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 occuring 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 underlying 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-occurance 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[1] (52.15%) (where they received payment at the rate of £5/hr) and via the in-school research participation system[2] (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

[1] https://www.prolific.co/

[2] https://keelepsychology.sona-systems.com/

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.

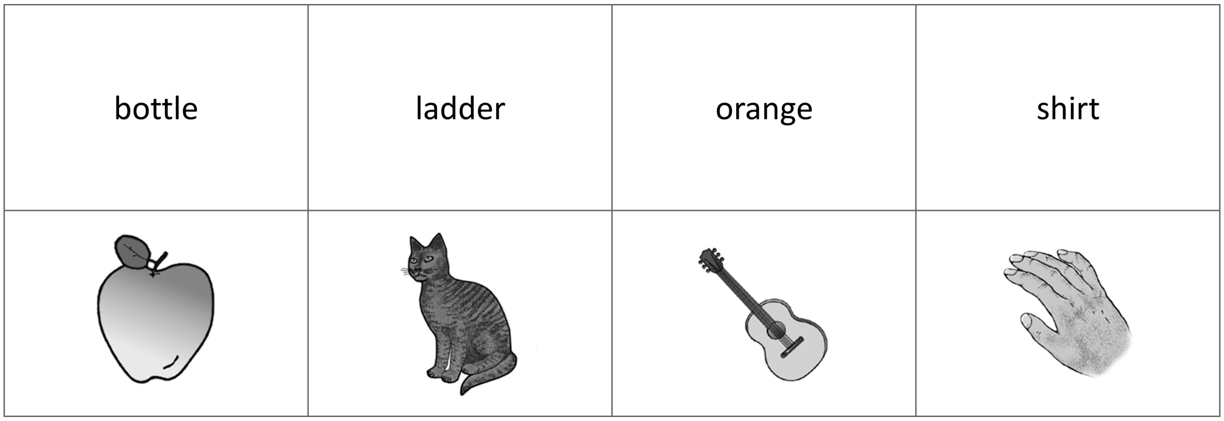

Figure 1: Example word and picture stimuli from the current study.

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[3]. 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 stim-

[3] https://www.qualtrics.com/uk/

uli, 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]. To further examine response patterns between pictures and words, ANOVAs were also run on 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.

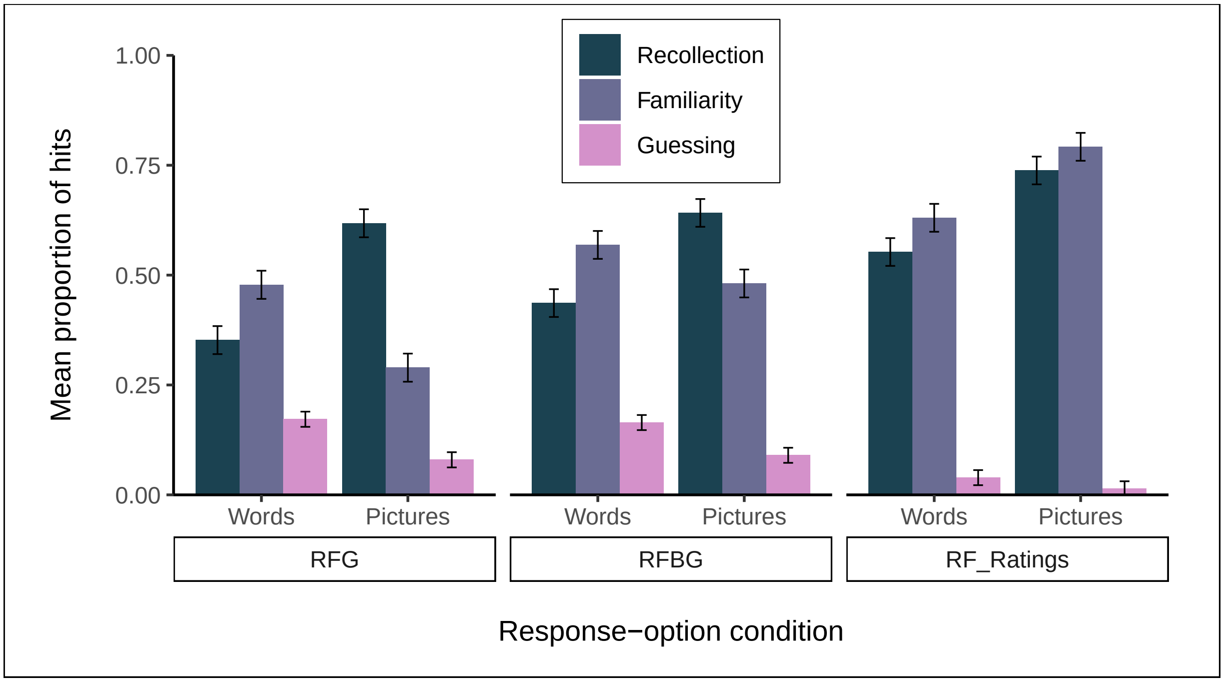

Figure 2: Proportion of hits assigned *Recollection*, *Familiarity*, and *Guessing*, by stimuli-format and response-option condition.

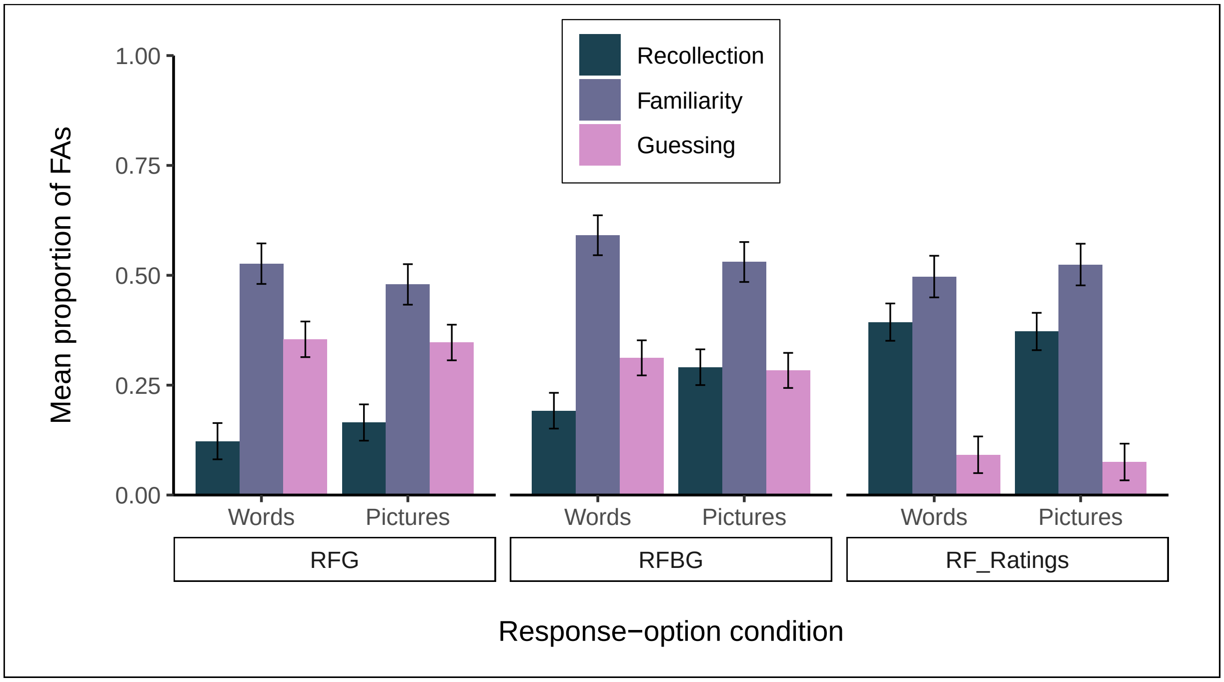

Figure 3: Proportion of FAs assigned *Recollection*, *Familiarity*, and *Guessing*, by stimuli-format and response-option condition.

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* $= 0.04$, $p = .331$, $\hat{\eta}^2_G = .004$.

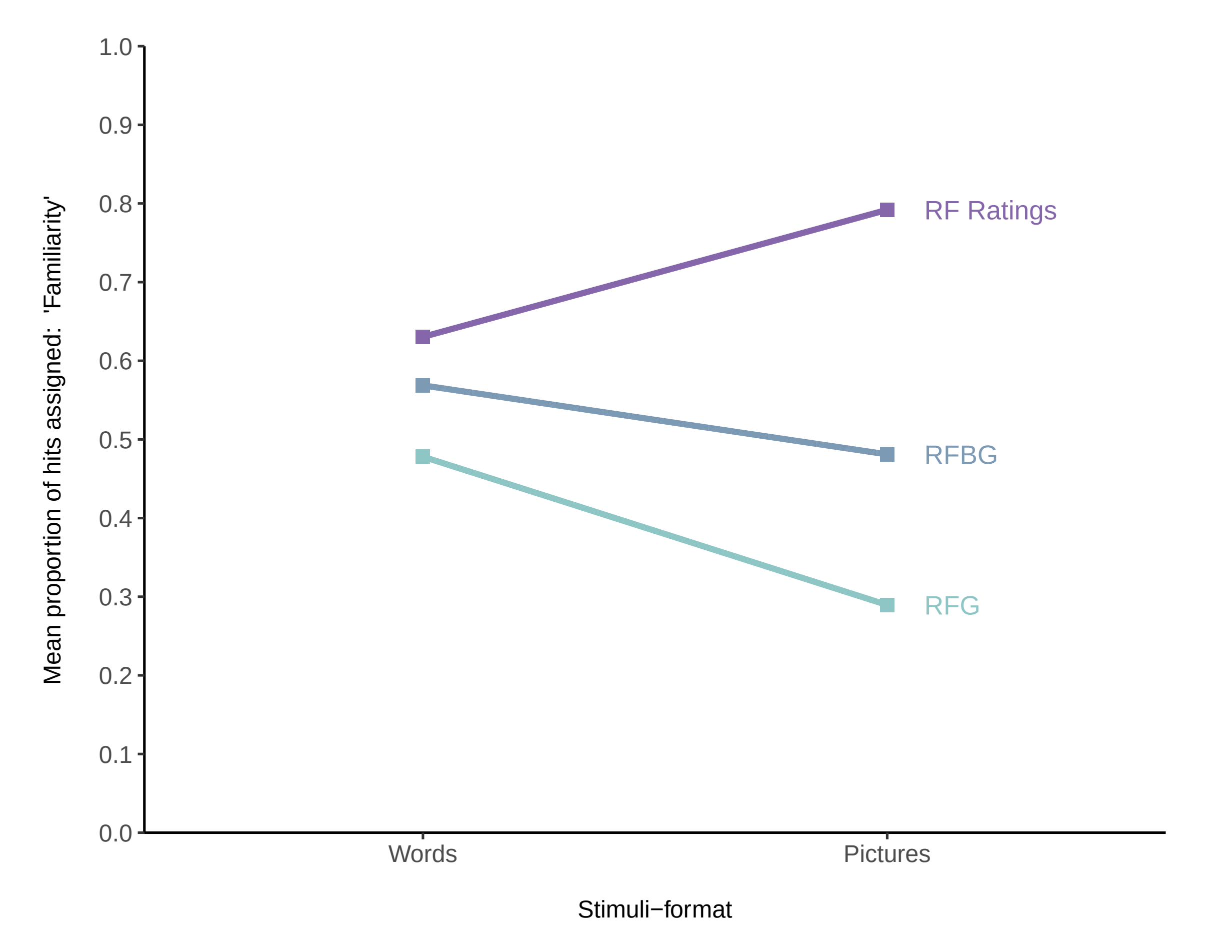

Figure 4: Interaction plot between stimuli-format and response-option condition for the mean proportion of hits assigned *Familiarity*.

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.

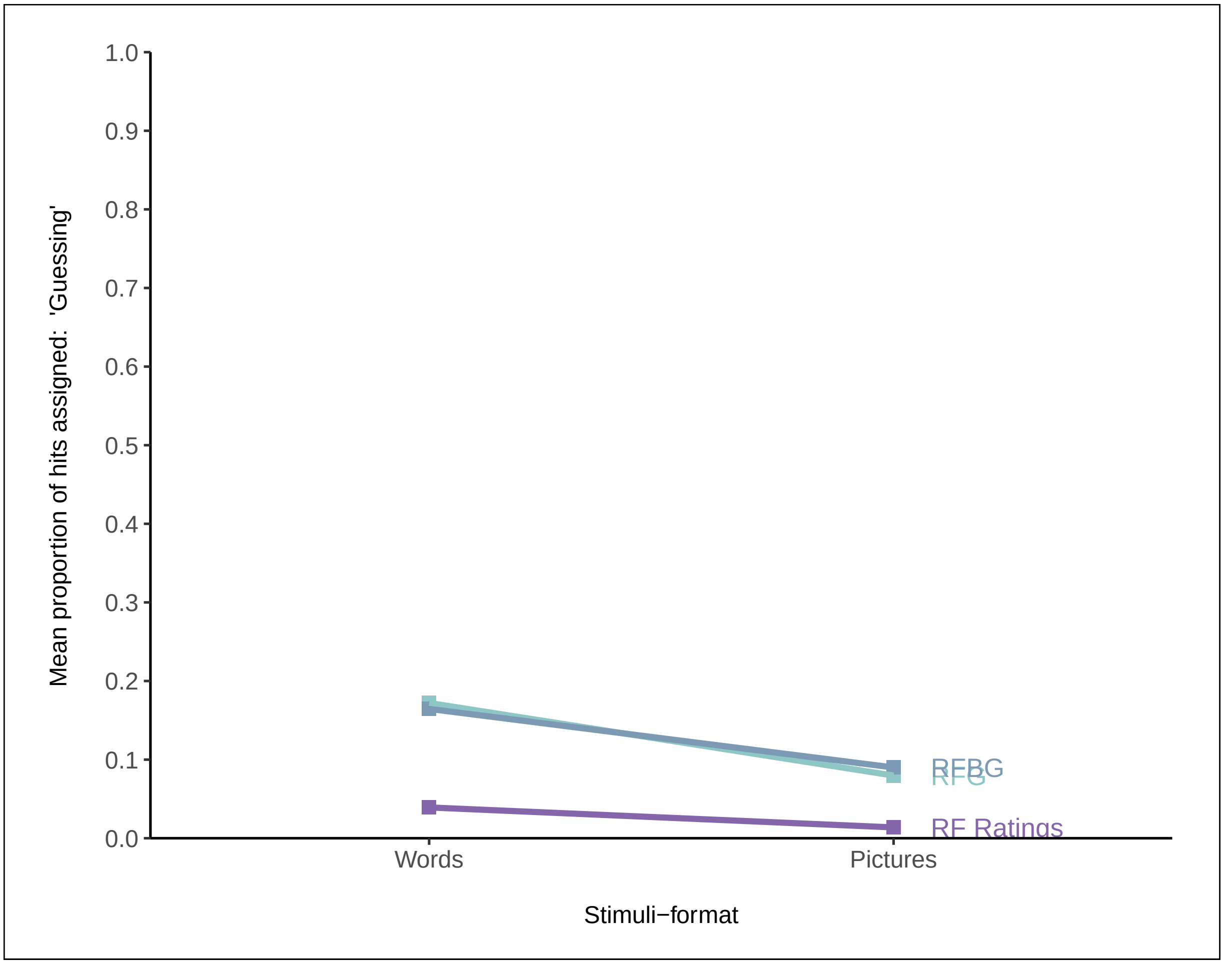

Figure 5: Interaction plot between stimuli-format and response-option condition for the mean proportion of hits assigned 'Guessing'.

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.

######—————————————-