How robust is familiar face recognition? A repeat detection study of >1,000 faces

Angus F Chapman*1, Hannah Hawkins Elder*1, & Tirta Susilo1

1School of Psychology, Victoria University of Wellington
*Co-first authors

In prep for Royal Society Open Science (registered report)

Correspondence: Tirta Susilo, School of Psychology, Victoria University of Wellington, PO Box 600, Wellington 6040, New Zealand. Email: tirta.susilo@vuw.ac.nz

Keywords: face recognition, face memory, familiarity, repeat detection

Abstract

Recent theories suggest that familiar faces have a robust representation in memory because they have been encountered over a wide variety of contexts and image changes (e.g., lighting, viewpoint, expression). In contrast, unfamiliar faces are encountered only once, and so they do not benefit from such richness of experience and are represented based on image-specific details. In this registered report, we used a repeat detection task to test whether familiar faces are recognised better than unfamiliar faces across image changes. Participants viewed a stream of more than 1000 celebrity face images for 0.5 sec each, any of which might be repeated at a later point and has to be detected. Some participants saw the same image at repeats while others saw a different image of the same face. A post-experimental familiarity check allowed us to determine which celebrities were and were not familiar to each participant. We had three predictions: (i) detection would be better for familiar than unfamiliar faces, (ii) detection would be better across same than different images, and (iii) detection of familiar faces would be comparable across same and different images, but detection of unfamiliar faces would be poorer across different images. We obtained support for the first two predictions but not the last. Instead, we found that repeat detection of faces, regardless of familiarity, was poorer across different images. Our study suggests that the robustness of familiar face recognition may have limits, and that under some conditions, familiar face recognition can be just as influenced by image changes as unfamiliar face recognition.
Introduction

The ability to recognise faces is crucial to successful functioning in the social environment. Our ability to recognise hundreds or even thousands of faces points to the existence of a sophisticated face recognition system. To study how this system works, scientists typically use two kinds of stimuli: familiar faces and unfamiliar faces. Familiar faces, by definition, are those we recognise and can identify. Many basic signatures of face processing apply to both types of stimuli, including perceptual markers like the face inversion effect (Yin, 1969; but see Megreya & Burton, 2006), the composite face effect (Young et al., 1987), the contrast negation effect (Kemp et al., 1996), the Thatcher effect (Thompson, 1980), and the face aftereffect (Webster & MacLin, 1999; Leopold et al., 2001), as well as brain markers from fMRI (activation in fusiform face area; Haxby et al., 2000), EEG (N170; Eimer, 2000), and MEG (M170; Ewbank, Smith, Hancock, & Andrews, 2008). These findings show that some processes in face recognition are fundamental and operate on all faces regardless of familiarity.

However, a great deal of evidence also shows that familiar and unfamiliar faces are processed in different ways. Familiar faces are recognised better and quicker than unfamiliar faces (Bruce & Young, 1986; Ellis, Shepherd, & Davies, 1979). Familiar faces are better recognised using internal features (e.g., eyes, mouth) than external features (e.g., hair) (Clutterbuck & Johnston, 2005; Ellis et al., 1979; Osborne & Stevenage, 2008; Young et al., 1985). Familiar faces are less likely to be missed in change detection and attentional blink paradigms (Jackson & Raymond, 2006; 2008). Finally, familiar and unfamiliar face recognition doubly dissociate – some acquired prosopagnosics are impaired with familiar but not unfamiliar faces (Bauer, 1984; Young et al., 1993) while others show the opposite pattern (Malone et al., 1982).

Taken together, these findings motivate the idea that familiar and unfamiliar faces rely on distinct mental representations (Johnston & Edmonds, 2009; Burton, Jenkins, & Schweinberger, 2011; Burton, 2013). Familiar faces are those we have seen many times in a wide range of contexts (e.g., different expression, viewing angle, lighting condition). Familiar faces are thought to rely on structural codes, which reflect some sort of an averaging process in which a robust representation of a face is constructed across multiple encounters that extracts the underlying stability of a face while filtering out irrelevant variability (Bruce & Young, 1986; Burton, Jenkins, Hancock, & White, 2005; Jenkins & Burton, 2011). Mental representations of familiar faces are therefore rich and encompassing because they integrate diverse experiences to create a stable impression of a specific person. On the other hand, unfamiliar faces are encountered only once, which means their representations can only comprise the one exposure. Unfamiliar faces are said to rely on pictorial codes, which involve storing the visual details of a particular face image (Bruce & Young, 1986). Unfamiliar face recognition is therefore more analogous to image matching, and as a result changes in specific image details typically leads to poor performance.

The distinction above predicts that changes to images will impair recognition of familiar faces less than unfamiliar faces. Consistent with this idea, many studies have found that compared to unfamiliar faces, familiar faces are more impervious to a wide variety of image changes including expression (Bruce, 1982; Patterson & Baddeley, 1977), lighting (Hill & Bruce, 1996), viewpoint (O'Toole et al., 1998), head orientation (Bruce, 1982; Patterson & Baddeley, 1977), image quality (Bruce, Henderson, Newman, & Burton, 2001; Burton, Wilson, Cowan, & Bruce, 1999), and environmental context (Davies & Milne, 1982). However, these studies tend to employ traditional tasks such as the old/new recognition task (presenting a subset of stimuli from the ‘study’ phase
interspersed with novel stimuli and asking participants to make an 'old/new' judgement; Armann et al., 2016; Bruce, 1982; Ellis et al., 1979; Yarmey, 1971), array face-matching task (asking participants to match a given target face to a different image of the target from a multiple-face array containing an image of the target and a number of novel face images; Megreya & Burton, 2006, 2007), or paired face-matching task (presenting participants with pairs of faces and asking them to make a 'same/different' judgement for each pair; Clutterbuck & Johnston, 2002; Megreya & Burton, 2006, 2007). Unlike real-world face learning, these tasks are typically short in duration (between 5-20 minutes) and use a relatively small number of stimuli (between 30-150) (Armann et al., 2016; Bruce, 1982; Ellis et al., 1979; Megreya & Burton, 2006). This means we don’t really know whether familiar face recognition remains robust to image changes under more demanding experimental conditions with significant memory load.

Here we investigate whether familiar face recognition remains robust to image changes with a challenging repeat detection task (Isola, Xiao, Torralba, & Oliva, 2011; Bainbridge, Isola, & Oliva, 2013; Bainbridge, 2017). Our task uses more than 1000 face images and takes almost 1 hr to complete. In this task, participants view a stream of face images and respond whenever they detect a repeated face. Participants don’t know which faces will be seen again, so they must attend to and memorise every face as a potential target. This task is similar to those used in studies of repetition priming (Bruce & Valentine, 1985; Johnston & Barry, 2001), in which presentation of face images enhances recognition of those faces on a later task. We use celebrity faces as stimuli to avoid the practical issues of devising a stimulus set that is specific to each participant (e.g., friends or family members). We perform a familiarity check at the end of the experiment to assess each participant’s familiarity with a subset of the stimuli and to classify which faces are familiar and which are not (Clutterbuck & Johnston, 2002).

Our critical manipulation involves the type of image shown for repeated faces (Figure 1). Repeated images could appear at one of two different times: vigilance repeats occurred shortly following the initial presentation of a face, and were included to assess participants’ engagement with the task; target repeats occurred much longer after the initial face was presented, and were the main concern of our analysis. In one condition, the repeats show the same image as the initial presentation, while in the other condition the repeats present different images of the same person. Our prediction is threefold. First, we expect better repeat detection when the target face is familiar rather than unfamiliar, regardless of image type. Second, we expect better detection when the target image is the same as the initial image than when it is a different image of the same face, regardless of familiarity. Third, we expect that detection of familiar faces will be less
influenced by image type than detection of unfamiliar faces is, and that detection of unfamiliar faces will be poorer when the target image is different than the initial face.

Method

Pilot Study.

We conducted a pilot study to estimate the effect of familiarity on repeat detection of faces. The full method and results are provided in the Supplementary Materials. There was a familiarity by image interaction, $F(1, 43) = 5.79, p = .020, \eta_p^2 = .119, BF_{10} = 2.84$, which revealed that unfamiliar faces were better detected with ‘same’ than ‘different’ images, $t(42.77) = 3.76, p < .001, d_s = 1.15, BF_{10} = 52.10$, while familiar faces were detected equally well regardless of image, $t(41.28) = 0.99, p = .327, d_s = 0.30, BF_{10} = 0.44$.

Power analysis.

No studies have compared familiar and unfamiliar faces using a repeat detection task, so we conducted power analysis based on the results of the pilot study. For the interaction effect, we converted $\eta_p^2 (0.119)$ to $\omega^2_p (0.096)$ to reduce bias in estimating effect size (Lakens, 2013), which corresponded to Cohen’s $f = 0.326$. A power analysis performed in GPower 3.1 with an effect size of 0.294 (which is 90% of the pilot effect size, to account for potential overestimation) gives a required sample size of 124 to achieve 90% power. Power analyses for the main effects of familiarity and image produced smaller required samples (main effect of familiarity: Cohen’s $f = 2.145, n = 6$; main effect of image: Cohen’s $f = 0.311, n = 112$). This means a sample size of 124 will have >90% power to test our predictions and capture all three effects (should they exist). We therefore planned to test participants until a sample of 124 non-excluded participants was reached.

We note that our effect size for the interaction between familiarity and image is smaller than those reported in past studies. For example, the effect sizes ($\eta_p^2$) for the interaction between familiarity and image were .27 and .22 in Armann et al. (2016), and .58 in Bruce (1982, Experiment 2). The effect size discrepancy between these studies and ours is likely due to task and/or stimulus differences, and possibly publication bias.

Participants.

One-hundred and forty-six first-year psychology students at Victoria University of Wellington participated for course credit. They were randomly assigned to the ‘same’ image ($n = 73$) or ‘different’ image ($n = 73$) conditions. Five participants (2 in the ‘same’ condition, 3 in the ‘different’ condition) did not complete the familiarity task due to a technical error, and their data was excluded. Seventeen participants (9 in the ‘same’ condition; 8 in the ‘different’ condition) were excluded based on pre-determined criteria (see Analytic Protocol). Our final sample therefore consisted of 124 participants, equally assigned to each condition. Participants were between 18 and 27 years of age ($M = 18.71, SD = 2.10$) and were predominantly women ($n = 98$). Most reported their ethnicity as New Zealand European/Pākehā ($n = 99$), with the remaining Māori ($n = 11$), Pacific ($n = 1$), Asian ($n = 4$), or as another unspecified ethnicity ($n = 9$).

Materials.

To create our stimulus set, we first drafted a list of 1006 celebrities who were actors, musicians, or other notable figures in popular culture. For each celebrity, we downloaded several face images from a Google Image search. Two images were selected for each celebrity to make up the stimulus set. The two images for each celebrity varied in terms of lighting, viewpoint, expression, and date taken. Images were cropped in an
oval to exclude non-face, background information (as in e.g., Bainbridge et al., 2013; Bainbridge, 2017). Each face was coded on the four demographic attributes: glasses (4.7% of faces); skin colour (10.1% dark, 89.9% light); gender (53.7% male, 46.3% female); and age (0.8% child, 91.6% adult, 7.6% elderly). We selected 100 celebrities as target repeats with a demographic breakdown that mirrors that of the full stimulus set to ensure that targets and non-targets cannot be easily discriminated using demographic clues. We also selected 25 extra celebrities from the remaining stimulus set to use as vigilance repeats.

**Procedure.**

**Repeat detection task.** Participants viewed a stream of 1131 images, each presented for 500 ms with a 1000 ms inter-trial interval. Of these images, 125 depicted a famous face shown earlier in the experiment, and participants were instructed to press the space bar when they detected these repeats. There are three types of trials. First are vigilance repeats (25 trials), in which a face was repeated quickly, within 4-8 trials of the initial presentation. These were included to gauge whether participants were focusing on the task. Second are target repeats (100 trials), in which a face was repeated, but the number of intervening faces was greater, within 91-109 trials of the initial presentation. Third are fillers (1006 trials), that include presentations of faces who were shown only once during the task (881 trials), but also the initial presentation of a target face in repeat trials (125 trials). For both vigilance and target repeats, the correct response for participants was to press the space bar, while for fillers the correct response was no response. Figure 1 shows our trial design. Participants were given a 30 second rest break every 162 images (∼4 min).

The critical manipulation was whether the repeated face (for both vigilance and target repeats) was the exact same image as the initial face (‘same’ condition), or whether it was a different image of the same face (‘different’ condition). This was manipulated between-subjects. The first image presented was randomly selected from the two images created per celebrity (see Materials), while the second image was determined by condition.

**Familiarity check.** Following the repeat detection task, we asked participants to identify 100 target faces and 100 demographic-matched filler faces. Participants were presented with each face image and asked to identify the person by typing their name in a text box. The face images shown were the same as the first presentation of the target in the repeat detection task. Participants were instructed that a sentence describing the famous face (e.g., a character’s name) was also acceptable. No time or word limit was placed on these responses. Responses were classified as “familiar” if participants provided the correct full name, made a minor error in naming, gave a specific example that identified the individual (e.g., a character played in a TV show or film), or provided a general category (e.g., actor, singer) that the famous individual belonged to. Responses that were blank, or otherwise incorrect were classified as “unfamiliar”. All non-blank responses (n = 7771) were scored by two coders (κ = .931) and any disagreements were resolved through discussion. In addition to identifying each of the target faces, participants were asked to rate their familiarity with the face on a 7-point Likert scale, from 1 (‘not at all familiar with this person’) to 7 (‘highly familiar with this person’). **Analytic protocol.**

The analytic protocol described here are as was used in our pilot study.

**Exclusion.** Participants were excluded based on specific criteria for both familiarity and repeat detection tasks. We excluded data at the subject level, but not at the level of individual trials. **Repeat detection task:** We first excluded participants based
on vigilance performance. Bainbridge and colleagues (2013) excluded participants who identified less than half of any 10 consecutive vigilance repeats. In our study, because vigilance repeats are more difficult for the ‘different’ group, we adjusted this procedure to avoid excluding unequal numbers of participants in each group. We excluded participants whose vigilance accuracy was >2 SDs below the mean for their condition. 

**Familiarity task:** Our analysis requires that participants are familiar with some target faces, so we also excluded participants who were familiar with <10% of target faces.

**Data checks.** To confirm data quality, we performed four data checks by testing whether: (1) the number of familiar faces identified differed between groups, (2) hits for vigilance repeats was higher than false alarms for fillers, (3) hits for vigilance repeats was higher for ‘same’ than ‘different’ groups, and (4) false alarms for fillers was comparable across groups.

**Main analysis.** All analyses were conducted in R (R Core Team, 2016). For the main analysis assessing familiarity on repeat detection for the two groups, we fit a linear mixed-effects model with the package lme4 (Bates, Maechler, Bolker, & Walker, 2015), and analysis of variance was conducted with lmerTest (Kuznetsova, Brockhoff, & Christensen, 2016) using Type III sums of squares and Satterthwaite approximation for degrees of freedom. To test our hypothesis that familiarity will influence detection of familiar and unfamiliar faces differently depending on the type of image, we conducted a 2 × 2 mixed ANOVA with image (same vs different) and familiarity (familiar vs unfamiliar, based on responses of each participant during the familiarity task) as fixed factors, with a random intercept modelled for each participant. Because the number of faces people recognise as familiar and unfamiliar are unlikely to be equal, a weighted contrast was used so that means reflect the true average across this factor. Analysis of null results was complemented by Bayesian statistics. Using the BayesFactor package (Morey & Rouder, 2015), we conducted Bayesian t-tests alongside the traditional t-tests; F-ratios were first converted to t-values (Dienes, 2014). We report the Bayes factor in favour of the alternative hypothesis (BF_{10}) based on analysis using a ‘default’ prior with scale = 0.707 (Morey, Romeijn, & Rouder, 2016).

**Data availability.** All data and analysis scripts are available at osf.io/6p2k4/?view_only=c15990c96bc0422e984dc804451355b6

### Results

#### Registered Analyses.

**Data Checks.** To ensure that data from the familiarity and repeat detection tasks were valid for further analysis, we conducted four checks. First, in the familiarity task, participants identified a reasonable number of faces (M = 24.9%, SD = 11.6, range = 10.0-70.0), which did not differ between groups, t(121.94) = 1.10, p = .273, d = 0.20, BF_{10} = 0.332. Second, in the repeat detection task, hit rates for vigilance repeats (M = 46.9%, SD = 23.6) was higher than false alarms for fillers (M = 9.9%, SD = 7.5), t(123) = 18.31, p < .001, d = 1.64, BF_{10} > 1000. This shows that performance on vigilance trials was not simply due to participants responding at random. Third, vigilance detection was greater with ‘same’ than ‘different’ images, t(119.12) = 10.34, p < .001, d = 1.87, BF_{10} > 1000, consistent with previous literature (Armann et al., 2016; Bainbridge, 2017; Bruce, 1982). Four, false alarms for fillers did not differ between groups, t(121.94) = 0.54, p = .591, d = 0.10, BF_{10} = 0.219, showing that false responses to fillers was unaffected by changes to the target images. Table 1 contains group means and SDs for data checks; Figure 2 plots individual data. These data checks produced similar results as we found in our pilot study.
Table 1. Group mean (SD) percentages of faces identified in the familiarity task, hit rates for vigilance repeats, and false alarms for fillers in the repeat detection task.

<table>
<thead>
<tr>
<th>Group</th>
<th>Familiarity task</th>
<th>Repeat detection task</th>
<th>Fillers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% Faces identified</td>
<td>Vigilance</td>
<td>Fillers</td>
</tr>
<tr>
<td>Same</td>
<td>23.8 (11.4)</td>
<td>63.0 (18.6)</td>
<td>9.5 (7.5)</td>
</tr>
<tr>
<td>Different</td>
<td>26.0 (11.7)</td>
<td>32.8 (12.2)</td>
<td>10.2 (7.4)</td>
</tr>
</tbody>
</table>

**Main Analysis.** We compared detection hit rates in each group for target faces that participants reported as familiar and unfamiliar. Results are presented in Table 2 and Figure 3. Familiar faces were detected more accurately ($M = 61.9\%$) than unfamiliar faces ($M = 25.5\%$), reflected by the main effect of familiarity, $F(1, 122) = 595.74$, $p < .001$, $\eta^2_p = .830$, $BF_{10} > 1000$. There was also a main effect of image, $F(1, 122) = 55.44$, $p < .001$, $\eta^2_p = .312$, $BF_{10} > 1000$, as repeat detection was better for the ‘same’ ($M = 44.7\%$) than the ‘different’ ($M = 24.6\%$) images. However, unlike in the pilot study, there was no interaction between familiarity and image, $F(1, 122) = 2.60$, $p = .110$, $\eta^2_p = .021$, $BF_{10} = 0.617$, although Bayes factors provided inconclusive evidence for this null effect.

Unfamiliar faces were better detected with ‘same’ than ‘different’ images, $t(102.59) = 7.15$, $p < .001$, $d = 1.30$, $BF_{10} > 1000$, and this was also true for familiar faces, $t(121.97) = 6.37$, $p < .001$, $d = 1.15$, $BF_{10} > 1000$.

Can our results be accounted for by the fact that the ‘different’ image condition is more difficult than the ‘same’ image condition? To address this issue, we restricted our analysis to a subset of participants in each condition such that they were equated on vigilance performance. To achieve this, we used an iterative approach where we selected the $n$ participants in the ‘different’ image condition who performed best on vigilance repeats, and the $n$ participants in the ‘same’ image condition who performed worst on vigilance repeats. We varied $n$ across the range of the sample ($n = 2$ to $n = 62$ in each condition) to find the point at which the difference in vigilance performance

---

*Figure 2.* Individual data for data checks. Dots are jittered horizontally to improve visibility. Left panel shows the number of familiar faces is comparable across groups. Right panel shows that hit rates for vigilance repeats is higher for ‘same’ than ‘different’ groups, whereas false alarms for fillers are similar. Error bars show 95% CI around the mean.
between groups was minimized. This point was reached by selecting the top 27 participants in the ‘different’ image condition, and bottom 25 participants in the ‘same’ image condition. This subset was equated on vigilance performance ($M = 45.2\%, SD = 9.5$), $t(49.99) = 0.26, p = .799, d_s = 0.07, BF_{10} = 0.286$, however those in the ‘different’ image condition were familiar with a greater number of the target faces ($M = 33.1\%, SD = 11.3$) than the ‘same’ image condition ($M = 20.8\%, SD = 6.9$), $t(43.51) = 4.77, p < .001, d = 1.35, BF_{10} = 825$.

Results for the analysis of hit rates are shown in Figure 4. We found a main effect of familiarity, $F(1, 50) = 353.01, p < .001, \eta^2_p = .876, BF_{10} > 1000$, but not image, $F(1, 50) = 0.30, p = .585, \eta^2_p = .006, BF_{10} = 0.315$. Crucially, the lack of interaction between familiarity and image remained, $F(1, 50) = 1.90, p = .174, \eta^2_p = .174, BF_{10} = 0.606$.

Detection of unfamiliar faces was unaffected by image type, $t(49.98) = 0.38, p = .705, d = 0.11, BF_{10} = 0.295$, nor was detection of familiar faces, $t(48.76) = 0.99, p = .325, d = 0.28, BF_{10} = 0.419$. Therefore, even when both groups performed equally on vigilance trials, familiarity with the target face improved detection similarly regardless of whether the repeated face was the same or different.

---

### Table 2. Group mean (SD) hit rates for detection of familiar and unfamiliar targets in the repeat detection task.

<table>
<thead>
<tr>
<th>Group</th>
<th>% Hits</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Familiar faces</td>
<td>Unfamiliar faces</td>
<td></td>
</tr>
<tr>
<td>Same</td>
<td>73.8 (20.6)</td>
<td>35.0 (17.6)</td>
<td></td>
</tr>
<tr>
<td>Different</td>
<td>50.1 (20.9)</td>
<td>16.1 (11.1)</td>
<td></td>
</tr>
</tbody>
</table>

---

*Figure 3. Individual data for correct detection of target repeats by familiarity and group. Dots are jittered horizontally to improve visibility. Familiar faces were detected better than unfamiliar faces in both groups, and faces were detected better by ‘same’ than by ‘different’ group. Error bars show 95% CI around the mean.*
**Exploratory Analyses.**

**More specific prior on interaction.** While the interaction between familiarity and image type in the main analysis was not significant, Bayes factors did not provide strong support for the null hypothesis. But because the Bayes factor for this effect was calculated without assuming the direction of the interaction, and because our data showed a pattern opposite to our prediction (and what we found in our pilot study), this value underestimates the true evidence against our registered hypothesis. If we specifically test the hypothesis that unfamiliar faces will be more affected by changes in image by specifying a one-sided test, Bayes factors show strong support for the null hypothesis, $BF_{10} = 0.079$. This analysis provides additional evidence that our findings are inconsistent with our registered predictions.

**d’**. Because we asked participants to indicate their familiarity with 100 of the filler faces, we were able to calculate false alarms (and hence d’) for familiar and unfamiliar faces. These results are presented in Table 3 and Figure 5. Participants were more sensitive to repeats of familiar ($M = 1.64$) than unfamiliar faces ($M = 0.71$), reflected by the main effect of familiarity, $F(1, 122) = 272.79, p < .001, \eta^2_p = .691, BF_{10} > 1000$. There was also a main effect of image, $F(1, 122) = 71.34, p < .001, \eta^2_p = .369, BF_{10} > 1000$, as repeat detection was better for the ‘same’ ($M = 1.29$) than the ‘different’ ($M = 0.60$) images. There was also an interaction between familiarity and image, $F(1, 122) = 6.79, p = .010, \eta^2_p = .053, BF_{10} = 3.96$. Participants were more sensitive with ‘same’ than ‘different’ images for both unfamiliar faces, $t(117.79) = 7.14, p < .001, d = 1.29, BF_{10} > 1000$, and familiar faces, $t(121.22) = 7.39, p < .001, d = 1.34, BF_{10} > 1000$, but this difference was greater for familiar ($M = 0.93$) than unfamiliar faces ($M = 0.69$).

![Figure 4. Individual data for correct detection of target repeats by familiarity and group in a subset of participants who were equated for performance on vigilance repeats. Dots are jittered horizontally to improve visibility. In both groups, detection was better for familiar than unfamiliar faces. Error bars show 95% CI around the mean.](image-url)
Table 3. Group mean (SD) d’ and false alarm rates for detection of familiar and unfamiliar targets in the repeat detection task.

<table>
<thead>
<tr>
<th>Group</th>
<th>d’</th>
<th>Unfamiliar faces</th>
<th>% False alarms</th>
<th>Unfamiliar faces</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same</td>
<td>2.09 (0.71)</td>
<td>1.02 (0.53)</td>
<td>11.2 (10.1)</td>
<td>9.9 (8.6)</td>
</tr>
<tr>
<td>Different</td>
<td>1.18 (0.66)</td>
<td>0.40 (0.43)</td>
<td>14.2 (12.1)</td>
<td>9.3 (9.0)</td>
</tr>
</tbody>
</table>

The finding of an interaction between familiarity and image with d’ but not hit rates can be explained by differences in false alarm rates (Figure 6). False alarms were more common to familiar (M = 12.7%) than unfamiliar faces (9.6%), F(1, 122) = 16.60, p < .001, $\eta^2_p = .120$, BF10 = 205. There was no effect of image type, $F(1, 122) = 0.59, p = .444$, $\eta^2_p = .005$, BF10 = 0.250. However, there was an interaction between familiarity and image, $F(1, 122) = 5.25, p = .024, \eta^2_p = .041$, BF10 = 2.01, such that false alarms were greater for familiar than unfamiliar faces when the images were ‘different’, t(61) = 4.28, p < .001, BF10 = 312, but there was no difference when for ‘same’ images, t(61) = 1.33, p = .188, BF10 = 0.322. Although this effect on false alarms was small, when combined with the effects on hit rate, it leads to the interaction found in d’.

**Familiarity rating.** Participants also rated their familiarity with each target face on a 7-point scale. This rating allows us to assess the effects of familiarity on repeat detection of faces with a more continuous scale, in addition to the dichotomous familiar/unfamiliar scoring by independent coders. To do this, we fit a binomial mixed-effects model predicting hit rates for target repeats, with fixed effects for self-reported familiarity as well, as the main effects and interactions for coder-scored familiarity and image condition (similar to the models used for hit rate and d’). We also fit random

![Figure 5](image-url) **Figure 5.** Individual data for d’ by familiarity and group. Dots are jittered horizontally to improve visibility. Participants were more sensitive to repeats of familiar faces than unfamiliar faces in both groups, and more sensitive to repeats in the ‘same’ than ‘different’ group. Error bars show 95% CI around the mean.
intercepts for each participant, participant by coder-scored familiarity (subject effects), and target image identity (item effects). Significance of fixed effects was determined through comparison of nested models.

This model was able to replicate the previous analysis on hit rate: there was a main effect of coder-scored familiarity, $\chi^2 = 179.73$, $p < .001$, and of image, $\chi^2 = 42.39$, $p < .001$, but no interaction between coder-scored familiarity and image, $\chi^2 = 0.17$, $p = .676$. Notably, there was no effect of self-reported familiarity, $\chi^2 = 0.19$, $p = .660$, suggesting that hit rates in each condition were independent of participants’ self-reported familiarity with the target faces. This data is shown in Figure 7, which clearly demonstrates that self-reported familiarity with the target faces was not predictive of actual repeat detection performance. Self-reported familiarity did not differ when coders scored the response as familiar ($M = 3.11$, $SD = 0.91$) and unfamiliar ($M = 3.16$, $SD = 0.93$), $t(123) = 1.33$, $p = .185$, $d = 0.24$, $BF_{10} = 0.237$. In fact, participants gave numerically higher ratings of familiarity when their responses were scored as unfamiliar, providing further evidence against the ability of self-reported familiarity to influence detection performance.

**Relationship between face knowledge and repeat detection.** There were substantial participant differences in knowledge of target and filler faces ($M = 26.0\%$, range = 8.5-73.0\%). To assess whether greater face knowledge improved detection performance, we correlated participants’ familiarity score with their $d’$ in the repeat detection task. We ran separate analyses for familiar and unfamiliar faces because familiar faces had higher overall $d’$ (Figure 5). Greater face knowledge was related to better detection of familiar faces, $r(122) = .204$, $p = .023$, although Bayes factors provided inconclusive evidence for this relationship, $BF_{10} = 2.05$. A similar relationship was found between face knowledge and detection of unfamiliar faces, $r(122) = .258$, $p = .004$, $BF_{10} = 8.99$. 

**Figure 6.** Individual data for false alarms by familiarity and group. Dots are jittered horizontally to improve visibility. Familiar faces were incorrectly detected more when in the ‘different’ group. Error bars show 95% CI around the mean.
These relationships appear to be driven by a reduction in false alarms as face knowledge increases, rather than greater hit rates: whereas the relationship between face knowledge and hits was not significant for either familiar, $r(122) = .025, p = .786$, $BF_{10} = 0.198$, or unfamiliar faces, $r(122) = .106, p = .243, BF_{10} = .357$, a negative relationship was found between face knowledge and false alarm rates for both familiar, $r(122) = -.252, p = .005, BF_{10} = 7.44$, and unfamiliar faces, $r(122) = -.188, p = .037$, although Bayes factors were inconclusive for the latter correlation, $BF_{10} = 1.42$.

Collectively, these data suggest that greater face knowledge (indexed by greater rates of identification in the familiarity task) improved discrimination of familiar and unfamiliar targets from fillers in the repeat detection task.

**Discussion**

Theories of face recognition propose that familiar and unfamiliar faces are represented differently in memory: unfamiliar faces are said to rely on *pictorial codes* (the visual details of a specific face image), while familiar faces are thought to rely on *structural codes* (a rich and robust representation of a face generated by averaging across multiple exposures to the same face; Bruce & Young, 1986; Burton, Jenkins, Hancock, & White, 2005; Jenkins & Burton, 2011). The robust representations of familiar faces ought to facilitate recognition because they filter out the idiosyncratic information unique to single exposures to a specific face. In this registered study, we examined this proposal with a repeat detection task – an intensive and demanding experimental paradigm more akin to face recognition in real life (Isola et al., 2011; Bainbridge et al., 2013; Bainbridge, 2017). We had three predictions. First, we predicted better detection of familiar faces. Second, we predicted better detection across same images. Third, we predicted an interaction between familiarity and image type.

Specifically, we expected stable detection of familiar faces regardless of repeated image, but poorer detection of unfamiliar faces when the repeat involves different images.

*Figure 7.* Group average data for correct detection of target repeats by group, rater-scored familiarity, and self-reported familiarity. Dot position is adjusted to improve visibility. Self-reported familiarity was unrelated to detection performance. Error bars show 95% CI around the mean.
As expected, detection was superior when participants were familiar with the target face, and when the face image they had to detect was identical to the initial presentation. Consistent with past research (Armann et al., 2016; Bruce, 1982), these findings indicate that we are better at recognising faces which we have had greater exposure to, and that recognising faces across same images is easier than recognising them across different images. Our results advance the literature by showing that these findings hold in a study with registered data collection and analytic protocols. However, our prediction of interaction between familiarity and image type was not supported. Instead, we found that repeat detection of faces, regardless of familiarity, was poorer across different images. This result is at odds with prior work showing the robustness of familiar face recognition to a wide range of image changes (e.g. lighting, viewpoint; Bruce, 1982; Hill & Bruce, 1996; O'Toole et al., 1998; Patterson & Baddeley, 1977).

What explains the discrepancy between our finding and those of past studies showing that familiar face recognition is robust to image changes? One possibility is task differences. Unlike many studies that lasted for up to 20 minutes and use between 30-150 face stimuli (Armann et al., 2016; Bruce, 1982; Ellis et al., 1979; Megreya & Burton, 2006), our task had participants remain vigilant and pay attention to more than 1,000 face images for almost 1 hour. If correct, our finding suggests that the robustness of familiar face recognition has limits. Future research could vary the parameters of the task (stimulus load, presentation duration, time between repeats, etc.) to further characterise the specific conditions under which familiar face recognition is no longer robust. Another possibility is that because most participants were unfamiliar with a majority of the faces (the average participant only knew about a quarter of the celebrities), they might have relied primarily on a pictorial strategy throughout the experiment. The advantage for familiar over unfamiliar faces might then be due to participants simply responding more often to faces they are familiar with, which is consistent with greater false alarms to familiar faces (3.1% increase), although familiar faces also produced greater hits (36.4% increase). This account assumes that familiar face recognition can rely on both structural and pictorial codes, depending on task requirement. Finally, it might also be the case that familiar face recognition relies not only on structural and pictorial codes, but also semantic knowledge associated with the face. Instead of memorising each face that was presented in the repeat detection task, our participants may have used this semantic information in addition to pictorial codes. Because semantic information is only available to familiar faces, they are better detected on repeats than unfamiliar faces. However, when semantic information is not available, detection still benefits from pictorial codes. If this interpretation is correct, our finding that hit rates were greater for familiar-but-different faces than for unfamiliar-but-same faces suggests that participants were more reliant on semantic information than on pictorial cues.

In addition to the registered results, we conducted two exploratory analyses. In one analysis, we found that participants’ self-rated familiarity with target faces (as opposed to our scoring/coding of their verbal responses) did not predict how well they detected those faces. This finding accords with recent studies showing that people’s judgments about various aspects of face recognition are not related to actual performance (Bainbridge 2017; Palermo et al., 2016), although it is possible that our participants’ notions of familiarity are idiosyncratic because our instruction did not specify what counts as “familiar” (i.e., one participant might have rated a face as ‘highly familiar’ having only seen it a few times, whereas another may need to have seen the face many more times before making the same judgement.). A second exploratory
analysis revealed that participants who knew more celebrities were more sensitive at the task (as measured by d’). This relationship holds regardless of familiarity and image conditions; participants with greater face knowledge were less likely to mischaracterise a new face as a repeat, even when that face was unfamiliar to them. One explanation might be that both familiar and unfamiliar faces rely on common perceptual representations, such as those proposed by face space (Valentine, 1991; Rhodes & Leopold, 2011) or holistic (Yin, 1969; Tanaka & Farah, 1993) accounts, the only difference being that familiar faces trigger connections with visual and semantic long-term memory.

A potential limitation of our study is fatigue. Participants might have become tired over time because our repeat detection task is long and demanding. This might have led to poorer detection towards the end of the study. However, these effects are unlikely to undermine our findings. Although target trials are more likely later in the repeat detection task (because of the various time lags between repeats), this should be equally true in all conditions, yet we still found a large difference in performance across conditions (from 16% in the unfamiliar/different condition to nearly 75% in the familiar/same condition). Similarly, participants might have responded less reliably in the post-experimental familiarity check. This would lead us to underestimate participants’ true level of familiarity (which could mask potential differences between familiar and unfamiliar faces), yet we still obtained a main effect of familiarity.

In sum, our study is the first to investigate the robustness of familiar face recognition with an intensive repeat detection task. The combination of a priori power analysis, large sample size, extensive stimulus set, and registered protocols makes our results particularly robust. Our study demonstrates that familiar faces are memorised better than unfamiliar faces, and that recognition is better when faces are repeated using the same image. However, our study suggests that while familiar face recognition benefits from structural codes not available to unfamiliar faces, the benefit may have limits. Finally, our study illustrates the value of the repeat detection task as an experimental paradigm that can shed new lights not only on familiar and unfamiliar face recognition, but also on other theoretical issues in the face processing literature.
Research Ethics. This study was approved by the Victoria University of Wellington Human Ethics Committee.

Data availability. Our data and analysis scripts are available on OSF at osf.io/6p2k4/

Competing interests. We declare no competing interests.

Authors’ contributions. AC, HHE, and TS designed the experiment. AC and HHE collected data and conducted statistical analyses. AC, HHE, and TS wrote the manuscript, and all gave final approval for publication.

Funding. No external sources of funding were provided for this research.
References


Dienes Z. Using Bayes to get the most out of non-significant results. Front Psychol. 2014;5:781. (doi:10.3389/fpsyg.2014.00781)


Supplementary Materials

Pilot Study

Method

Except where noted, our method in the pilot study was the same as in the registered study.

Participants. Sixty-four first-year psychology students at Victoria University of Wellington participated for course credit. They were randomly assigned to the ‘same’ image (n = 31) or ‘different’ image (n = 33) conditions. One participant in the different image condition was excluded from analyses because s/he did not respond during the repeat detection task. Eight participants (three in ‘same’ condition; five in ‘different’ condition) did not complete the familiarity task because of a technical error, so their data was excluded. Our sample therefore consisted of 55 participants.

Procedure. Familiarity task. In the pilot study, participants identified only the 100 target faces, and did not provide subjective familiarity ratings. All non-blank responses (n = 1533) were scored by two coders (κ = .951) and a third coder resolved any disagreements.

Technical issue. A programming error in the familiarity task prevented us from assessing the effects of familiarity as completely as we had hoped. All participants were shown 100 faces, but they were drawn out of the entire target set of 200 images (two images for each face). This means some participants saw some faces twice and didn’t see others at all. On average, participants saw 75.32 unique faces (SD = 2.57, range = 70-82). Familiarity for faces that participants saw twice was scored as the higher of the two responses. This error was fixed for our registered study.

Pilot data and analysis scripts are available at osf.io/6p2k4/
Results

Exclusions. Nine participants (3 in the ‘same’, 6 in the ‘different’ groups) were excluded for low familiarity, and one participant in the ‘same’ condition was excluded for poor performance on vigilance repeats. The final sample consisted of 45 participants (21 in the ‘different’, and 24 in the ‘same’ condition).

Data checks. To ensure that data from the familiarity and repeat detection tasks were valid for further analysis, we conducted four checks. First, in the familiarity task, participants identified a reasonable number of faces ($M = 30.4\%, SD = 12.6$, range = 10.7-55.8$%), which did not differ between groups, $t(42.90) = 0.84$, $p = .405$, $d_s = 0.26$, $BF_{10} = 0.39$. Second, in the repeat detection task, hit rates for vigilance repeats ($M = 46.4\%, SD = 19.6$) was higher than false alarms for fillers ($M = 12.5\%, SD = 9.8$), $t(44) = 10.98$, $p < .001$, $d_s = 1.64$, $BF_{10} > 1000$. This shows that performance on vigilance trials was not simply due to participants responding at random. Third, vigilance detection was greater with ‘same’ than ‘different’ images, $t(41.58) = 5.85$, $p < .001$, $d_s = 1.78$, $BF_{10} > 1000$, consistent with previous literature (Armann et al., 2016; Bainbridge, 2017; Bruce, 1982). Four, false alarms for fillers did not differ between groups, $t(42.72) = 0.55$, $p = .582$, $d_s = 0.17$, $BF_{10} = 0.33$, showing that false responses to fillers was unaffected by changes to the target images. Table S1 contains group means and SDs for data checks; Figure S1 plots individual data.

<table>
<thead>
<tr>
<th>Group</th>
<th>Familiarity task</th>
<th>Repeat detection task</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% Faces identified</td>
<td>Vigilance</td>
</tr>
<tr>
<td>Same</td>
<td>29.0 (13.6)</td>
<td>58.3 (17.0)</td>
</tr>
<tr>
<td>Different</td>
<td>32.1 (11.3)</td>
<td>32.8 (12.2)</td>
</tr>
</tbody>
</table>

*Table S1.* Group mean (SD) percentages of faces identified in the familiarity task, hit rates for vigilance repeats, and false alarms for fillers in the repeat detection task.

![Figure S1](image-url)  
*Figure S1.* Individual data for data checks. Dots are jittered horizontally to improve visibility. Left panel shows the number of familiar faces is comparable across groups. Right panel shows that hit rates for vigilance repeats is higher for ‘same’ than ‘different’ groups, whereas false alarms for fillers are similar. Error bars show 95% CI around the mean.
Main analysis. We compared detection hit rates in each group for target faces that participants reported as familiar and unfamiliar. Results are presented in Table S2 and Figure S2. Familiar faces were detected more accurately ($M = 62.1\%$) than unfamiliar faces ($M = 28.0\%$), reflected by the main effect of familiarity, $F(1, 43) = 256.57, p < .001, \eta^2_p = .856, BF_{10} > 1000$. There was also a main effect of image, $F(1, 43) = 6.38, p = .015, \eta^2_p = .129, BF_{10} = 3.54$, as repeat detection was better for the ‘same’ ($M = 44.4\%$) than the ‘different’ ($M = 32.3\%$) images. Critically, these main effects were qualified by a familiarity by image interaction, $F(1, 43) = 5.79, p = .020, \eta^2_p = .119, BF_{10} = 2.84$. While unfamiliar faces were better detected with ‘same’ than ‘different’ images, $t(42.77) = 3.76, p < .001, d_s = 1.15, BF_{10} = 52.10$, familiar faces were detected equally well regardless of image, $t(41.28) = 0.99, p = .327, d_s = 0.30$, although Bayes factors provided only weak evidence for this null effect, $BF_{10} = 0.44$. This result shows that recognition of familiar faces is robust to image change, but recognition of unfamiliar faces is poorer when the target image is different than the initial face.

<table>
<thead>
<tr>
<th>Group</th>
<th>% Hits</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% Hits</td>
<td>Familiar faces</td>
<td>Unfamiliar faces</td>
</tr>
<tr>
<td>Same</td>
<td>64.5 (15.9)</td>
<td>35.5 (14.0)</td>
<td></td>
</tr>
<tr>
<td>Different</td>
<td>59.6 (17.0)</td>
<td>20.4 (13.1)</td>
<td></td>
</tr>
</tbody>
</table>

*Table S2.* Group mean (SD) hit rates for detection of familiar and unfamiliar targets in the repeat detection task.

*Figure S2.* Individual data for correct detection of target repeats by familiarity and group. Dots are jittered horizontally to improve visibility. Familiar faces were detected equally well by both groups but unfamiliar faces were detected better by ‘same’ than by ‘different’ group. Error bars show 95% CI around the mean.
Can our results be accounted for by the fact that the ‘different’ image condition is more difficult than the ‘same’ image condition? To address this issue, we restricted our analysis to a subset of participants in the ‘same’ condition who performed below the group median on vigilance repeats (≤58.0%, n = 12), and participants in the ‘different’ condition who performed above the group median (≥32.0%, n = 13). This subset was equated on vigilance performance (M = 42.2%, SD = 10.8), t(21.48) = 1.23, p = .233, d_s = 0.56, BF_{10} = 0.64, and were familiar with a similar number of faces (M = 28.9%, SD = 12.0), t(20.99) = 1.72, p = .101, d_s = 0.79, BF_{10} = 1.07. Results are shown in Figure S3. We found a main effect of familiarity, F(1, 23) = 179.37, p < .001, η^2_p = .886, BF_{10} > 1000, but not image, F(1, 23) = 0.20, p = .656, η^2_p = .009, BF_{10} = 0.40. Crucially, the familiarity by image interaction holds, F(1, 23) = 9.88, p = .005, η^2_p = .300, BF_{10} = 9.29. As in the full sample, unfamiliar faces were detected better with ‘same’ (M = 32.4%) than ‘different’ images (M = 22.7%), t(22.84) = 2.07, p = .049, d_s = 0.95, BF_{10} = 1.66. In contrast, familiar faces were detected equally well with ‘same’ (M = 57.4%) and ‘different’ images (M = 63.0%), t(22.30) = 0.97, p = .344, d_s = 0.44, BF_{10} = 0.51. Although Bayes factors provide only weak evidence in support of these conclusions, the pattern is similar to that in the full sample; even when both groups performed equally on vigilance trials, familiar face recognition remained robust to changes in the image, while unfamiliar face recognition was poorer when the repeated images were different.

**Figure S3.** Individual data for correct detection of target repeats by familiarity and group in a subset of participants who were equated for performance on vigilance repeats. Dots are jittered horizontally to improve visibility. Participants in the ‘different’ group suffered in detection for unfamiliar faces more so than those in the ‘same’ group. Error bars show 95% CI around the mean.