Contents

```
##    Clear environment and list required packages:
      rm(list=ls())
      packages <- c("tidyverse", "dplyr", "captioner", "tinytex", "kableExtra", "formattab
                    "psycho", "afex", "emmeans", "janitor", "psych", "rstatix", "ggpubr",
                    "gridExtra", "bibtex", "apa", "car")
      packages_to_install <- packages[!(packages %in% installed.packages()[,"Package"])]
      if(length(packages_to_install)) install.packages(packages_to_install)
      invisible(suppressPackageStartupMessages(lapply(packages, library, character.only =
      #devtools::install_github("crsh/papaja")
      library(papaja)
      #tinytex::install_tinytex()
##    Write references file for all of the packages used, and general R ref:
      #write.bib(packages, "resources/packages.bib")
      #write.bib(citation(), "resources/base.bib")
      options(scipen=999)   # Turn off scientific notation for numbers printed to the Cons
##############################################################################################
##  Define function to get package version:
    packages_versions <- function(p) {
        paste(packageDescription(p)$Package, packageDescription(p)$Version, sep = " ")}
##  Define function used in data tidying:
    coalesce_by_column <- function(df) {return(dplyr::coalesce(!!! as.list(df)))}
##  Define function to format t-test results:
    ttest.writeup <- function(i){
        paste0("*t*(", i$df, ") = ",
            round(i$statistic, 2), ", *p* ",
                ifelse(i$p.adj < .001, "< .001",
```

```
                    paste0("= ", round(i$p.adj, 2)))))}



    independent.ttest.writeup <- function(i){
        paste0("*t*(", (i$n1 + i$n2) - 2, ") = ",
            round(i$statistic, 2), ", *p* ",
                ifelse(i$p.adj < .001, "< .001",
                    paste0("= ", round(i$p.adj, 2)))))}



##  Captioner:
    figure_numbering <- captioner()
    table_numbering <- captioner(prefix = "Table")
    appendix_numbering <- captioner(prefix = "Appendix", type = c("C")) ## Use letters ra
    appendix_numbering("spelling.corrections",
                    "Manipulations / spelling corrections of naming responses (Experime
    appendix_numbering("exp2.full.norms",
                    "Normative data for newly curated photograph items (Experiment 2).'
```

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

Background

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, & Jen-

nings, 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 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 par-

ticular 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;
    - iii) 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]; 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.

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

[1] https://www.prolific.co/
[2] https://keelepsychology.sona-systems.com/

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

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[3]. Subjects initially com-

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

pleted 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 with *R* (R Core Team, 2020) using the *afex* (v0.28-0; Singmann, Bolker, Westfall, Aust, & Ben-Shachar, 2020) and *rstatix* packages (v0.6.0; Kassambara, 2020).

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 options at test, a series of 2 (stimuli format: words, pictures) x 3 (response option 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.

The ANOVA on *d'* scores demonstrated a significant main effect of stimuli-format, $F(1, 183) = 278.32$, $p < .001$, $\eta_p^2 = .60$, a PSE was evident, with pictures ($M = 1.73$) producing significantly better discrimination between hits and FAs than words ($M = 0.92$), $t(185) = 16.77$, $p < .001$; $d = 1.23$, 95% CI [1.06, 1.44]. 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, $F(1, 183) = 131.77$, $p < .001$, $\eta_p^2 = .42$, with pictures ($M = 0.62$) showing a higher number of overall hits compared to words ($M = 0.47$), $t(185) = 11.55$, $p < .001$; $d = 0.85$, 95% CI [0.68, 1.05]. Similarly, the ANOVA on the proportion of FAs showed a significant main effect of stimuli-format, $F(1, 183) = 61.18$, $p < .001$, $\eta_p^2 = .25$, with pictures ($M = 0.12$) producing significantly fewer FAs than words ($M = 0.21$), $t(185) = -7.81$, $p < .001$; $d = -0.57$, 95% CI [-0.69, -0.45]. 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.

Table 2: 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.

	Hits	FAs	d'
Stimuli format			
Words	0.47	0.21	0.92
Pictures	0.62	0.12	1.73
Response option			
RFG	0.62	0.19	1.44
RFBG	0.54	0.16	1.28
RF-Ratings	0.48	0.14	1.24

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.

Recollection (hits): For R hits, there was a significant interaction between stimuli format and response option option condition, $F(2, 183) = 3.62$, $p = .029$, $\eta_p^2 = .04$ (see Figure 4).

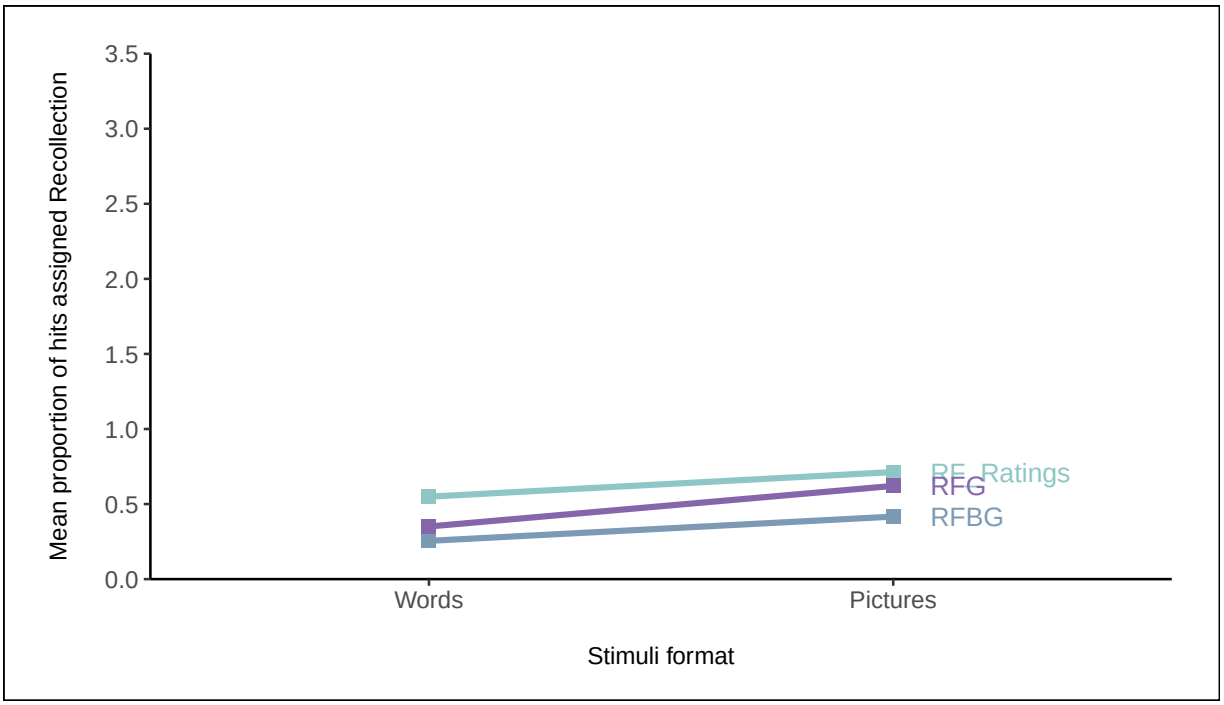


Figure 4: Interaction plot between stimuli format and response option for the mean proportion of hits assigned Recollection.

Pictures produced a higher proportion of hits assigned *Recollection* than words in the RFG group (words [*M*= 0.35] vs. pictures [*M*= 0.62], *t*(183) = -8.18, *p* < .001), RFBG group (words [*M*= 0.25] vs. pictures [*M*= 0.42], *t*(183) = -5.09, *p* < .001) and RF-Ratings group (words [*M*= 0.55] vs. pictures [*M*= 0.71], *t*(183) = -5.12, *p* < .001).

The interaction is evident following comparisons of the same stimuli format across response option conditions. For words, the proportion of R hits was significantly higher in the RF-Ratings group compared to both the RFG group (RFG [*M* = 0.35] vs. RF-Ratings [*M* = 0.55], *t*(279.16) =

4.07, p = .001) and the the RFBG group (RFBG [M = 0.25] vs. RF-Ratings [M = 0.55], t(279.16) = 6.15, p < .001). The RFG and RFBG groups did not significantly differ in the proportion of word hits assigned *Recollection* (RFG [M = 0.35] vs. RFBG [M = 0.25], t(279.16) = -1.97, p = .755).

For pictures, the RF-Ratings group again showed a significantly higher proportion of R hits than the RFBG group (RFBG [M = 0.42] vs. RF-Ratings [M = 0.71], t(279.16) = 6.20, p < .001). While words produced a significant difference between the RFG and RF-Ratings group for R hits, the same was not found for pictures (RFG [M = 0.62] vs. RF-Ratings [M = 0.71], t(279.16) = 1.89, p = .900). Instead, the RFG group also produced a significantly higher proportion of R hits than the RFBG group (RFG [M = 0.62] vs. RFBG [M = 0.42], t(279.16) = -4.20, p = .001).

Familiarity (hits): For F hits, there was a significant interaction between stimuli format and response option option condition, F(2, 183) = 28.27, p < .001, η_p^2 = .24 (see Figure 5).

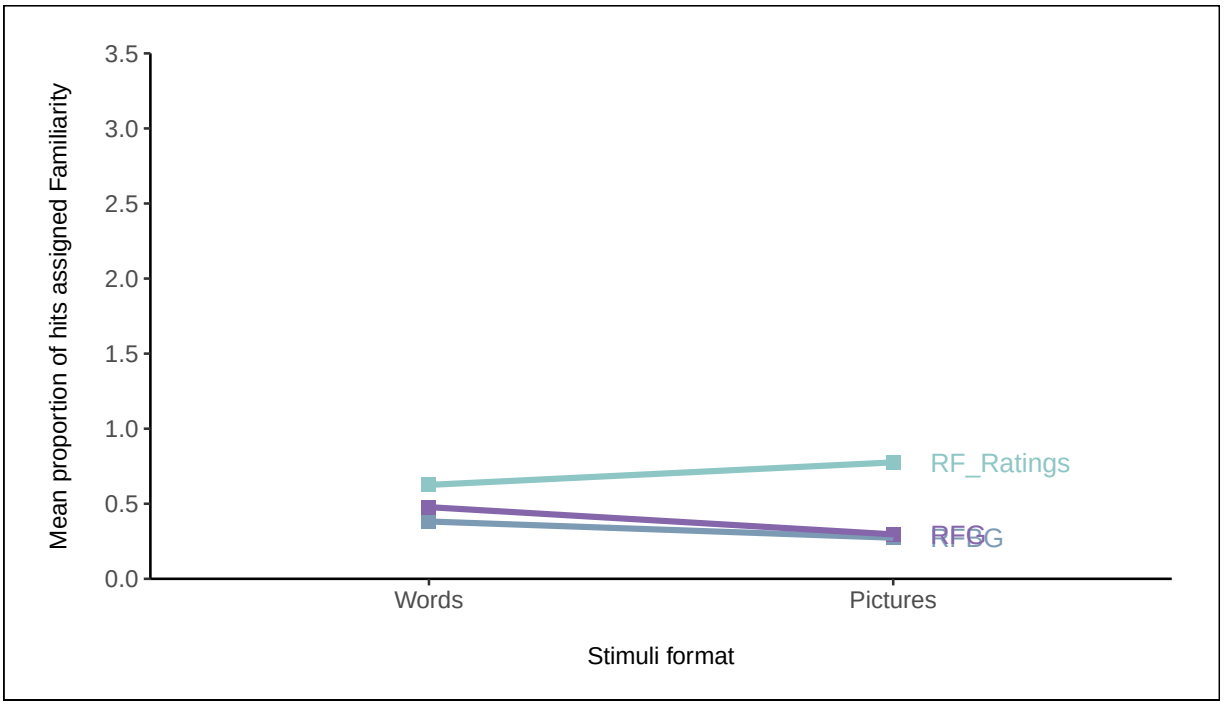


Figure 5: Interaction plot between stimuli format and response option for the mean proportion of hits assigned Recollection.

Pictures produced a higher proportion of hits assigned *Familiarity* than words in the RF-Ratings

group only (words [*M*= 0.63] vs. pictures [*M*= 0.78], *t*(183) = -4.62, *p* < .001). In both the RFG and RFBG groups, words produced a higher proportion of hits assigned *Familiarity* than pictures (RFG: words [*M*= 0.48] vs. pictures [*M*= 0.3], *t*(183) = 5.41, *p* < .001; RFBG: words [*M*= 0.38] vs. pictures [*M*= 0.27], *t*(183) = 3.36, *p* = .014).

Comparisons of the same stimuli format across response option conditions showed that, for words, the proportion of F hits was significantly higher in the RF-Ratings group compared to both the RFG group (RFG [*M* = 0.48] vs. RF-Ratings [*M* = 0.63], *t*(302.47) = 3.31, *p* = .016) and the the RFBG group (RFBG [*M* = 0.38] vs. RF-Ratings [*M* = 0.63], *t*(302.47) = 5.56, *p* < .001). The RFG and RFBG groups did not significantly differ in the proportion of word hits assigned *Familiarity* (RFG [*M* = 0.48] vs. RFBG [*M* = 0.38], *t*(302.47) = -2.14, *p* = .499).

For pictures, the RF-Ratings group again showed a significantly higher proportion of F hits than both the RFG group (RFG [*M* = 0.3] vs. RF-Ratings [*M* = 0.78], *t*(302.47) = 10.70, *p* < .001) and the RFBG group (RFBG [*M* = 0.27] vs. RF-Ratings [*M* = 0.78], *t*(302.47) = 11.44, *p* < .001). Similar to words, the RFG and RFBG groups did not significantly differ in the proportion of F hits produced by pictures (RFG [*M* = 0.3] vs. RFBG [*M* = 0.27], *t*(302.47) = -0.50, *p* > .999).

Guessing (hits): For G hits, there was a significant interaction between stimuli format and response option option condition, *F*(2, 183) = 3.99, *p* = .020, η_p^2 = .04 (see Figure 6).

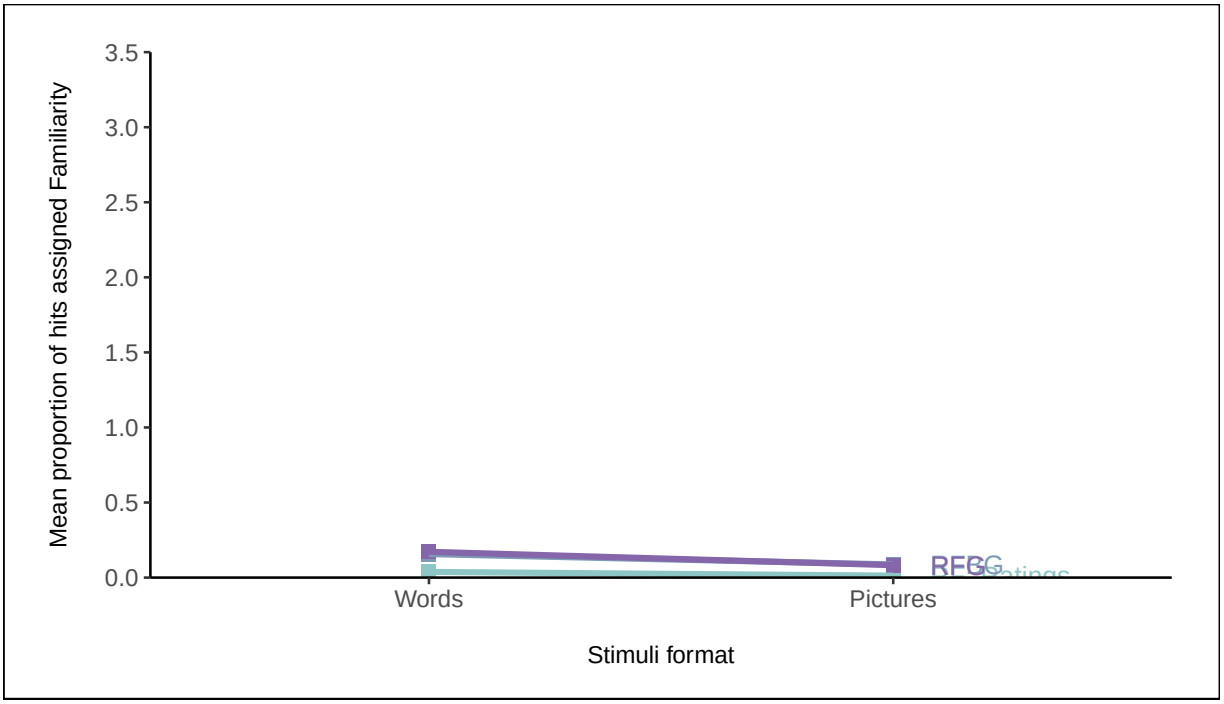


Figure 6: Interaction plot between stimuli format and response option for the mean proportion of hits assigned Recollection.

Words produced a higher proportion of hits assigned *Guessing* than pictures in both the RFG group (words [*M*= 0.17] vs. pictures [*M*= 0.08], $t(183) = 5.28$, $p < .001$) and RFBG group [*M*= 0.16] vs. pictures [*M*= 0.09], $t(183) = 4.26$, $p < .001$). Words and pictures did not significantly differ in RF-Ratings group (words [*M*= 0.04] vs. pictures [*M*= 0.01], $t(183) = 1.51$, $p > .999$).

Comparisons of the same stimuli format across response option conditions showed that, for words, the proportion of G hits was significantly lower in the RF-Ratings group compared to both the RFG group (RFG [*M* = 0.17] vs. RF-Ratings [*M* = 0.04], $t(286.62) = -5.56$, $p < .001$) and the RFBG group (RFBG [*M* = 0.16] vs. RF-Ratings [*M* = 0.04], $t(286.62) = -5.03$, $p < .001$). The RFG and RFBG groups did not significantly differ in the proportion of word hits assigned *Guessing* (RFG [*M* = 0.17] vs. RFBG [*M* = 0.16], $t(286.62) = -0.64$, $p > .999$).

For pictures, the RF-Ratings group again showed a significantly lower proportion of G hits than the RFBG group (RFBG [*M* = 0.09] vs. RF-Ratings [*M* = 0.01], $t(286.62) = -3.15$, $p = .027$), however, the comparison with the RFG group did not reach significance (RFG [*M* = 0.08] vs. RF-

Ratings [M = 0.01], t(286.62) = -2.89, p = .063). Again, the RFG and RFBG groups did not significantly differ in the proportion of G hits produced by pictures (RFG [M = 0.08] vs. RFBG [M = 0.09], t(286.62) = 0.20, p > .999).

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.

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. There-

fore, 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

Background

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 7 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.

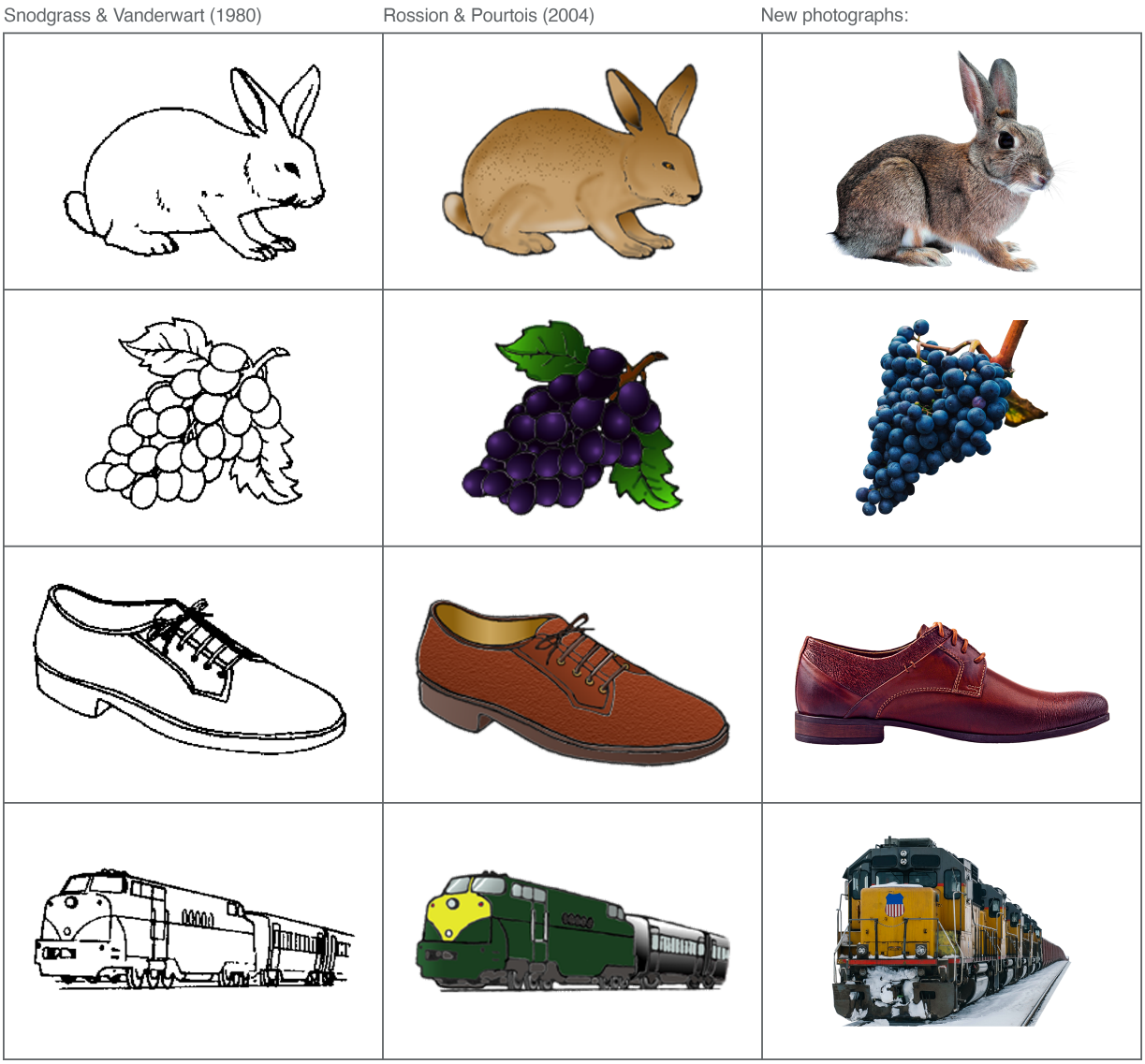


Figure 8: Examples of matching pictures across Snodgrass & Vanderwart (1980), Rossion & Pourtois (2004), and photographs from the current study. Greyscale versions of the drawings and photographs are not presented in this example.

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[4] (where they received payment at the rate of £5/hr), and via the in-school research participation system[5] (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 shaded 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

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

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

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 shaded 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[6]; Pexels[7]) for photographs that depicted the same everyday objects as the shaded 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 unfeathered). 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 shaded 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 9). The primary aim of the current study was

[6]https://unsplash.com/
[7]https://www.pexels.com/

to obtain photographs that could be clearly distinguished as a unique stimuli format among words and shaded 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).

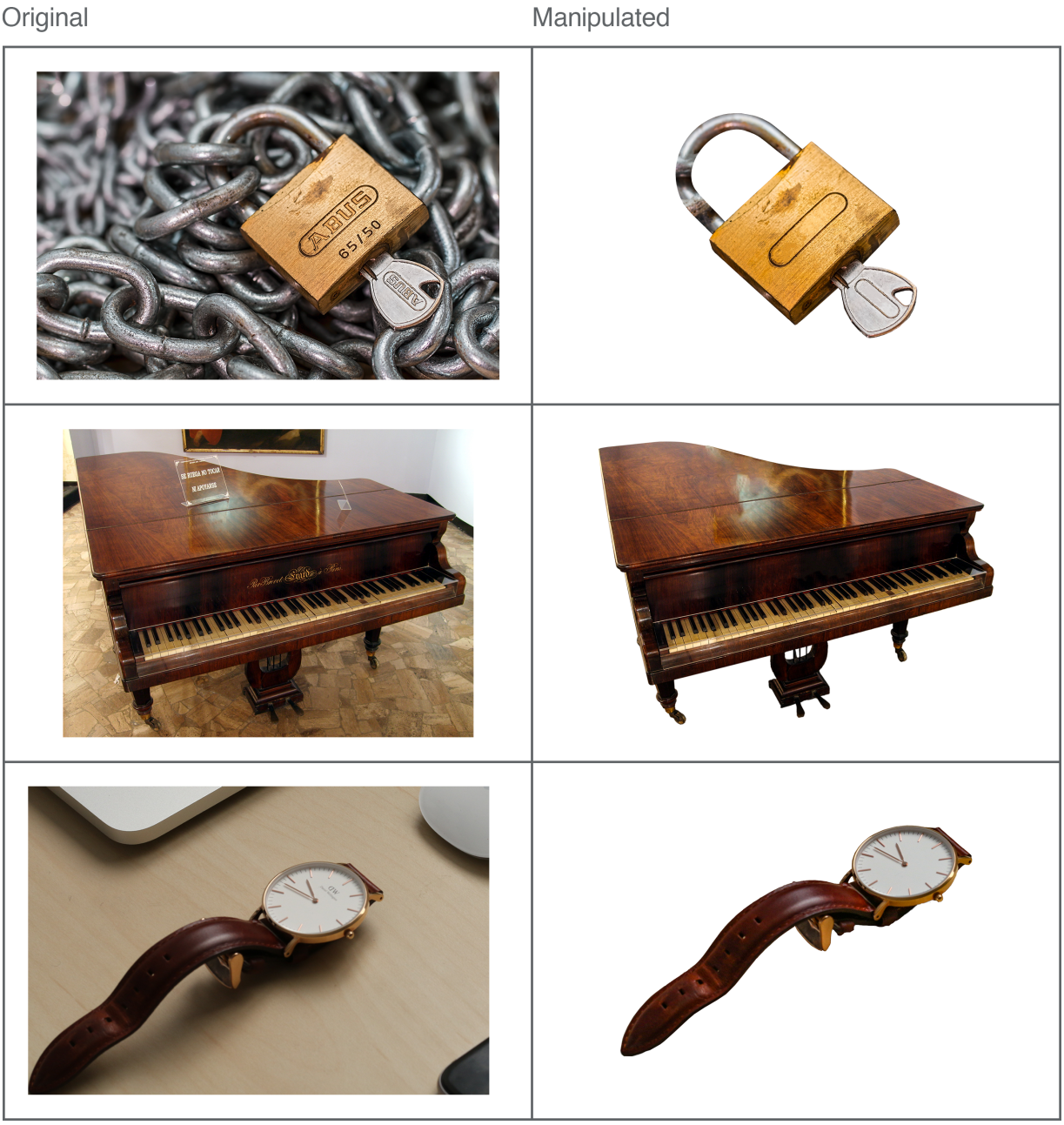


Figure 9: Examples of background and text removal in photograph items.

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[8] - a survey platform that allowed for straightforward collection of consent, demographics, and computer compatibility data, and ii) Pavlovia[9] - 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

[8]https://www.qualtrics.com/uk/
[9]https://pavlovia.org/

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.

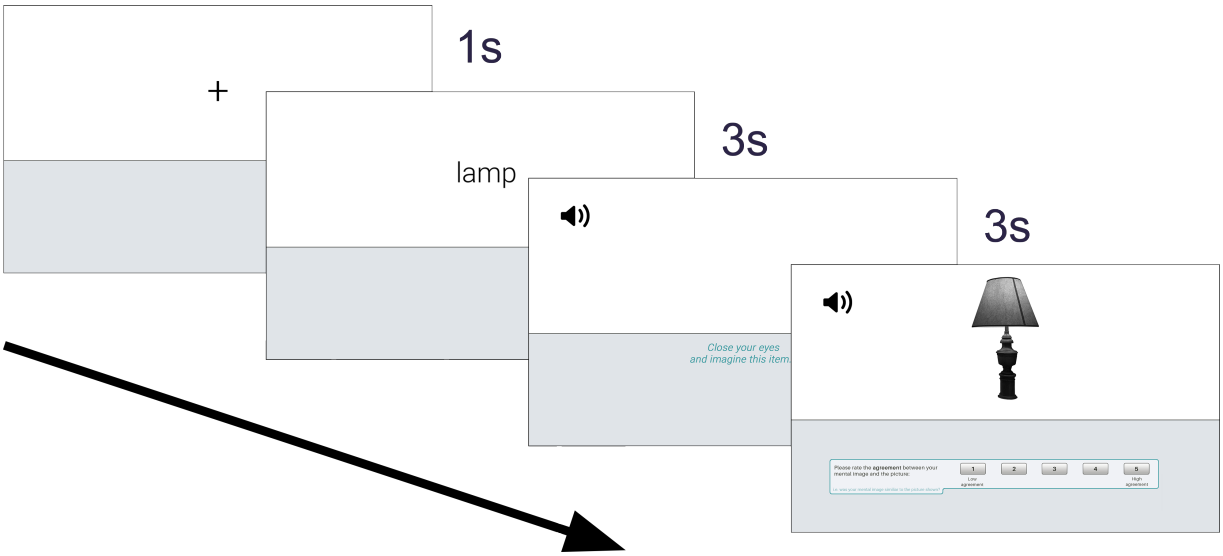


Figure 10: Data collection procedure for Mental Imagery Agreement responses.

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 congruent 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 shaded 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 poten-

tial 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 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 image, 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[10] (73.37%), where payment at the rate of £5/hr was given, and via the in-school research participation system[11] (15.38%), where they received course participation credits. A small number of participants were also recruited from Psychological Research on the Net[12] (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.

Materials A total of 126 innocuous, everyday objects (e.g. clock, rabbit, shoe) were presented across three individual stimuli formats: written words, shaded drawings, and photographs. The

[10] https://www.prolific.co/
[11] https://keelepsychology.sona-systems.com/
[12] https://psych.hanover.edu/research/exponnet.html

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)***

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 shaded 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.

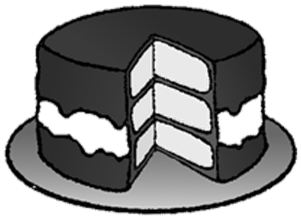
Words:	Drawings:	Photographs:
guitar		
mouse		
pumpkin		

Figure 11: Examples of the word, greyscale drawing, and greyscale photograph stimuli utilised in the current experiment.

Procedure Data was collected online using Gorilla[13] - 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

[13]https://gorilla.sc/

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, 1.7], and grey drawings (*M*= 0.76), *t*(160) = -8.04, *p* < .001; *d* = -0.63, 95% CI [-0.8, -0.47]. 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 [0.88, 1.31]). There were no significant interaction effects between stimuli format and response option condition, *F*(1.76, 280.25) = 0.58, *p* = .540, η_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, η_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 [-1.08, -0.79], and grey drawings (*M*= 0.09), *t*(160) = 5.59, *p* < .001; *d* = 0.44, 95% CI [0.31, 0.59]). 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 [-0.8, -0.45]. There were no significant interaction effects between stimuli format and response option condition, *F*(1.43, 226.74) = 1.17, *p* = .299, η_p^2 < .01.

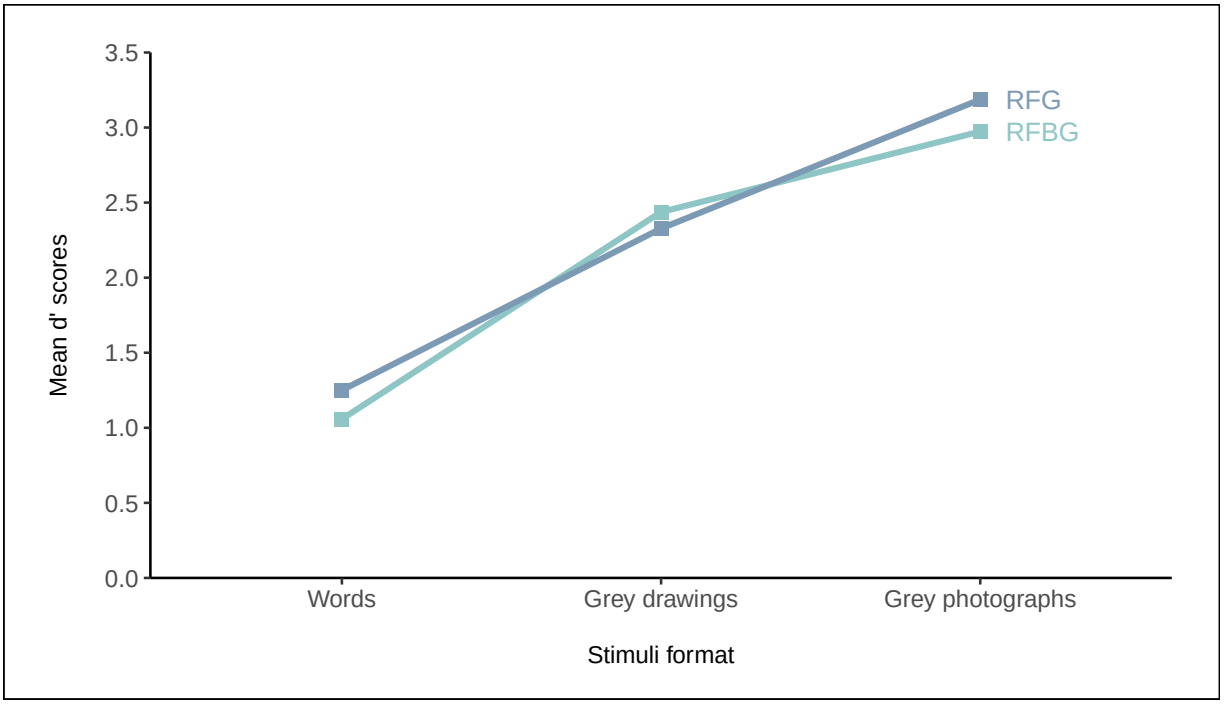


Figure 12: Interaction plot between stimuli format and response option for d' scores.

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, η_p^2 = .02 (see Figure 12). 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 13).

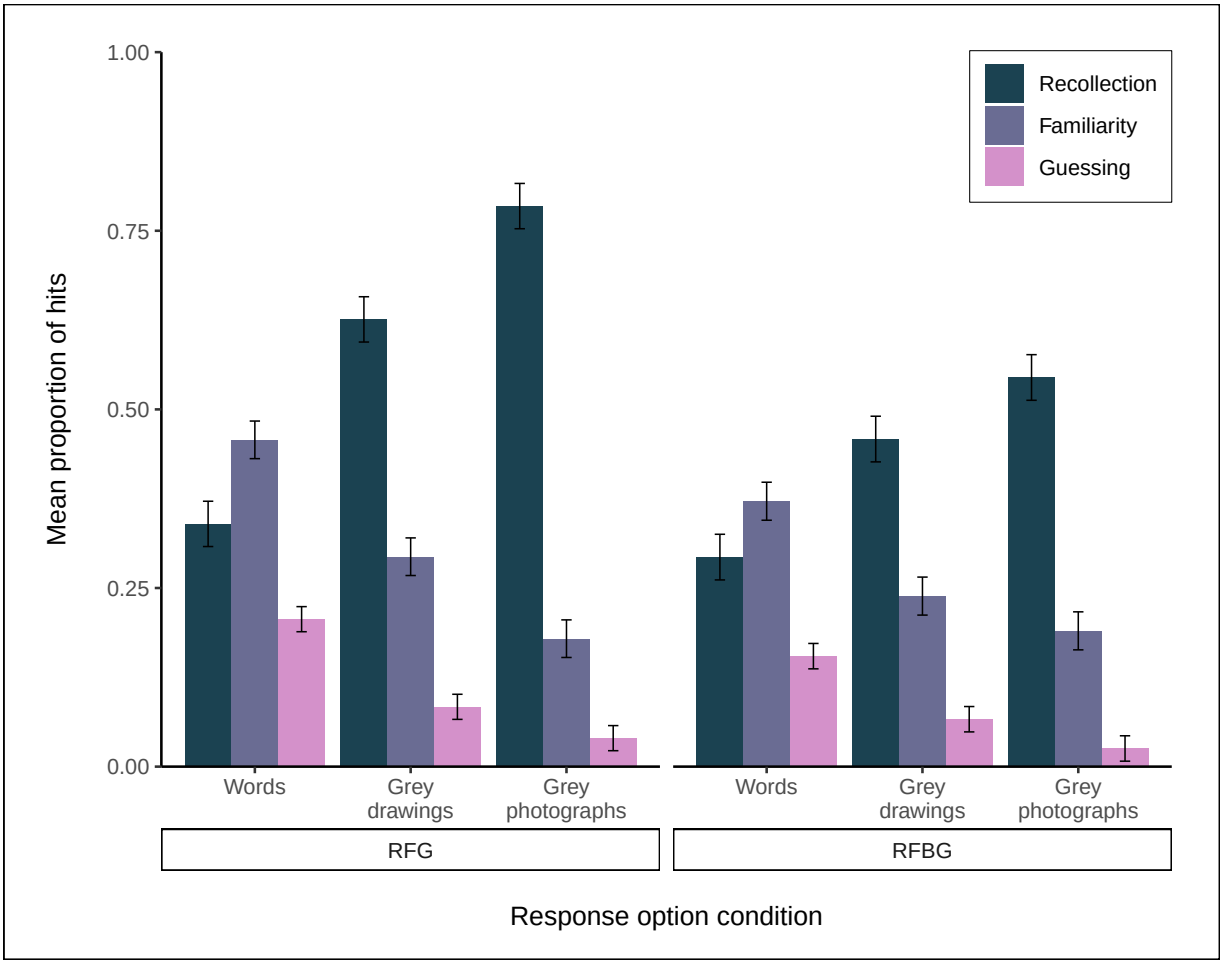


Figure 13: Proportion of hits assigned Recollection, Familiarity, and Guessing, by stimuli format and response option condition.

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 14).

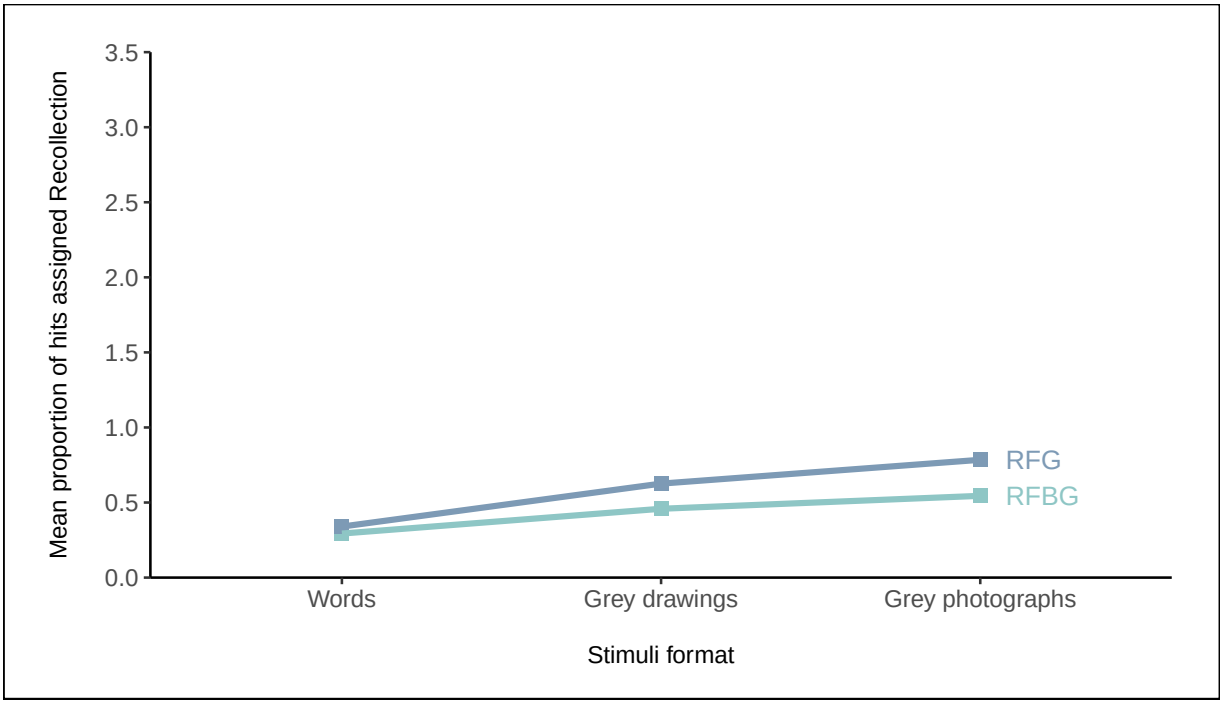


Figure 14: Interaction plot between stimuli format and response option for the mean proportion of hits assigned Recollection.

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 15).

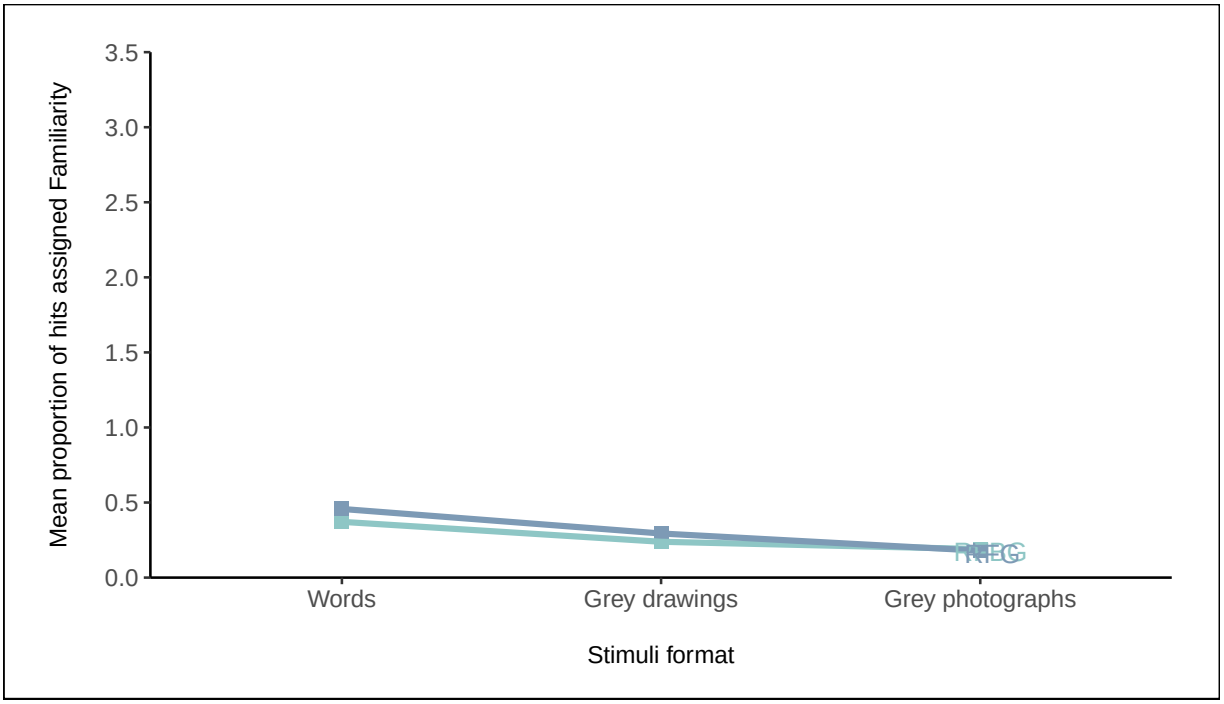


Figure 15: Interaction plot between stimuli format and response option for the mean proportion of hits assigned Familiarity

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, -0.64]) and grey drawings (*M*= 0.08; *t*(160) = 5.85, *p* < .001; *d* = 0.46, 95% CI [0.35, 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 [-0.73, -0.49]. 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 16).

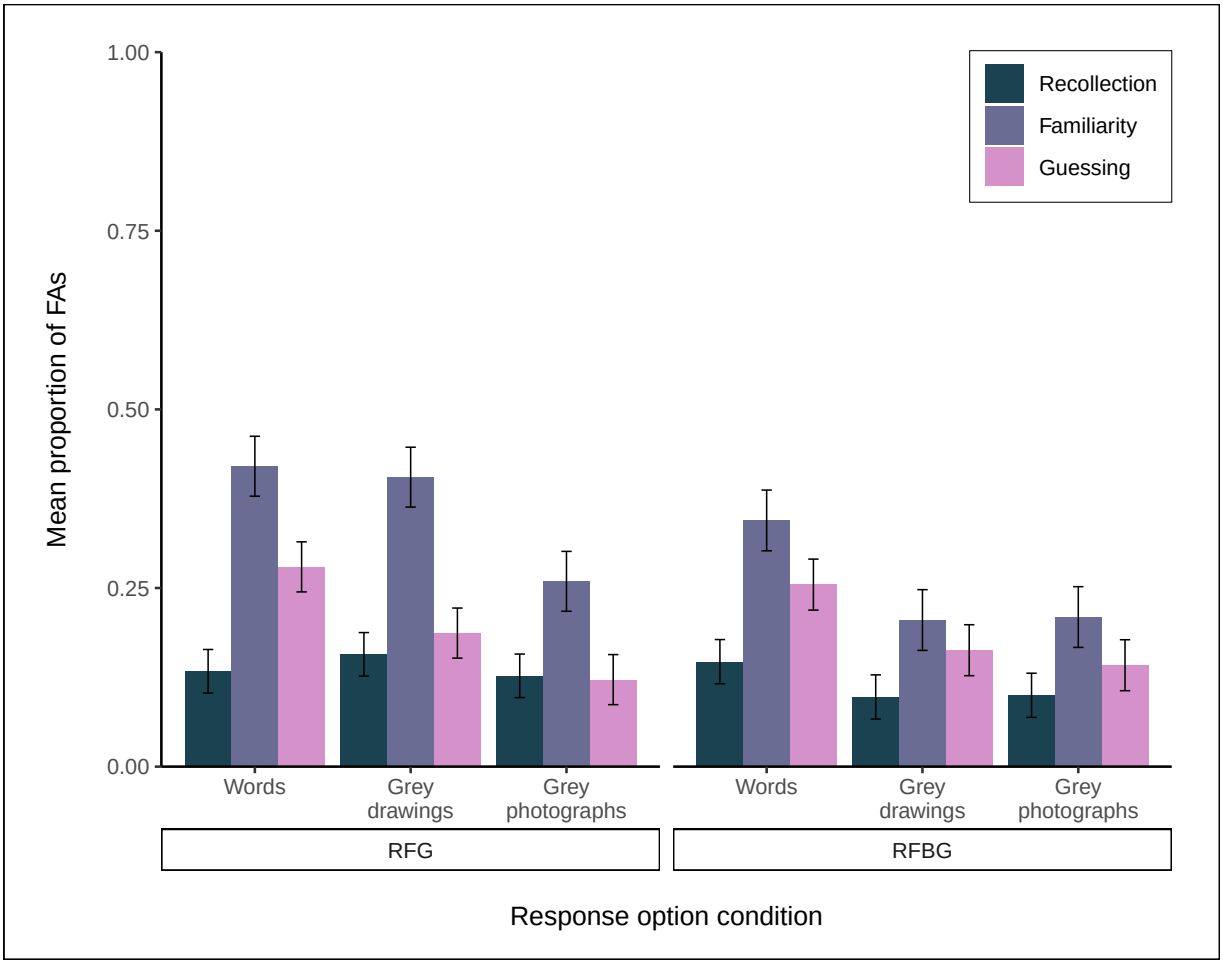


Figure 16: Proportion of FAs assigned Recollection, Familiarity, and Guessing, by stimuli format and response option condition.

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, -0.18]. 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 [-0.31, -0.0073]. 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, 0.32]. 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, -0.15]). 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 [-0.39, -0.07]. 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 [-0.03, 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 FAs***	$F(1, 159) = 6.30$, $p = .013$, $\eta_p^2 = .04$	*
Mean proportion: ***Guessing FAs***	$F(1, 159) = 0.08$, $p = .772$, $\eta_p^2 < .01$	

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 Chapter 3, the role of stimuli distinctiveness on recognition was explored by curating a new set of detailed, real-world object photographs, and comparing recognition performance toward these items with matched shaded drawn images and written word counterparts. Standard picture superiority effects (PSEs) were observed, with both types of picture producing enhanced recognition performance compared to word stimuli, however, a clear *photograph* superiority effect was also evident when comparisons were made between the photographs and drawings. Such results were attributed to the physical distinctiveness account of the PSE, whereby increased item-to-item variability results in items being remembered better than those with low variability. Studies have previously offered support for this hypothesis by attempting to manipulate the level of variability between items; Ensor et al. (2019) demonstrated how the PSE could be eliminated by i) increasing the variability between word stimuli, where each item was made more distinctive by manipulations to font, size, and colour, and ii) reducing the variability between picture stimuli, where shaded drawing items consisted of objects with highly similar shapes, size, and orientation. The current programme of research extended support for this hypothesis by increasing the variability between picture stimuli through the creation of a set of photographic stimuli where differences in detail, texture, and shading were intended to be more apparent across items than across drawings. While the *photograph* superiority effect (PhSE) demonstrated in the previous study highlights the importance of real world texture and shading in recognition memory, there are other inconsistencies across recognition memory studies with regard to distinctiveness - namely, whether images are presented to participants in greyscale or colour.

The extent to which colour affects the memorability of stimuli is unclear, though there is evidence to suggest this additional layer of information may increase the perceived distinctiveness of items. In a scene 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 real-world scene

photographs (e.g. train stations, city streets etc.). At test, participants were presented with two photographs side-by-side (one target + one similar lure) and asked to select the item shown during the study phase. The congruency of colour was manipulated, such that photographs were either presented in the same colour modality across both study and test (1. greyscale at study / greyscale at test; 2. colour at study / colour at test) or different (3. greyscale at study / colour at test; 4. colour at study / greyscale at test). Results showed that colour information produced superior recognition performance when presented congruently across encoding and retrieval. Interestingly, this colour benefit was *only* evident in the congruent condition; if colour images were paired with greyscale images - regardless of the order at encoding and test - recognition benefits were absent. A source judgement question at test, probing whether the colour format for each item was the same or different as at study showed similar performance across all four conditions (though performance was particularly poor for items presented in colour); such findings indicate the recognition benefits of the congruent colour condition were not a result of accessing memory for the colour information itself, but rather colour indirectly highlighted certain features within the photographs that were not otherwise noticed as prominently in greyscale, and thus the colour photographs were overall more distinctive as a result.

The current experiment aims to establish whether colour information facilitates successful recognition using object stimuli. The methodology of *Experiment 3* is precisely replicated - whereby recognition is tested across word, drawing, and photograph stimuli - though rather than greyscale items, the sole manipulation of the current study is the addition of colour to the two types of image stimuli. Initial analyses will determine whether the distinctiveness effects of the previous study are also evident when colour items are utilised, and in particular, whether a *photo* superiority effect remains, or if this finding is unique to greyscale stimuli. It is acknowledged that colour may not affect the two types of image in a uniform manner. The highly distinctiveness nature of the photograph items may plateau in such a way that performance can no longer be noticeably enhanced; any additional distinctiveness generated by colour information might show little benefit if performance is already sufficiently high. If this is the case, however, the effects of the colour manipulation should still be evident when examining recognition perfor-

mance for drawings, where there is more room for such benefits to become apparent. Following this initial analysis, exploratory comparisons will be made to compare the data from the current experiment with that from *Experiment 3*, in an effort to highlight any key differences introduced by the addition of colour. A secondary aim from previous experiments remains, whereby the role of different available response options (*Recollection*/*Familiarity*/*Guessing* or *Recollection*/*Familiarity*/*Both*/*Guessing*) will be examined to establish whether RFG response patterns are differentially affected following the introduction of colour. The following hypotheses are proposed:

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 image 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 image 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[14] (85.98%), where payment at the rate of £5/hr was given, and via the in-school research participation system[15] (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 ($\alpha = .05$, two-tailed), 79 subjects per group would be necessary ($N = 158$) in a 3x2 mixed ANOVA.

[14]https://www.prolific.co/
[15]https://keelepsychology.sona-systems.com/

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, shaded drawings, and photographs. Words and shaded 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) shaded 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 17 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 coun-

terbalancing 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[16]. 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).

[16]https://gorilla.sc/

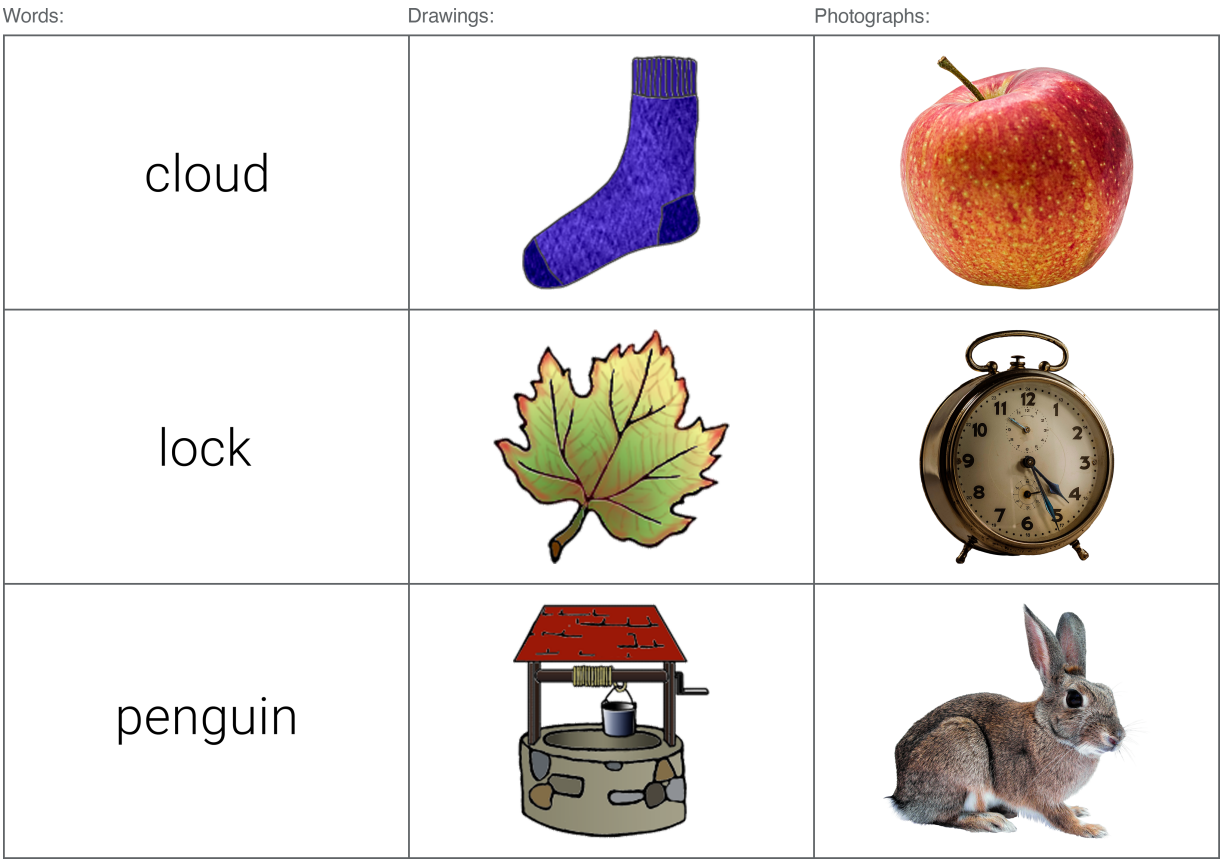


Figure 17: Examples of the word, colour drawing, and colour photograph stimuli utilised in the current experiment.

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 with *R* (R Core Team, 2020) using the *afex* (v0.28-0; Singmann et al., 2020) and *rstatix* packages (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.

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 ($F(2, 318) = 0.20$, $p = .817$, $\eta_p^2 < .01$).

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					**Experiment 4: Colour**				
		Hits	FAs	d'			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

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 [-1.02, -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 [0.37, 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, -0.89]. 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 18).

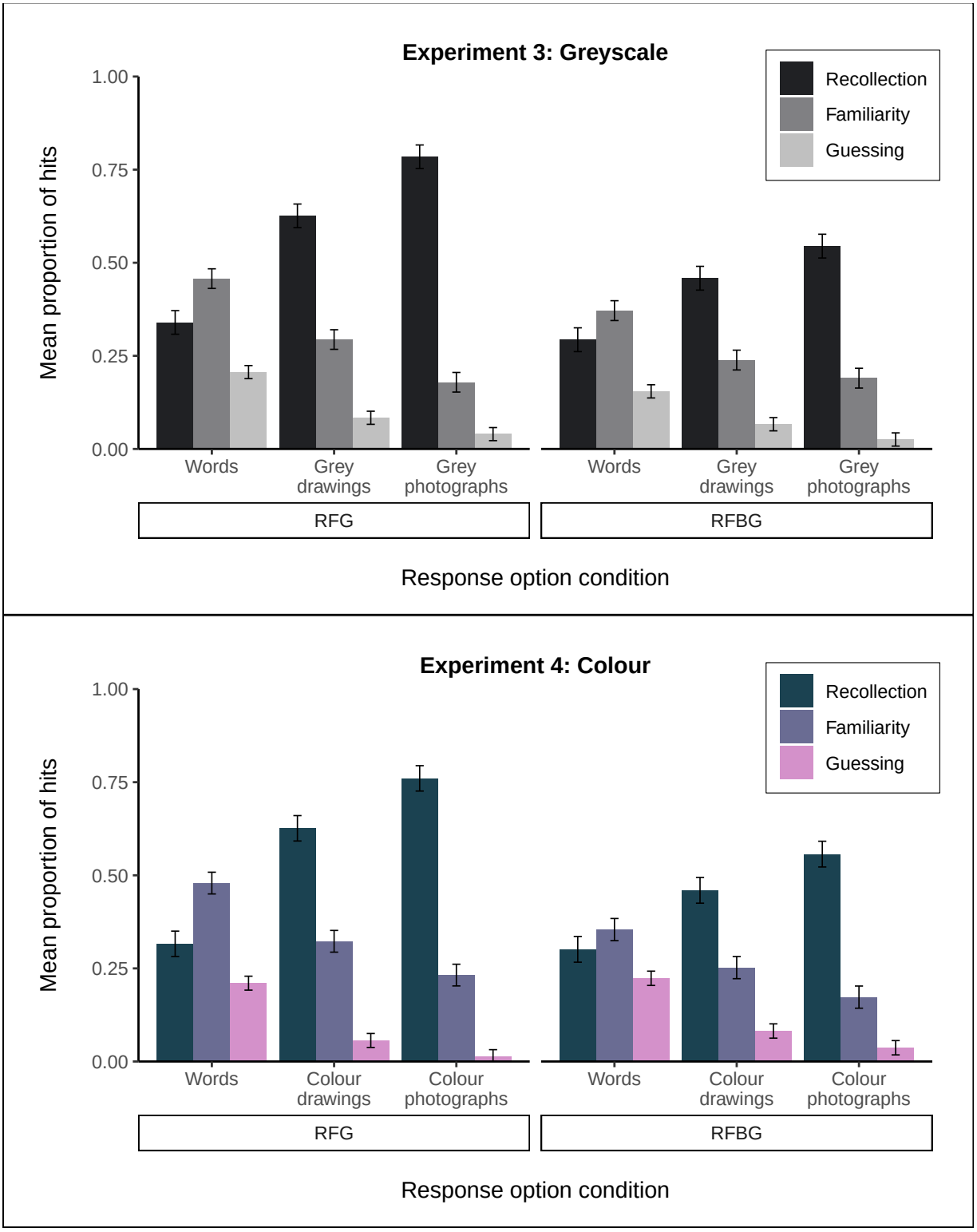


Figure 18: Proportion of hits assigned Recollection, Familiarity, and Guessing, by stimuli format and response option condition.

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).

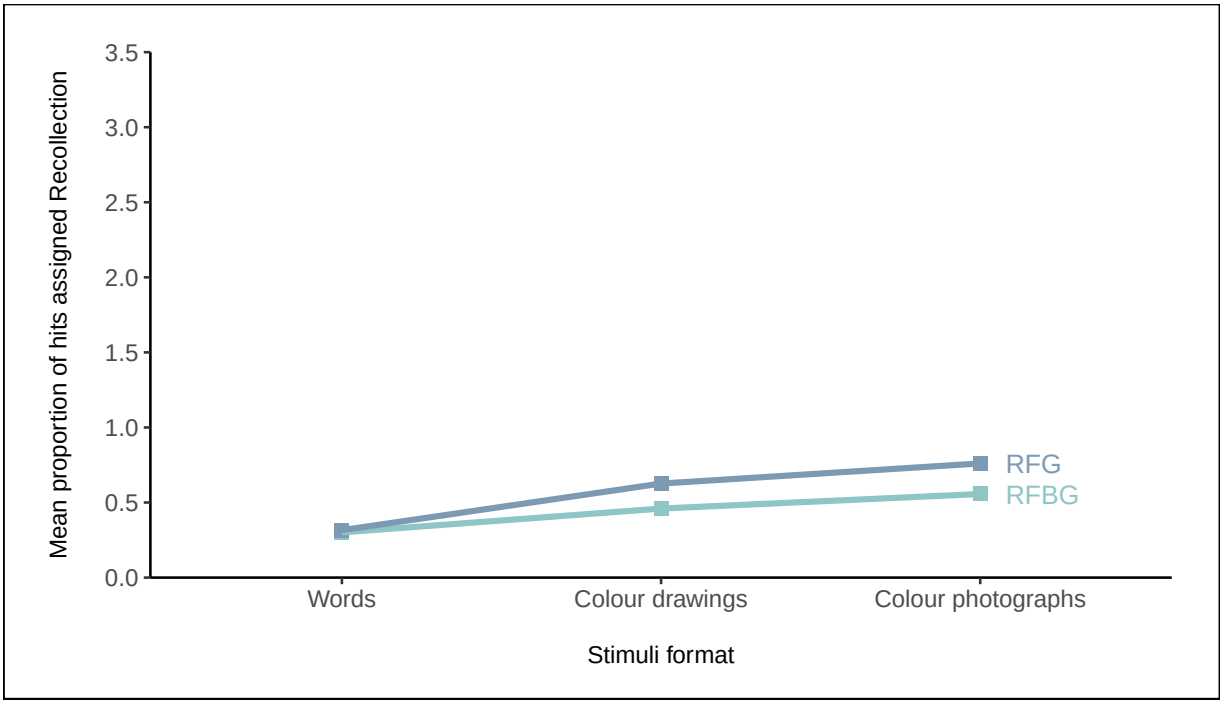


Figure 19: Interaction plot between stimuli format and response option for the mean proportion of hits assigned Recollection.

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 [-0.84, -0.49], and colour drawings (*M*= 0.29), *t*(160) = 5.97, *p* < .001; *d* = 0.47, 95% CI [0.31, 0.65]. 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, -0.3]. 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, 0.55]). 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.

Separate 3 (stimuli format: words, drawings, photographs) x 2 (response option condition: RFG-judgements, RFBG-judgements) mixed ANOVAs were also conducted on the mean proportion of FAs assigned Recollection, Familiarity, and Guessing (see Figure 20).

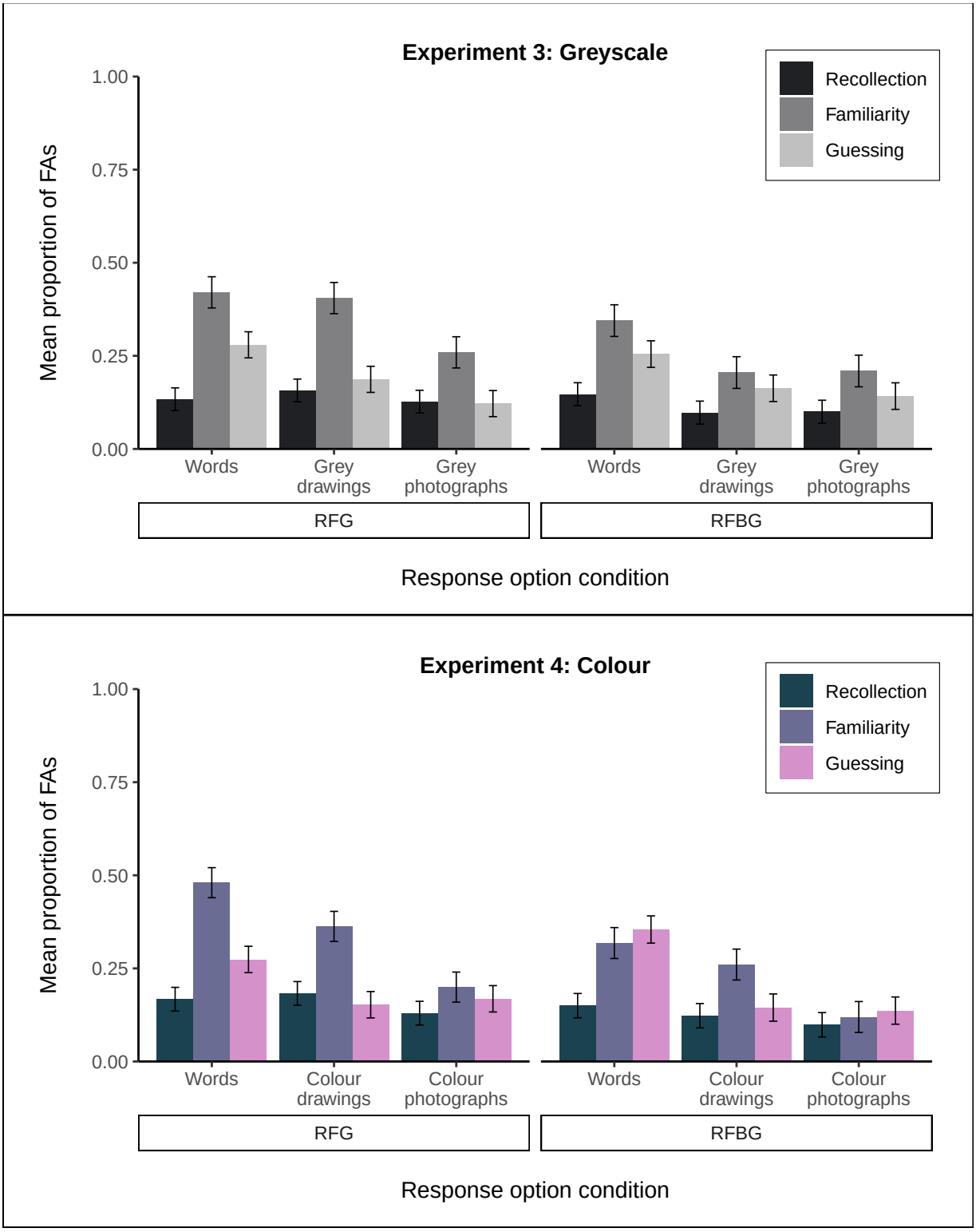


Figure 20: Proportion of FAs assigned Recollection, Familiarity, and Guessing, by stimuli format and response option condition.

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 [-0.69, -0.36]. 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, -0.05]. 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 [0.22, 0.53]. 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 [-0.53, -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 [-0.54, -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 [-0.16, 0.14]). 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 18 and Figure 20 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-

termine whether the addition of colour information had any differential effects compared to the findings of *Experiment 3*. The only main effect of response option was observed in the ANOVA for the mean proportion of FAs ($F(1, 159) = 1.03$, $p = .312$, $\eta_p^2 < .01$), whereby a higher proportion of FAs was evident in the RFG group ($M = 0.13$) compared to the RFBG group ($M = 0.1$), $t(457.34) = -2.43$, $p = .016$, $d = 0.22$. Such results do not support Hypothesis 3, where significant main effects of response option were expected in the ANOVAs for the mean proportion of hits and mean proportion of FAs assigned *Familiarity*. Similarly, the expected interactions for mean *d'* scores and proportion of hits assigned *Familiarity* were also absent. The only consistent finding across *Experiment 3* and the current experiment (with regard to response option) comes from the ANOVA for the mean proportion of hits assigned *Recollection* - where a significant interaction between stimuli format and response option was apparent in both experiments; in both experiments, the proportion of R hits did not differ across RFG and RFBG groups for word stimuli, whereas for drawings and photographs, the RFG group produced a significantly higher proportion of R hits than the RFBG group.

Discussion

The aim of the current experiment was to determine the effects of colour on recognition performance. In *Experiment 3*, recognition performance was found to improve in a linear manner as the intended distinctiveness of greyscale items increased, from words to shaded drawings to photographs. While standard PSEs were expected, the most notable finding of the previous study was the superiority of photographs compared to drawings - a result that offered support for the physical distinctiveness account of the PSE over dual-coding theory. In the current experiment, the physical distinctiveness of image stimuli was manipulated through the addition of colour; it was hypothesised that colour would provide an additional layer of distinctive information that would further facilitate successful recognition. When comparing the results of the current study with those obtained in *Experiment 3*, this hypothesis appears wholly unsupported; colour had little to no impact on the proportion of hits, FAs, and overall recognition accuracy.

The current results continue to be best explained by the physical distinctiveness account of the

PSE. Across each of the key variables, photographs demonstrated clear recognition benefits compared to drawings, replicating that found in *Experiment 3* and further supporting the notion that the current set of stimuli appear to effectively capture different 'levels' of stimuli distinctiveness. The additional real-world detail present in the photographs, and the increased variability between items as a result, provides memorial advantages over simple illustrated representations of objects. The key manipulation in the current study - the addition of colour - was unsuccessful in providing an additional layer of distinctiveness that would further enhance recognition for the objects compared to when they are presented in greyscale. Failure to discern any impact of colour does not seem attributable to ceiling effects, whereby performance is already so high that benefits are imperceivable, since responses toward drawings were equally unaffected by the manipulation.

(colour = no effect) Stróżak

######———————————————-

Chapter 5: Manipulating the distinctiveness of word stimuli

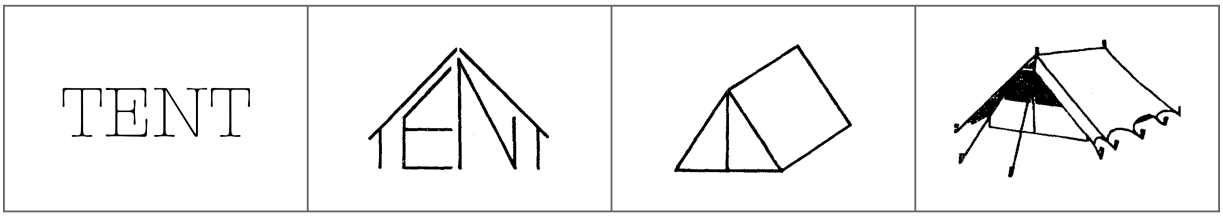
Background

In the previous two experiments, picture superiority effects (PSEs) were examined by systematically comparing recognition performance across three distinct stimuli formats. Word stimuli acted as a baseline to allow PSEs to emerge, while shaded drawings and photographs provided varying levels of detail upon which the magnitude of picture superiority could be compared. Not only were standard PSEs observed, but clear *photo* superiority effects (PhSE) were also apparent; in *Experiment 3* (greyscale pictures) and *Experiment 4* (colour pictures), photographs produced clear recognition benefits over drawings; in both studies, photograph items produced a significantly higher proportion of hits, lower proportion of FAs, and higher *d'* scores compared to drawings. Such findings are best understood by physical distinctiveness accounts of the PSE (Mintzer & Snodgrass, 1999). Increased item-to-item variability between the detailed photograph items acted to enhance their memorability in comparison to drawings, where this item-to-item variability was less pronounced. The findings from these experiments cannot readily be accounted for by the dual-coding theory (DCT; Paivio, Rogers, & Smythe, 1968; Paivio, 1991, 2007) of the PSE, where one would expect both drawings and photographs to similarly generate imagined and verbal representations of the stimulus, and thus their superiority over words should be the same (with no additional advantage for photographs). By manipulating the format of pictures, the previous experiments were able to offer convincing support for the physical distinctiveness account of the PSE, however, further investigation is warranted to determine which factors contribute most to item distinctiveness. For example, the presence of real world textures and details appears to play a key role in item memorability, establishing a photograph superiority compared to drawn illustrations. On the other hand, comparisons of the results from *Experiment 3* (greyscale pictures) and *Experiment 4* (colour pictures) reveal the addition of colour had very little impact on recognition performance, contrary to predictions. If the physical distinctiveness account of the PSE is the most persuasive, it should be possible to determine the point at which distinctiveness / the PSE breaks down, and identify the most salient factors.

Another approach that might narrow this pursuit is turning the focus to word stimuli, and determining how certain manipulations could inform our understandings of the PSE. Only a few studies to date have pursued the thread of manipulating word stimuli. Paivio et al. (1968) examined the effects of adding colour to word stimuli, under the assumption colour would enhance the physical "vividness" of items, and thus improve recall as a result. Participants studied one of four possible stimuli formats (greyscale words / colour words / greyscale drawings / colour drawings) before completing a written free recall test. Results showed there was no difference of this word manipulation, and black words were recalled at the same rate as those presented in colour, prompting a DCT interpretation of the PSE. In another free recall task, Peeck, Van Dam, & Uhlenbeck (1977) manipulated the memorability of word stimuli by utilising pictograms; individual letters that make up the name of an object were presented in such a way that the outline of the entire word resembled the object itself (for example, the letters: "T" "E" "N" "T" were drawn to resemble the shape of a tent when viewed as a whole (see Figure 21 for example items). Subjects studied four stimuli formats: i) regular words; ii) pictogram words; iii) regular pictures (simple line drawings); and iv) impoverished pictures (drawings made up of a very simple shapes, with no detail present), to compare the effects of increasing "perceptual richness" (regular words / pictograms) and decreasing (regular pictures / impoverished pictures). The pictograms were recalled significantly more than the regular word items - a result attributed to the importance of the physical characteristics of to-be-remembered stimuli. However, as recall for the regular pictures and impoverished pictures did not significantly differ, DCT interpretations were also discussed. According to DCT, all regular word stimuli have the ability to generate both verbal and imagery codes - though the generation of dual codes is assumed to happen more often in pictures, since it requires far less effort to internally label a picture (and generate an additional verbal code) than form a mental image from a written label (and generate an additional imagery code). Regardless, when dual codes are formed for words, it is assumed that the generated imagery code is indeed most likely a mental picture of the words referent (e.g. a mental image of the tent last used when camping), rather than any specific image of how the

physical letters appeared during presentation (e.g. the text was black and the letters were close together). This assumption makes the findings of Peeck et al. (1977) a little more difficult to interpret; given subjects are not constructing an image of the presentation of the word at study, it is not immediately clear why recall should be better for the pictograms over the regular words. The authors suggest such pictograms could be considered a unique class of picture stimuli, and while subjects were not generating an image of their physical appearance, the pictograms more readily aroused the expected mental images than did regular words (i.e. perhaps the resemblance to the referent object made conjuring a mental image less effortful).

Peeck, Van Dam, & Uhlenbeck (1977):



Haber & Myers (1982):

Ensor, Surprenant, & Neath (2019):

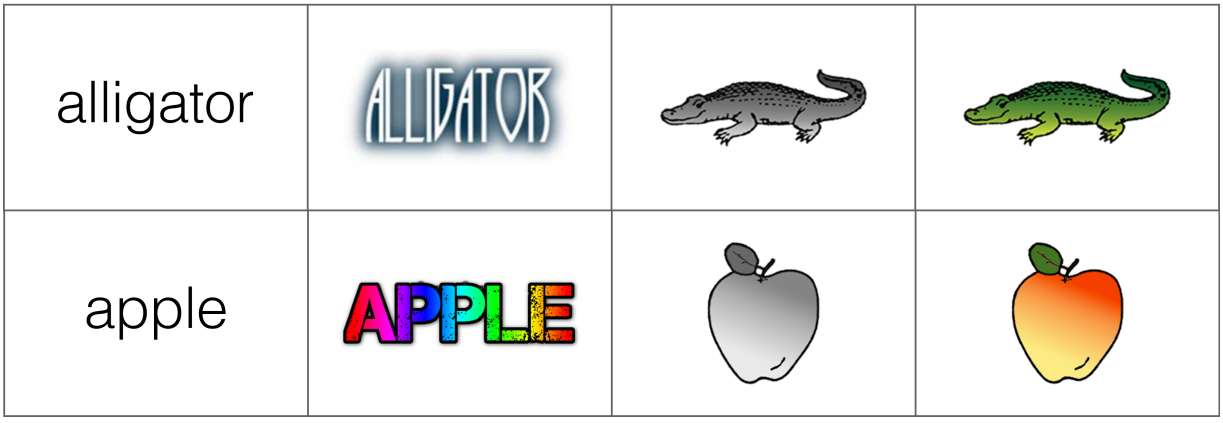


Figure 21: Example stimuli taken from Peeck et al. (1977), Haber & Myers (1982), and Ensor et al. (2019).

More recent work has examined the manipulation of word distinctiveness in recognition memory paradigms. In a effort to disentangle the prevalent theories of the PSE, Ensor et al. (2019) increased the item-to-item variability of word stimuli by utilising unique typefaces, a range of colour combinations, inconsistent sizes, and selective use of capitalisation (see Figure 21 for example items). In recognition memory paradigms, word stimuli are typically presented in a plain black, readable typeface (e.g. Arial, Times), with a focus on consistency between items (i.e. aside from the word itself, all other variables are often equal); the manipulation by Ensor et al. (2019) increased the variability across items under the assumption distinctiveness would be enhanced. Comparing recognition memory performance for such items against regular words and pictures produces some distinct and contrasting predictions. Operating under the physical distinctiveness account of the PSE, recognition should be enhanced for the more-distinct word stimuli in comparison to regular word stimuli; as such, when comparing these items with pictures, PSEs should be reduced or even eliminated as the differences in distinctiveness that drive the PSE are lessened. DCT, on the other hand, predicts no elimination of the PSE; manipulations to the colour, font, and capitalisation of word stimuli do nothing to negate the additional imagery code that is readily available for pictures. Thus, pictures will continue to show memorial advantages irrespective of how word stimuli - constrained by a single, verbal code - are presented. Recognition performance was compared in an old new paradigm for low distinctive words (black, plain font), high distinctive words (varying fonts, colours, size, and capitalisation), low distinctive pictures (greyscale drawings) and high distinctive pictures (colour drawings). Results showed that, while a standard PSE was observed between greyscale pictures and black words, there was no PSE between greyscale pictures and colour words, indicating the formats were similarly memorable as a result of the manipulation. The colour word stimuli were also shown to attenuate the PSE when comparisons were made with colour pictures, and when trial numbers were increased, the PSE was eliminated here too. Such findings offer convincing support for the physical distinctiveness account, in that increasing the distinctiveness of word stimuli acts to eliminate picture superiority. As such, it is difficult to reconcile these results in a dual-coding framework, which predicts any word items to consistently show diminished performance com-

pared to pictures.

The current study has two primary aims: i) replicate the elimination of the PSE demonstrated by Ensor et al. (2019); ii) examine whether the PhSE is also attenuated / eliminated when comparing photograph items with distinctive word stimuli. A secondary aim is to further characterise the role of colour in the previously established *photo* superiority effect (PhSE). Since previous analyses could only be performed between drawings and photographs of the same colour modality (i.e. i) drawings without vs. photographs without in *Experiment 3*; and ii) drawings with vs. photographs with in *Experiment 4*), the extent to which colour is involved in the PhSE, if at all, remains mostly unclear. In a scene recognition paradigm (utilising pictures of city streets, parks etc.), Suzuki & Takahashi (1997) demonstrated enhanced recognition performance when colour - as opposed to greyscale - items were shown at study and test. Interestingly, this benefit was only apparent when study and test conditions congruently presented items in colour; any recognition enhancements attributed to colour disappeared when blocks were incongruent (e.g. items presented in grey at study, and the same items presented in colour at test). Suzuki & Takahashi (1997) also found that subjects' memory for the colour mode they had seen items in at study was poor; in other words, participants could not remember well whether they had studied items in colour or greyscale, despite exhibiting better recognition for congruent colour items. Since the benefits of colour information could not be attributed to any conscious recall of the actual colours in the pictures, it was instead hypothesised colour information acts to indirectly highlight certain features in a picture - i.e. details that were not otherwise noticed as prominently in greyscale - and thus increases their distinctiveness (and memorability) as a result. It is unclear whether colour information provides similar benefits in object recognition, or if this effect applies exclusively to real-world photographs of scenes (presumably comprised of a multitude of items).

If colour enhancements akin to those demonstrated by Suzuki & Takahashi (1997) are indeed also present in object recognition, it is possible that the PhSE could be eliminated as a result. In the same way the PSE is eliminated between words and drawings when word stimuli are made more distinctive (Ensor et al., 2019), the PhSE may too be eliminated when comparisons are made between colour drawings (whose distinctiveness is presumably somewhat enhanced by

colour) and greyscale photographs (whose real-world details and textures increase their distinctiveness in relation to drawings, but colour enhancements are absent). Indeed, there is some evidence to support this notion; the normative data obtained by Rossion & Pourtois (2004) for their revised set of Snodgrass & Vanderwart (1980) object drawings revealed clear object naming benefits when items were presented in colour. When subjects were asked to provide single-word, unambiguous labels toward a number of illustrated object stimuli, naming agreement scores across participants were significantly higher for items presented in colour, as opposed to illustrations shown in greyscale (and simple un-shaded outlines). Reaction times were also significantly faster for the colour items compared to the other types. While these effects were most pronounced when items had a diagnostic colour (e.g. fruits), items depicting man-made objects without diagnostic colours also showed such benefits. The researchers concluded colour information clearly plays a role in facilitating basic everyday item recognition, however, similar benefits are not consistently supported in other experimental paradigms involving memory. In a free recall task, Paivio et al. (1968) compared PSE effects for greyscale and colour items and found no effect of colour - in fact, recall was generally poorer for the colour items. In a recognition memory paradigm, Ensor et al. (2019) demonstrated some minor differential effects of colour; in their first experiment, an elimination of the PSE was demonstrated when their distinctive words were compared with *greyscale* drawings (suggesting both formats exhibited similar levels of distinctiveness), but the PSE was only attenuated when their distinctive words were compared with *colour* drawings (suggesting the colour drawings were presumably more distinct than the greyscale drawings). However, this minor benefit of colour information disappeared when the number of study and test trials were doubled, as the PSE was then eliminated for greyscale *and* colour drawings.

It may be that any distinctiveness benefits afforded by colour are particularly susceptible to methodological circumstances, or indeed are simply negligible - a notion supported thus far in the current programme of research. Visual inspection of the results from *Experiment 3* (greyscale pictures) and *Experiment 4* (colour pictures) suggest no enhancements as a result of colour information with both presentation types displaying highly similar PSE and PhSE magnitudes,

proportions of hits, FAs, and mean *d'* scores, and patterns of RFG responding. As photographs were consistently recognised better than drawings, however, we cannot interpret the lack of a colour enhancement as evidence of DCT in the same was as Paivio et al. (1968). Rather, the findings instead characterise colour as a factor that does not contribute to the physical distinctiveness of an object, contrary to what we might assume. An elimination of the PhSE therefore seems unlikely; if colour information shows no discernible benefit in recognition memory paradigms, performance toward greyscale photographs will likely remain superior in comparison to colour drawings.

Based on results of the previous experiments, and the research discussed above, the following hypotheses are proposed:

1. Recognition will be enhanced for word stimuli characterised by distinct variations to colour, font, and capitalisation across items, compared to those consistently presented in the same black typeface. For colour words, this enhancement is expected to manifest as the following compared to greyscale words:
    - i) higher overall *d'* scores;
    - ii) higher proportion of correct hits;
    - iii) lower proportion of false alarms.

Such findings will establish a successful replication of the Ensor et al. (2019) manipulation, and allow for the examination of PSE elimination effects in the current paradigm.

2. A picture superiority effect (PSE) will be evident, whereby recognition for either picture type (drawings / photographs) will be enhanced in comparison to greyscale words. Drawings and photographs are expected to show the same performance benefits as above compared to greyscale words.
3. PSEs will be eliminated when recognition is compared for colour words and shaded drawings:

- Elimination of the PSE when colour words are compared with greyscale drawings;
- Attenuation or elimination of the PSE when colour words are compared with colour drawings.

Ensor et al. (2019) observed differences in the comparison between colour words and colour drawings, depending upon the number of trials presented to participants (attenuation: 40 trials at study + 80 at test; elimination: 80 trials at study + 160 at test). As the current study is situated between these (60 trials at study + 120 at test), either outcome is expected.

4. A *photograph* superiority effect (PhSE) will also be apparent:
    - Greyscale photographs will produce better recognition than greyscale drawings (as in *Experiment 3*);
    - Colour photographs will produce better recognition than colour drawings (as in *Experiment 4*);
    - Greyscale photographs will produce better recognition than colour drawings (the PhSE will not be eliminated when drawings are shown in colour).
5. PSEs between colour words and photographs will persist:
    - While the recognition benefits afforded to colour words may eliminate or attenuate the PSE when comparisons are made with drawing stimuli, this will not be sufficient to eliminate the effect when comparisons are made with photographs.

Experiment: Manipulating word distinctiveness

Method

Participants

A total of 173 participants completed the online experiment, all between the ages 18-59 years (see Table 11 for a breakdown of the age/gender of the current sample). The majority of participants reported English as being their first language (87.86%). Participants were primarily

recruited from the in-house research participation system[17] (91.33%), though some were also sourced from the voluntary participation website Prolific Academic[18] (4.05%), where payment at the rate of £5/hr was given, and from other online social media (e.g. Facebook; 4.62%). Sample size was calculated a-priori using G*Power[19] (Faul, Erdfelder, Buchner, & Lang, 2009) to detect a small effect size of Cohen's f = 0.1 with 95% power (α = .05, two-tailed), 166 subjects would be necessary in a one-way repeated measures ANOVA.

Table 11: Gender and age (*SD*) of the current sample.

Gender	N	Age	(SD)
Female	*139*	20.63	*(6.09)*
Male	*31*	21.26	*(7.97)*
Female/Non-binary	*1*	19.00	*(0)*
FTM Transgender	*1*	23.00	*(0)*
Unspecified	*1*	32.00	*(0)*
Total	***173***	**20.81**	***(6.46)***

Materials

Three stimuli formats were utilised in the current experiment - words, drawings, and photographs - each of which had colour and greyscale variations (see Figure 22 for example stimuli). Shaded drawings were sourced from Rossion and Pourtois (2004) and consisted of shaded illustrations of innocuous, everyday objects (e.g. clock, rabbit, shoe). The photographs were those curated in *Experiment 2* - high quality photographs that similarly depicted the same everyday objects as those found in the Rossion & Pourtois (2004) drawings, with the objects of interest isolated from their original backgrounds. Word stimuli consisted of the written-word labels for each of the objects; the greyscale word stimuli consisted of a simple sans-serif typeface (Roboto light; 54pt) presented in plain black ink, while the colour variations contained manipulations to font, size, colour, and capitalisation. In their experiment, Ensor et al. (2019) made comparisons to

[17] https://keelepsychology.sona-systems.com/
[18] https://www.prolific.co/
[19] https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower

the same set of drawings used in the current study; as such, many of their colour word items could be re-purposed for use in the current study. However, as not *all* of the current pool of items were represented in this previous experiment, new colour word items were also created using the same design source (Cool Text Graphics Generator[20]). Every effort was made to match the font styles of the newly created words (i.e. objects that had not been present in the Ensor et al. (2019) study) with those that had been utilised previously for different objects. All stimuli were separately placed on a 500x500px blank canvas using Adobe Photoshop 2021 (22.0.0 Release); some items were constrained to this canvas size by their height, while others were constrained by their width, depending on the dimensions of the object. All items were exported as .pngs files for presentation by the online survey platform, where they were presented at their actual size (i.e. without scaling) to ensure consistency across participants.

Design

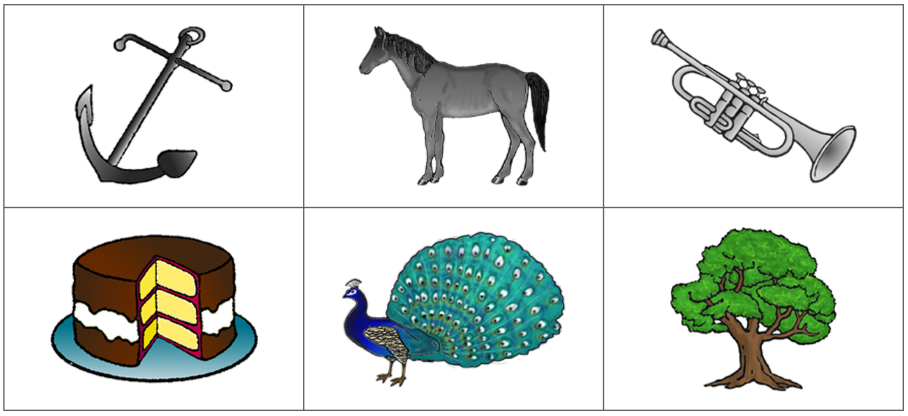
A one-way repeated measures design was utilised, consisting of a within-subjects factor of stimuli format with six levels: i) greyscale words; ii) colour words; iii) greyscale drawings; iv) colour drawings; v) greyscale photographs; vi) colour photographs. Participants completed a single study block (60 items) and a single test block (120 items), both of which were comprised of an equal number of items for each of the six stimuli formats. The particular format objects were presented in (for example, whether the item “penguin” was presented as a colour word or as a greyscale photograph) was counterbalanced across participants via blocked randomisation. All study / test counterbalancing routes were of equal length.

[20] https://cooltext.com/

Word stimuli:

flag	penguin	whistle
CANDLE	piano	tiger

Drawing stimuli:



Photograph stimuli:

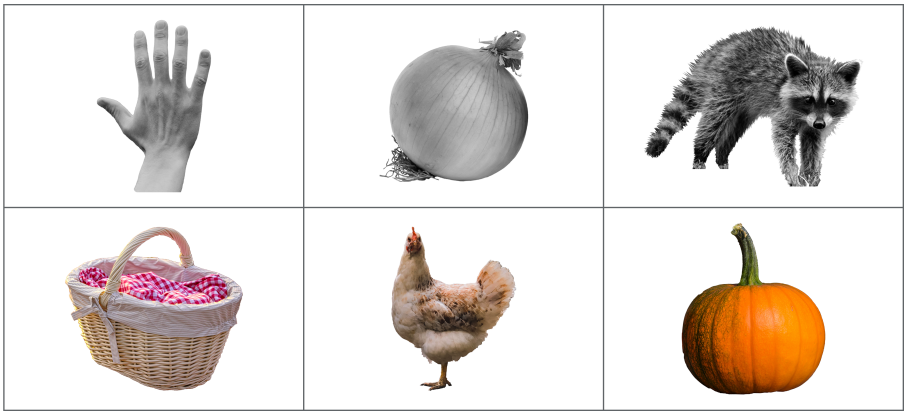


Figure 22: Examples of the six stimuli formats utilised in the current experiment: greyscale and colour variations of words, drawings, and photographs.

Procedure

As in previous experiments, data was collected online using the Gorilla[21] experiment platform. Participants completed three self-paced phases: i) study phase, ii) distractor task, and iii) recognition test. In the study phase, subjects were instructed to learn each of the items presented in preparation for a later memory test. Word, drawing, and photograph items were displayed one-at-a-time at random on-screen, and for each subjects required to report whether the current format was a word, drawing, or photograph. No distinction was made between greyscale and colour items in the wording of these response options (i.e. participants should respond "Word" to greyscale words *and* colour words), since this encoding judgement was utilised simply to ensure subjects directed their attention toward the to-be-remembered stimuli. Following the study phase, subjects completed a distractor task consisting of simple multiple choice mathematical questions (e.g. 6 x 4 = ?), before finally being presented with the recognition test. Subjects were again presented with items from each of the six stimuli formats. shown one-at-a-time at random on-screen. While half of the test items had been presented earlier in the study phase, the other half were new and had not been studied. First, subjects were first required to make *Old*/*New* judgements based on whether they recognised each of the items. If subjects experienced no recognition toward an item (and thus selected *New*) they would simply continue to the next item. If subjects did experience recognition (and thus selected *Old*), they were presented with a follow-up screen probing how they had arrived at this decision; subjects were asked to report whether they had recognised the item via *Recollection*, *Familiarity*, or were simply *Guessing* that it was old. Within participants, the format items were presented in remained the same across study and test (e.g. if the item "penguin" was shown in grey word format at study, it was also shown as a grey word at test).

Data processing

Subjects were excluded from analyses on the basis of two key criteria; 1) less than 90% accuracy during the encoding task (N= 14; where they were asked to report whether each item was shown

[21] https://gorilla.sc/

as a word, drawing, or photograph); 2) extreme z-scores (*N*= 1; i.e. those presenting z-scores of +/- 3 for total hits, total FAs, or overall recognition [hits minus FAs]).The subjects to meet these criteria were considered outliers and excluded from analysis, leaving a total of 158 datasets.

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 *Recollection*, *Familiarity*, and *Guessing*. Comparisons are made in relation to the first two experiments of Ensor et al. (2019), the methodologies of which were identical to one another, apart from the number of trials involved (one experiment consisting of 40 trials at study and 80 at test, and the other consisting of 80 trials at study and 160 at test). With regard to trial numbers, the current study sits between the two Ensor et al. (2019) experiments, with 60 trials at study and 120 at test. All analyses were conducted with *R* (R Core Team, 2020) using the *afex* (v0.28-0; Singmann et al., 2020) and *rstatix* packages (v0.6.0; Kassambara, 2020).

Results

To address each of the hypotheses, a series of one-way repeated measures analyses of variance (ANOVA) - using six factors (grey & distinctive words, grey & colour drawings, grey & colour photographs) - were conducted on the mean proportions of hits, mean proportion of FAs, and mean *d'*scores. Sphericity was assessed using Mauchly's Test; when the assumption of sphericity was violated, Greenhouse-Geisser corrections were applied.

To compare the results of the current study with previous findings, the raw data from the first two experiments of Ensor et al. (2019) was re-analysed in order to calculate statistics that were not reported in their paper.

Picture superiority (PSE) Standard PSEs were apparent in the ANOVAs on the mean proportion of hits, mean proportion of FAs, and mean *d'* scores. Grey words (M = 0.58) showed a significantly lower proportion of hits compared to all types of picture: grey drawings (M = 0.75), $t(157) = 9.04$, $p < .001$, $d = 0.72$; grey photographs (M = 0.91), $t(157) = 17.42$, $p < .001$, d = 1.39; colour drawings (M = 0.78), $t(157) = 10.37$, $p < .001$, $d = 0.83$; and colour photographs (M = 0.91), $t(157) = 17.53$, $p < .001$, $d = 1.39$. Grey words (M = 0.18) also showed a significantly higher proportion of FAs compared to all types of picture: grey drawings (M = 0.08), $t(157)$ = -6.60, $p < .001$, d = -0.52; grey photographs (M = 0.04), $t(157) = -9.31$, $p < .001$, $d = -0.74$; colour drawings (M = 0.05), $t(157) = -8.45$, $p < .001$, $d = -0.67$; and colour photographs (M = 0.02), $t(157) = -10.72$, $p < .001$, $d = -0.85$. Finally, mean *d'* scores were also significantly lower for grey words (M = 1.49) compared to all types of picture: grey drawings (M = 2.65), $t(157)$ = 12.37, $p < .001$, $d = 0.98$; grey photographs (M = 3.62), $t(157) = 24.28$, $p < .001$, $d = 1.93$; colour drawings (M = 2.89), $t(157) = 14.80$, $p < .001$, $d = 1.18$; and colour photographs (M = 3.83), $t(157) = 24.61$, $p < .001$, $d = 1.96$.

Photograph superiority (PhSE) Examining greyscale items only, a PhSE was apparent across each of the variables (as demonstrated previously in *Experiment 3*). Grey photographs (M = 0.91) showed a significantly higher proportion of hits compared to grey drawings (M = 0.75), $t(157) = -10.75$, $p < .001$, $d = -0.86$; grey photographs (M = 0.04) showed a significantly lower proportion of FAs compared to grey drawings (M = 0.08), $t(157) = 3.86$, $p < .001$, d = 0.31; and grey photographs (M = 3.62) showed significantly higher mean *d'* scores than grey drawings (M = 2.65), $t(157) = -11.37$, $p < .001$, $d = -0.90$. Examining colour items only, a PhSE was similarly apparent across each of the variables (as demonstrated previously in *Experiment 4*). Colour photographs (M = 0.91) showed a significantly higher proportion of hits compared to colour drawings (M = 0.78), $t(157) = -10.03$, $p < .001$, $d = -0.80$; colour photographs (M = 0.02) showed a significantly lower proportion of FAs compared to colour drawings (M = 0.05), $t(157)$ = 4.57, $p < .001$, d = 0.36; and colour photographs (M = 3.83) showed significantly higher mean *d'* scores than colour drawings (M = 2.89), $t(157) = -10.81$, $p < .001$, $d = -0.86$.

To examine the role of colour vs. greyscale in the PhSE, colour drawings were compared with greyscale photographs to examine whether the PhSE was eliminated. Colour drawings (M = 0.78) showed a significantly lower proportion of hits compared to grey photographs (M = 0.91), $t(157) = -9.62$, $p < .001$, $d = -0.77$. Colour drawings also (M = 2.89) showed a significantly lower mean *d'* scores compared to grey photographs (M = 3.62), $t(157) = -7.51$, $p < .001$, d = -0.60. However, the PhSE was eliminated when comparing colour drawings (M = 0.05) and grey photographs (M = 0.04) on the mean proportion of FAs, where there was no difference between the formats, $t(157) = 1.74$, $p = .083$, $d = 0.14$.

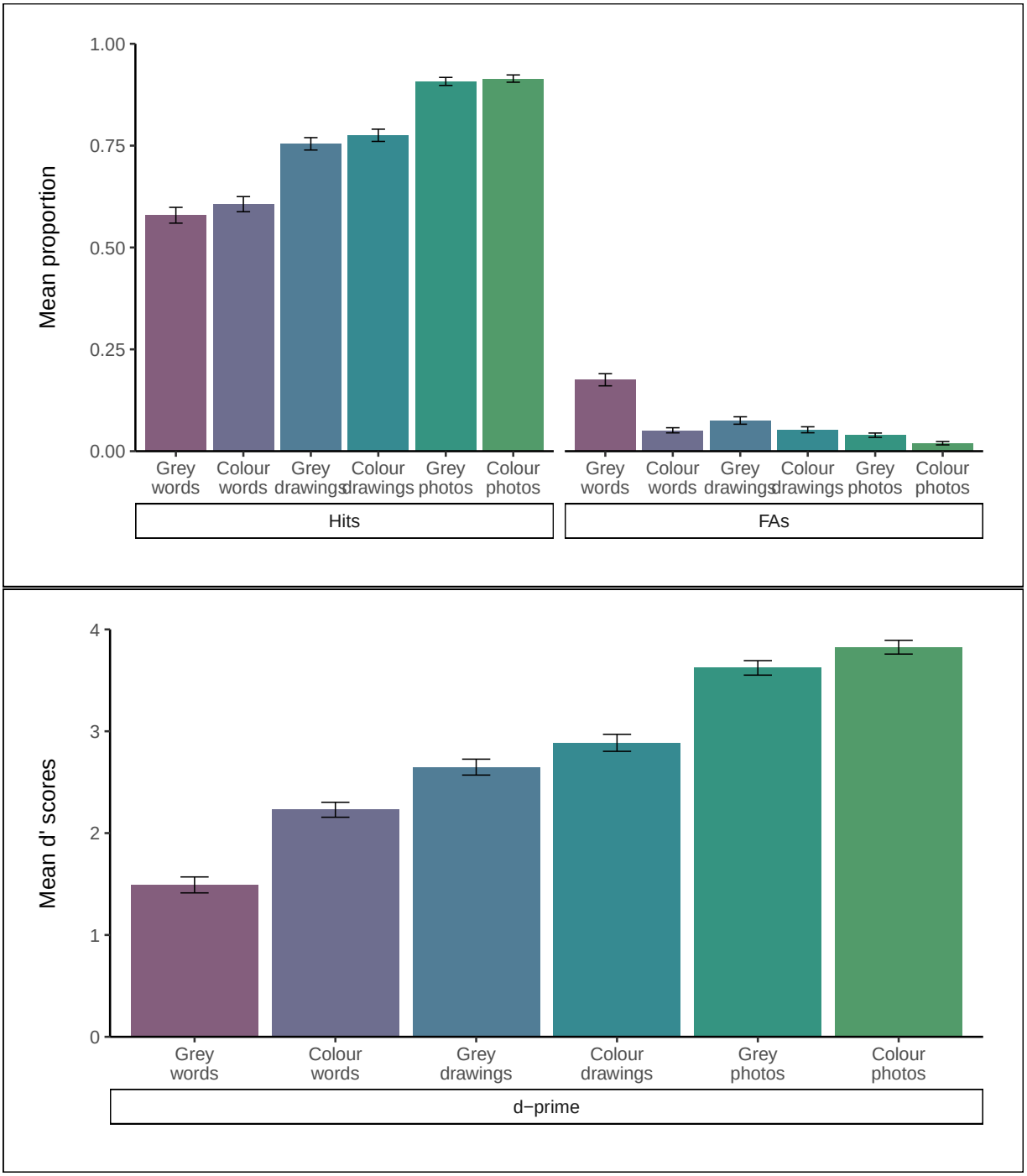


Figure 23: Mean proportion of hits, FAs, and d' scores by stimuli format.

Manipulation check Throughout each of the conducted ANOVAs, there was consistently a significant main effect of stimuli format: mean proportion of hits [$F(3.95, 619.49) = 137.94$, $p < .001$, $\eta_p^2 = .47$], mean proportion of FAs [$F(2.83, 443.69) = 52.71$, $p < .001$, $\eta_p^2 = .25$], mean *d'*-scores [$F(4.74, 744.81) = 190.22$, $p < .001$, $\eta_p^2 = .55$]. Planned comparisons following these significant results, however, did not always demonstrate that the distinctive word stimuli were indeed any more distinctive than the regular word stimuli. Primarily, there was no difference in the proportion of overall hits between distinctive ($M = 0.61$) and regular words ($M = 0.58$), $t(157) = 1.26$, $p = .211$, $d = 0.10$ - recognition performance (i.e. the number of items subjects were able to correctly identify) was the same regardless of the format word stimuli were presented in. Subjects did, however, make significantly fewer FAs as a result of the manipulation (distinctive words $M = 0.05$, regular words $M = 0.18$; $t(157) = -8.70$, $p < .001$, $d = -0.69$), offering some support to the manipulation, as subjects were better able to correctly reject lure items when they were presented in a distinctive format, rather than regular format, highlighting some performance benefits. Furthermore, distinctive words ($M = 2.23$) also showed significantly higher *d'* scores than regular words ($M = 1.49$), $t(157) = 7.96$, $p < .001$, $d = 0.63$; subjects were better able to accurately discriminate between hits and FAs when responding to the distinctive words over regular words.

```
## Warning: Use of `exp5__figure__rfg.hits.data$ymin` is discouraged. Use `ymin`
## instead.
```

```
## Warning: Use of `exp5__figure__rfg.hits.data$ymax` is discouraged. Use `ymax`
## instead.
```

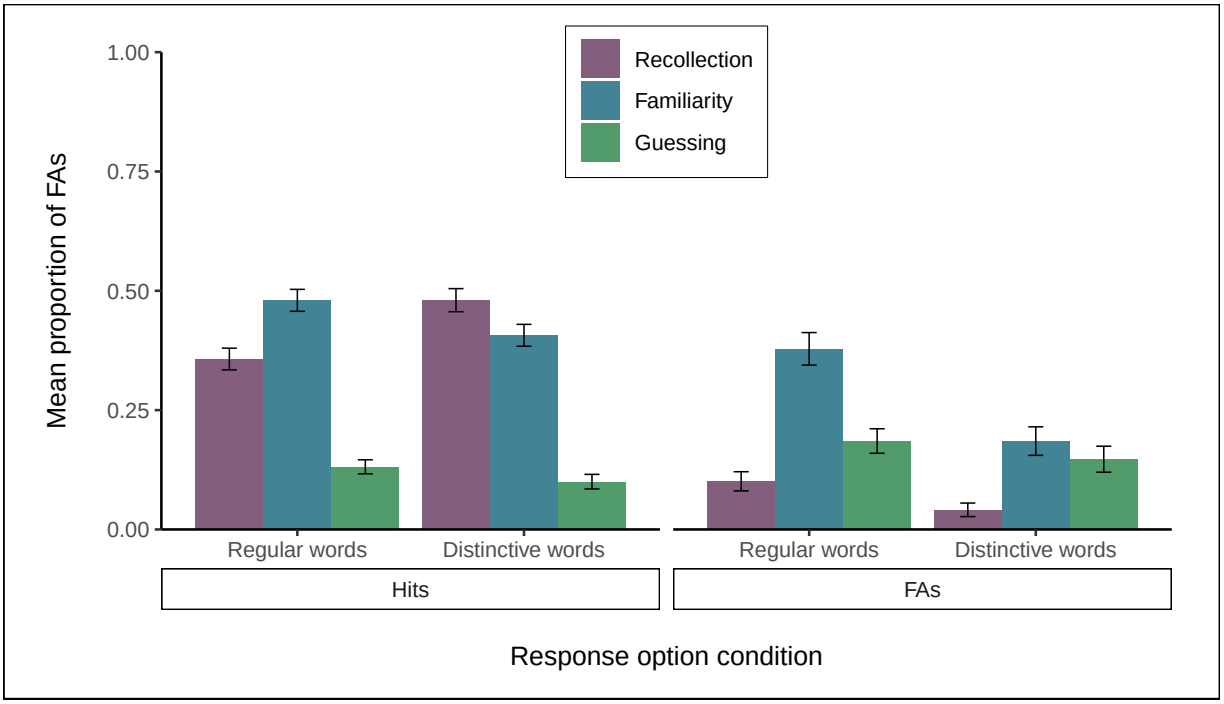


Figure 24: Mean proportion of hits and FAs assigned Recollection, Familiarity, and Guessing by stimuli format.

Analysis of the Ensor et al. (2019) data showed evidence of the distinctiveness manipulation. For the mean proportion of hits, there were significant main effects of stimuli format in their first [$F(3, 87) = 10.91$, $p < .001$, $\eta_p^2 = .27$] and second [$F(3, 87) = 8.37$, $p < .001$, $\eta_p^2 = .22$] experiments. Follow-up comparisons showed distinctive words resulted in significantly more hits than regular words across both experiments (see Table 13). For the mean proportion of FAs, the first experiment of Ensor et al. (2019) showed no significant main effect of stimuli format [$F(3, 87) = 1.41$, $p = .245$, $\eta_p^2 = .05$], however, the second experiment did [$F(3, 87) = 6.05$, $p < .001$, $\eta_p^2 = .17$]; though follow-up comparisons showed no difference in the proportion of FAs between distinctive and regular words. Finally, for mean *d'* scores, there was a significant main effect of stimuli format in both the first [$F(3, 87) = 10.99$, $p < .001$, $\eta_p^2 = .27$] and second [$F(3, 87) = 19.01$, $p < .001$, $\eta_p^2 = .40$] experiments. Follow-up comparisons showed evidence of the distinctiveness manipulation, whereby distinctive words produced significantly higher *d'* scores compared to regular words in both experiments (see Table 12).

Table 12: Mean proportion of hits for regular and distinctive word stimuli, across the current experiment and Ensor et al. (2019).

```
## Usually it is recommended to use column_spec before collapse_rows, especially in LaTeX,
```

		Grey words	Distinctive words	Planned comparisons
Hits				
Ensor, Surprenant, & Neath (2019)	Trials: 40 study / 80 test:	0.72	0.84	*t*(29) = 3.14, *p* = .004, *d* = 0.57
	Trials: 80 study / 160 test:	0.63	0.78	*t*(29) = 4.16, *p* < .001, *d* = 0.76
Current study	Trials: 60 study / 120 test:	0.58	0.61	*t*(157) = 1.26, *p* = .211, *d* = 0.10
FAs				
Ensor, Surprenant, & Neath (2019)	Trials: 40 study / 80 test:	0.15	0.08	*t*(29) = -1.94, *p* = .062, *d* = -0.35
	Trials: 80 study / 160 test:	0.29	0.23	*t*(29) = -1.48, *p* = .149, *d* = -0.27
Current study	Trials: 60 study / 120 test:	0.18	0.05	*t*(157) = -8.70, *p* < .001, *d* = -0.69
d'				
Ensor, Surprenant, & Neath (2019)	Trials: 40 study / 80 test:	2.18	2.97	*t*(29) = 3.69, *p* < .001, *d* = 0.67
	Trials: 80 study / 160 test:	1.09	1.89	*t*(29) = 7.35, *p* < .001, *d* = 1.34
Current study	Trials: 60 study / 120 test:	1.49	2.23	*t*(157) = 7.96, *p* < .001, *d* = 0.63