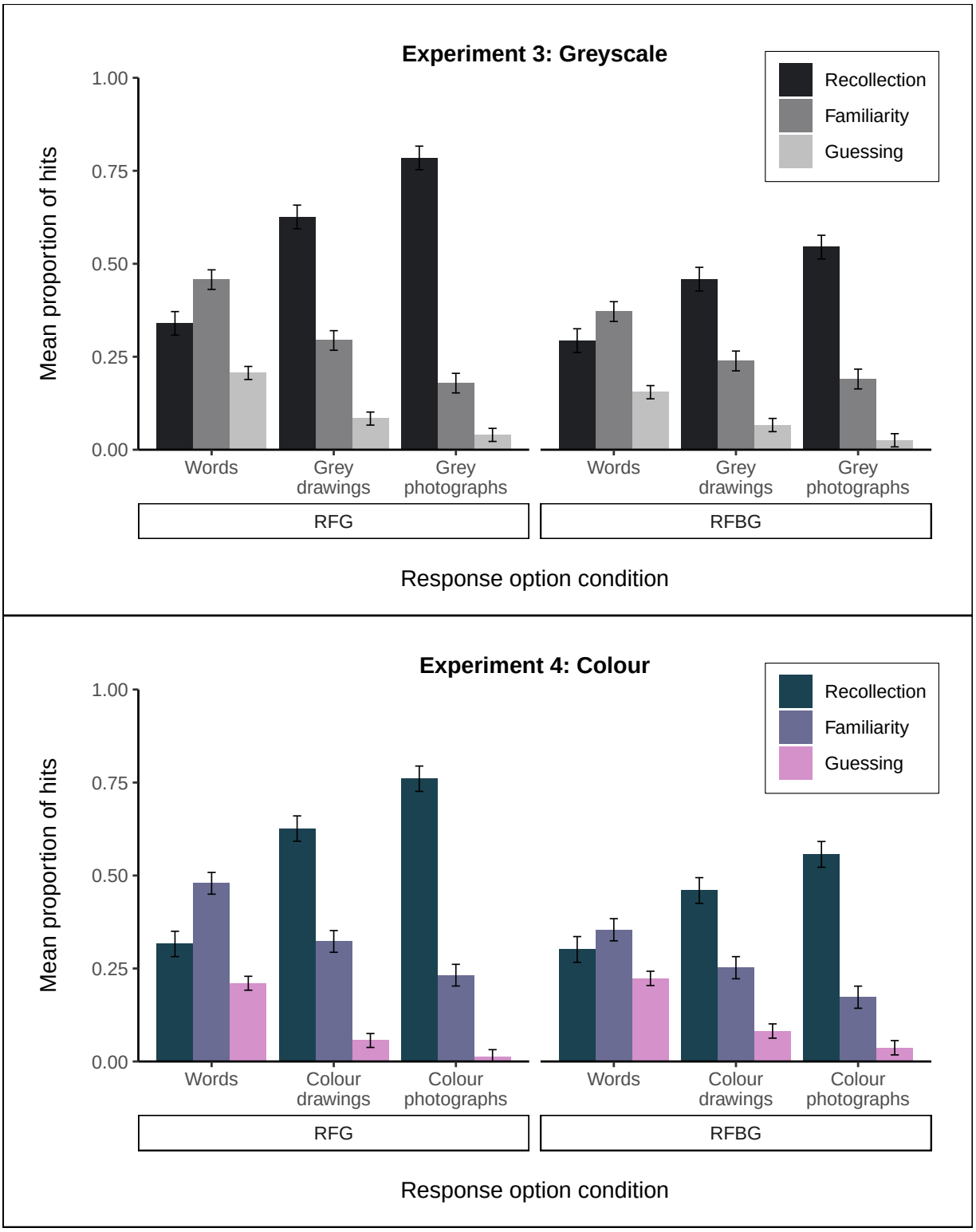


Figure 17: Proportion of hits assigned Recollection, Familiarity, and Guessing, by stimuli format and response option condition.

Recollection (hits): Results from the ANOVA on the mean proportion of hits assigned Recollection showed a significant interaction between stimuli format and response option condition, $F(1.39, 221.56) = 10.79$, $p < .001$, $\eta_p^2 = .06$ (see Figure 19).

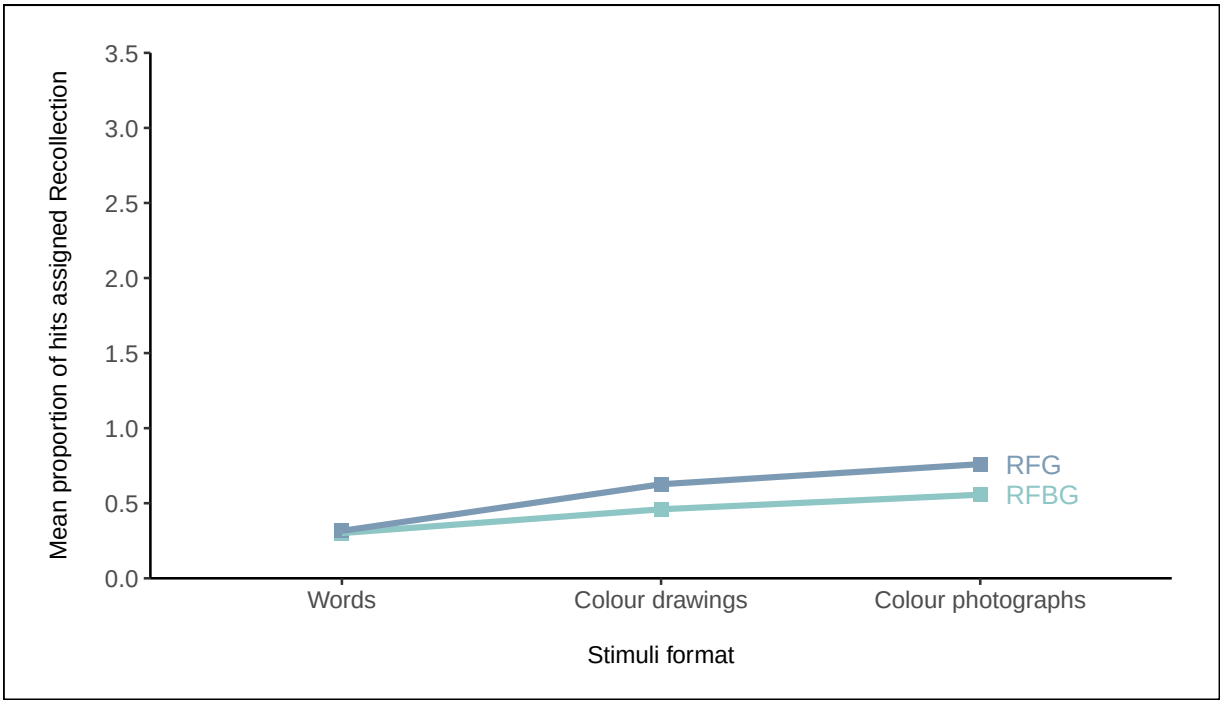


Figure 19: Interaction plot between stimuli format and response option for the mean proportion of hits assigned Recollection.

Comparisons across stimuli formats showed colour photographs produced a significantly higher proportion of R hits than both words and colour drawings in both the RFG group (colour photographs [*M*= 0.76] vs. words [*M*= 0.32], $t(318) = -15.02$, $p < .001$; colour photographs [*M*= 0.76] vs. colour drawings [*M*= 0.63], $t(318) = -4.53$, $p < .001$) and the RFBG group (colour photographs [*M*= 0.56] vs. words [*M*= 0.3], $t(318) = -8.17$, $p < .001$; colour photographs [*M*= 0.56] vs. colour drawings [*M*= 0.46], $t(318) = -3.11$, $p = .031$). Likewise, colour drawings produced a significantly higher proportion of R hits in comparison to Words in both the RFG (colour drawings [*M*= 0.63] vs. words [*M*= 0.32], $t(318) = -10.49$, $p < .001$) and RFBG conditions (colour drawings [*M*= 0.46] vs. words [*M*= 0.3], $t(318) = -5.06$, $p < .001$).

The interaction is evident following comparisons of the same stimuli format across response option conditions. The RFG group produced a significantly higher proportion of R hits than the

RFBG group for colour photographs (RFG [*M* = 0.76] vs. RFBG [*M* = 0.56], *t*(274.37) = -4.19, *p* = .001) and for colour drawings (RFG [*M* = 0.63] vs. RFBG [*M* = 0.46], *t*(274.37) = -3.43, *p* = .011). However, this was not the case for words, where there was no difference in the proportion of R hits between the RFG (*M* = 0.32) and RFBG groups (*M* = 0.3; *t*(274.37) = -0.30, *p* > .999).

Familiarity (hits): Results from the ANOVA on the mean proportion of hits assigned Familiarity again showed a significant main effect of stimuli format *F*(1.49, 236.15) = 50.18, *p* < .001, η_p^2 = .24. Colour photographs (*M*= 0.2) produced significantly fewer F hits than both words (*M*= 0.42), *t*(160) = -8.34, *p* < .001; *d* = -0.66, 95% CI [-0.87, -0.49], and colour drawings (*M*= 0.29), *t*(160) = 5.97, *p* < .001; *d* = 0.47, 95% CI [0.32, 0.64]. The colour drawings (*M*= 0.29) also showed significantly fewer F hits compared to words (*M*= 0.42), *t*(160) = -5.77, *p* < .001; *d* = -0.45, 95% CI [-0.64, -0.28]. There were no significant interaction effects between stimuli format and response option condition, *F*(1.49, 236.15) = 1.33, *p* = .263, η_p^2 < .01.

Guessing (hits): The ANOVA on the mean proportion of hits assigned Guessing demonstrated a significant main effect of stimuli format *F*(1.29, 204.70) = 82.24, *p* < .001, η_p^2 = .34. Colour photographs (*M*= 0.02) produced significantly fewer G hits in comparison to both words (*M*= 0.22; *t*(160) = -10.18, *p* < .001; *d* = -0.8, 95% CI [-0.92, -0.7]) and colour drawings (*M*= 0.07; *t*(160) = 5.5, *p* < .001; *d* = 0.43, 95% CI [0.32, 0.56]). The colour drawings (*M*= 0.07) also showed a significantly lower proportion of G hits compared to words (*M*= 0.22; *t*(160) = -8.41, *p* < .001; *d* = -0.66, 95% CI [-0.79, -0.54]. There were no significant interaction effects between stimuli format and response option condition *F*(1.29, 204.70) = 0.09, *p* = .824, η_p^2 < .01.

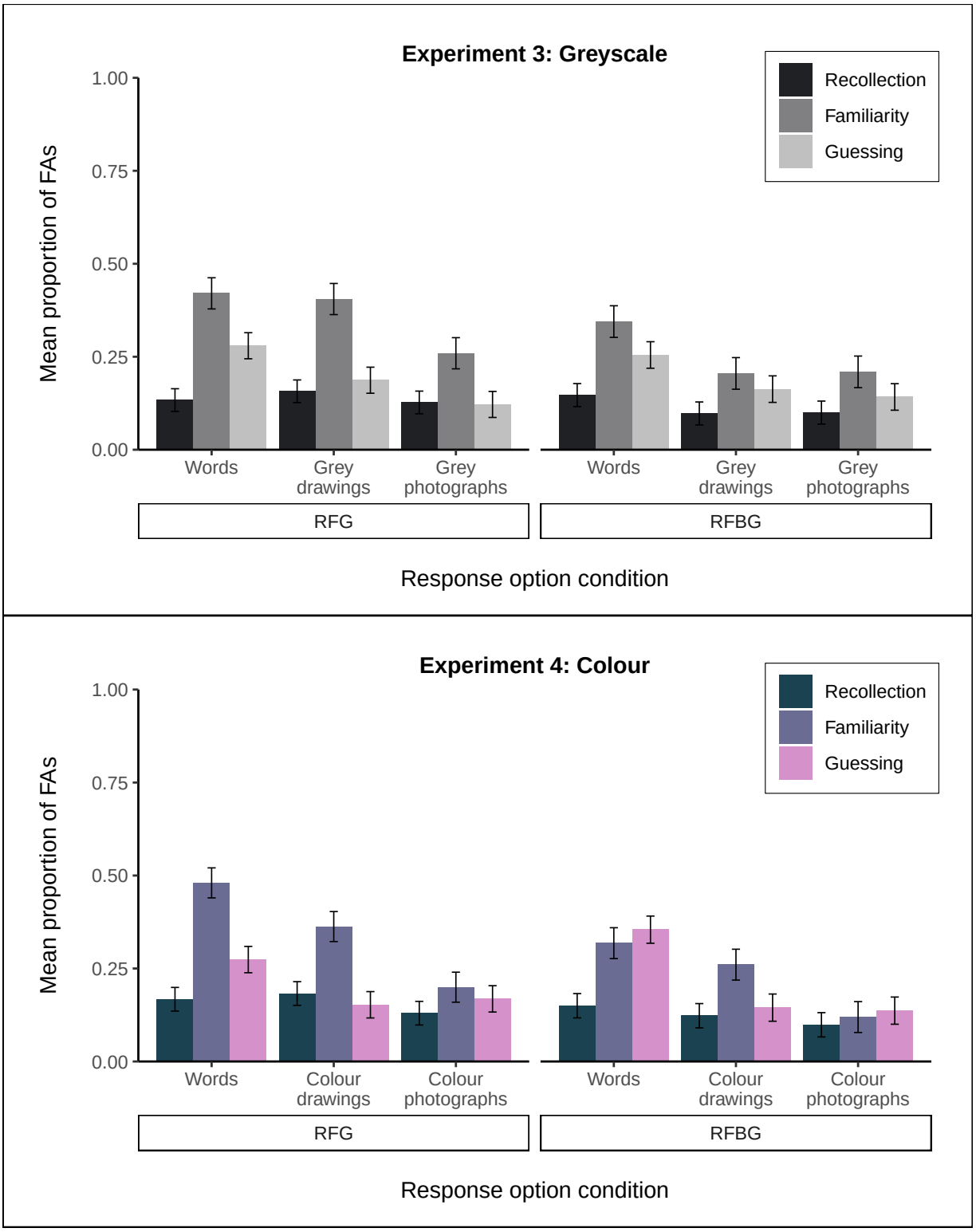


Figure 18: Proportion of FAs assigned Recollection, Familiarity, and Guessing, by stimuli format and response option condition.

Recollection (FAs): For FAs assigned *Recollection*, there was no significant main effect of stimuli format [$F(2, 318) = 1.58$, $p = .207$, $\eta_p^2 < .01$ or interaction [$F(2, 318) = 0.32$, $p = .727$, $\eta_p^2 < .01$].

Familiarity (FAs): The ANOVA for FAs assigned *Familiarity* showed a significant main effect of stimuli format, $F(2, 318) = 22.92$, $p < .001$, $\eta_p^2 = .13$. Colour photographs ($M = 0.16$) produced significantly fewer F FAs than words ($M = 0.4$), $t(160) = -6.41$, $p < .001$; $d = -0.51$, 95% CI [-0.68, -0.34]. Likewise, colour drawings ($M = 0.31$) also showed a significantly lower proportion of FAs compared to words ($M = 0.4$), $t(160) = -2.45$, $p = 0.05$; $d = -0.19$, 95% CI [-0.35, -0.04]. However, there was no significant difference in the proportion of FAs assigned Familiarity between colour photographs ($M = 0.16$) and colour drawings ($M = 0.31$), $t(160) = 4.65$, $p < .001$; $d = 0.37$, 95% CI [0.21, 0.52]. There were no significant interaction effects between stimuli format and response option condition, $F(2, 318) = 0.70$, $p = .498$, $\eta_p^2 < .01$.

Guessing (FAs): The ANOVA on the mean proportion of FAs assigned *Guessing* demonstrated a significant main effect of stimuli format $F(2, 318) = 16.11$, $p < .001$, $\eta_p^2 = .09$. Colour photographs ($M = 0.15$) produced significantly fewer G FAs in comparison to words ($M = 0.31$), $t(160) = -4.5$, $p < .001$; $d = -0.35$, 95% CI [-0.51, -0.21]). Likewise, colour drawings ($M = 0.15$) also showed a significantly lower proportion of FAs compared to words ($M = 0.31$), $t(160) = -4.85$, $p < .001$; $d = -0.38$, 95% CI [-0.55, -0.23]. However, there was no significant difference in the proportion of FAs assigned Guessing between colour photographs ($M = 0.15$) and colour drawings ($M = 0.15$), $t(160) = -0.15$, $p = 1$; $d = -0.01$, 95% CI [-0.17, 0.15]). There were also no significant interaction effects between stimuli format and response option condition $F(2, 318) = 1.57$, $p = .210$, $\eta_p^2 < .01$.

Visual inspection of the data in Figure 17 and Figure 18 demonstrates a highly similar pattern of responding between *Experiment 3* and the current study, and suggest the addition of colour had little impact on RFG response patterns.

Response option availability

In each of the aforementioned ANOVAs, the role of response option was also examined to de-