White, Burton, Jenkins & Kemp "Redesigning photo-ID to improve unfamiliar face matching performance": Supplementary materials

For the reasons outlined in the manuscript, our main dependent variables were accuracy scores on match and mismatch trials. Here, for the interested reader, we provide Signal Detection Theory statistics (A' and B") for all experiments, and analysis of additional dependent variables of reaction time and confidence data that were collected in Experiment 2. We also include details of the procedure used to collect similarity rating data in Experiment 2. These data were collected and analysed in an attempt to clarify the cognitive process underlying the image-array advantage reported in this experiment. Finally, we include details of an experiment that was described in a previous version of this manuscript. The results of this study were essentially a replication of Experiment 2 and so were not included is the paper (see Replication Experiment, Pg 6).

EXPERIMENT 1

Two-way ANOVA for sensitivity scores (*A*') showed significant main effects of both Familiarity [*F* (1,87) = 106, $p < 0.01$] and Image Type [*F* (1,87) = 8.44, $p < 0.01$], but no significant interaction ($F < 1$). Bias data (*B*") also showed main effects of Familiarity [*F* (1,87) = 35.2, $p < 0.01$], and Image Type [*F* (1,87) = 11.2, $p < 0.01$], but no significant interaction [*F* (1,87) = 3.30, $p > 0.05$].

	Unfamiliar		**Familiar**	
	Single Photo	*Average*	*Single Photo*	*Average*
A'	.842 (*.093*)	.868 (*.068*)	.938 (*.083*)	.952 (*.062*)
B''	.304 (*.458*)	.051 (*.465*)	-.239 (*.514*)	-.148 (*.571*)

Table 1. Signal Detection Measures for the face matching task in Experiment 1 (Standard Deviations in parenthesis)

EXPERIMENT 2

Signal Detection Statistics

Summary statistics for sensitivity (*A*') and bias (*B*") measures in Experiment 2 are shown in Table 2. ANOVA for sensitivity scores shows a main effect of Array Size

[*F* (3,69) = 10.9, $p < 0.01$], but no significant effect of Study Time [*F* (1,69) = 1.49, $p < 0.01$], and no interaction between these factors ($F < 1$). Bias data also show a main effect of Array Size [*F* (3,69) = 16.5, $p < 0.01$], no significant effect of Study Time ($F < 1$) and a no interaction [*F* (6, 207) = 2.04, $p > 0.05$].

		Array Size			
		1	*2*	*3*	*4*
A'	3 seconds	.863 (*.059*)	.902 (*.076*)	.891 (*.079*)	.918 (*.059*)
	6 seconds	.865 (*.062*)	.891 (*.053*)	.900 (*.056*)	.890 (*.058*)
	9 seconds	.890 (*.061*)	.913 (*.054*)	.912 (*.052*)	.926 (*.060*)
B''	3 seconds	.113 (*.328*)	-.082 (*.382*)	-.269 (*.487*)	-.392 (*.455*)
	6 seconds	-.010 (*.455*)	-.093 (*.487*)	-.340 (*.551*)	-.157 (*.563*)
	9 seconds	.059 (*.515*)	-.263 (*.473*)	-.239 (*.549*)	-.427 (*.447*)

Table 2. Signal Detection Measures for the face matching task in Experiment 2 (Standard Deviations in parenthesis)

Confidence Ratings

Confidence data for correct trials Experiment 2 are shown in Table 3. For correct responses, a three-way ANOVA revealed a significant main effect of Array Size [*F* (3,69) = 47.7, $p < 0.05$], but no significant main effect of Trial Type or Study Time (*F*s < 1). The three-way interaction between these factors was non-significant [*F* (12, 207) = 1.40, $p > 0.05$], as were the two-way interactions between Trial Type and Study Time [*F* (2, 69) = 1.25, $p > 0.05$], and between Array Size and Study Time [*F* (6, 207) = 1.86, $p > 0.05$]. Thus, the Study Time manipulation had no significant effect on performance.

The interaction between Trial Type and Array Size was significant [*F* (3, 207) = 11.0, $p < 0.05$]. Simple Main Effects revealed a significant effect of Array Size for match trials [*F* (3, 213) = 19.5, $p < 0.05$], but not mismatch trials [*F* (3,213) = 2.39, $p > 0.05$]. The effect of Trial Type was significant for single-photo arrays [*F* (1,71) = 7.34, $p < 0.05$, $d = 0.644$], two-photo arrays [*F* (1,71) = 37.4, $p < 0.05$, $d = 1.45$], three-photo arrays [*F* (1,71) = 48.0, $p < 0.05$, $d = 1.65$], and four-photo arrays [*F* (1,71) = 86.7, $p < 0.05$, $d = 2.21$], with participants giving higher confidence ratings for match trials.

We carried out planned comparison t-tests between successive Array Sizes to determine if confidence increased as a function of Array Size. For match trials, mean confidence was higher for single photos than for two-photo arrays [t (71) = 4.70, $p <$ 0.05, d = 0.350], but there were no differences between either two-photo and three-photo arrays [t (71) = 1.26, $p >$ 0.05, d = 0.350], or three-photo and four-photo arrays ($t < 1$). For mismatch trials, mean confidence was higher for single photos than for two-photo arrays [t (71) = 4.70, $p <$ 0.05, d = 0.350], but there were no differences between either two-photo and three-photo arrays [t (71) = 1.26, $p >$ 0.05, d = 0.085], or three-photo and four-photo arrays ($t < 1$). For incorrect trials, there were a large proportion of missing trials (i.e. for cells where participants did not make incorrect responses), which precluded reliable analysis.

Response Latency

Response latency data for correct trials in Experiment 2 are shown in Table 3. A three-way ANOVA revealed a significant main effect of Trial Type with faster responses for 'match' decisions than for 'mismatch' decisions [F (1, 69) = 47.7, $p <$ 0.05]. There were no significant main effects of Array Size [F (3, 69) = 1.76, $p >$ 0.05] or Study Time [F (2, 69) = 2.18, $p >$ 0.05]. The three-way interaction between these factors was not significant [F (12, 207) = 1.06, $p >$ 0.05], and neither were the two-way interactions between Trial Type and Study Time [F (2, 69) = 1.16, $p >$ 0.05], or between Array Size and Study Time [F (6, 207) = 1.69, $p >$ 0.05]. Thus, Study Time did not affect matching performance on this measure.

The interaction between Trial Type and Array Size was significant [F (3, 207) = 6.26, $p <$ 0.05]. Simple Main Effects revealed significant effects of Array Size for both match trials [F (3, 213) = 3.64, $p <$ 0.05], and mismatch trials [F (3, 213) = 4.27, $p <$ 0.05]. The effect of Trial Type on response latency was non-significant for single-photos ($F < 1$), but was significant for two-photo arrays [F (1,71) = 19.0, $p <$ 0.05, d = 1.034], three-photo arrays [F (1, 71) = 19.4, $p <$ 0.05, d = 1.047], and four-photo arrays [F (1, 71) = 8.57, $p <$ 0.05, d = 0.696], with participants taking longer to respond in mismatch trials.

		Match Trials				Mismatch Trials			
		1-photo	2-photo	3-photo	4-photo	1-photo	2-photo	3-photo	4-photo
Response Latency (ms)	*3 seconds*	2291 (2106)	1616 (1142)	1531 (1153)	2019 (1903)	1905 (1281)	2280 (1901)	2280 (1249)	2045 (1047)
	6 seconds	1316 (970)	1160 (415)	1327 (874)	1235 (972)	1569 (1205)	1749 (1051)	1927 (2128)	2404 (1739)
	9 seconds	1242 (832)	1142 (561)	1166 (849)	1158 (495)	1237 (514)	1784 (1435)	1894 (1699)	1747 (1135)
	Overall	1660 (1475)	1318 (791)	1384 (968)	1525 (1301)	1595 (1079)	1934 (1502)	2154 (1715)	2081 (1355)
Confidence (correct)	*3 seconds*	73.8 (11.4)	79.6 (9.5)	79.0 (11.9)	79.6 (10.7)	73.4 (9.5)	72.3 (11.4)	72.5 (9.9)	72.2 (10.6)
	6 seconds	75.5 (15.1)	76.9 (13.6)	79.6 (12.7)	80.9 (12.2)	71.9 (15.6)	70.9 (14.2)	71.0 (14.7)	68.6 (15.8)
	9 seconds	79.3 (10.5)	84.7 (9.5)	85.6 (9.0)	84.6 (8.6)	72.9 (11.9)	72.1 (15.0)	76.5 (10.8)	72.4 (12.0)
	Overall	76.1 (12.5)	80.3 (11.4)	81.2 (11.6)	81.6 (10.6)	72.7 (12.4)	71.8 (13.4)	73.3 (12.1)	71.1 (12.9)
Confidence (incorrect)	*3 seconds*	63.8 (20.8)	65.0 (18.9)	63.9 (10.7)	58.9 (17.8)	59.4 (15.1)	51.4 (15.0)	58.6 (26.5)	66.9 (16.0)
	6 seconds	56.4 (18.1)	64.3 (19.4)	58.0 (20.9)	66.3 (18.5)	59.5 (17.0)	59.5 (22.3)	62.1 (29.1)	53.8 (24.1)
	9 seconds	56.1 (17.1)	65.7 (13.4)	64.6 (11.1)	61.1 (19.1)	61.9 (16.4)	61.5 (15.1)	61.2 (22.0)	53.5 (33.1)
	Overall	58.1 (16.6)	53.2 (19.9)	55.7 (26.3)	59.6 (26.0)	53.9 (20.7)	61.0 (17.5)	62.6 (17.2)	61.1 (20.9)

Table 3. Mean confidence and response latency data for the face matching task in Experiment 2 (Standard Deviations in parenthesis)

	Match Trials				Mismatch Trials			
	1-photo	2-photo	3-photo	4-photo	1-photo	2-photo	3-photo	4-photo
Response Latency	3453 (1093)	3423 (1054)	3442 (1034)	3852 (1270)	3521 (1355)	4101 (1608)	4259 (1396)	5219 (2150)
Confidence (correct)	73.6 (10.2)	77.2 (10.6)	78.9 (9.7)	78.7 (10.4)	70.2 (12.8)	69.3 (15.0)	71.4 (14.6)	68.3 (14.1)
Confidence (incorrect)	56.3 (21.8)	57.0 (21.7)	54.2 (25.4)	56.6 (23.5)	63.8 (18.1)	64.4 (16.6)	66.0 (16.8)	62.5 (14.6)

Table 7. Mean confidence and response latency data for the Replication Experiment (Standard Deviations in parenthesis)

Similarity rating analysis

In Experiment 2, the effect of multi-photo arrays on face matching shows that providing more samples of variation in appearance improves face matching accuracy. However, as we note in the manuscript (Experiment 2, Results section) there are two different classes of process that could account for this benefit. On the one hand, participants could base their decision on the array photo that is most similar to the target, with the benefit of multiple photographs arising simply from the increased likelihood that a similar image will be found (pairwise match). An alternative explanation is that the viewer extracts an abstract representation from the array photos (abstractive match).

We attempted to distinguish between these alternative accounts by collecting similarity ratings for each array photo and target image pair. Using these ratings, we conducted a post-hoc analysis of the performance data in Experiment 2, to determine the relative extent to which trial performance (i.e. correct/ incorrect) was predicted by i) similarity of the *Most Similar* array photo to the target/foil image, and ii) the *Average* similarity of the array photos to the target/foil image.

Participants

Twenty-eight undergraduate students (17 female, mean age 19.3 years, SD = 2.1) from the University of New South Wales participated in return for course credit.

Design and Procedure

For each identity in Experiment 2, we collected pairwise ratings of similarity from the target image (Match trials) to each of the array photos, and also from the foil identity to each of the array photos (Mismatch trials). Image pairs were presented in a different random order for each participant (blocked by identity), and participants rated the similarity of two images on a scale labeled from 1 (Very Dissimilar) to 100 (Very Similar). Due to the large number of pair wise comparisons (800 in total), participants each provided ratings to half of the identities.

Analysis and Discussion

For each trial in Experiment 2 (72 participants x 160 trials = 11, 520 trials in total), we calculated i) the similarity of the Most Similar array photo to the target/foil, and ii) the Average similarity of the array to the target/foil. We then calculated the extent to which each of these measures was diagnostic of trial performance (correct/incorrect). To do this, we calculated ROC curves for each Array Size, with similarity ratings as input parameter, and trial outcome (correct/incorrect) as the classification variable. We then calculated the Area Under the ROC Curves (AUC) for each condition, conducting this analysis separately for Match and Mismatch trials (see Table 4).

	Array Size			
	1	*2*	*3*	*4*
Most Similar	.776 (*.015*)	.743 (*.021*)	.779 (*.022*)	.757 (*.024*)
Average	.776 (*.015*)	.774 (*.019*)	.783 (*.022*)	.761 (*.023*)
Most Similar	.601 (*.020*)	.670 (*0.17*)	.706 (*0.16*)	.550 (*.026*)
Average	.601 (*.020*)	.683 (*.017*)	.706 (*.016*)	.546 (*.025*)

Table 4. Area Under ROC Curve (AUC) values for the post-hoc analysis, representing the diagnostic value of array-to-photo similarity measures in predicting trial accuracy from Experiment 2 (bootstrapped standard error in parenthesis). Higher AUC scores represent better prediction of trial accuracy (chance = .50)

Overall, this analysis shows that similarity scores predicted match trial performance (hits) better than mismatch trial performance (false positives). However, comparison between the AUC scores for most similar image similarity and average array similarity show no obvious advantage for either measure in predicting trial accuracy. Moreover, it may be impossible to distinguish between these competing accounts based on the current data, because Most Similar and Average similarity ratings were highly correlated (pooled Spearman's $r = 0.91$). Therefore, we recommend that future research addresses this question directly, by designing photo arrays in such a way that Average and Most Similar models generate opposing predictions of performance behavior.

EXPERIMENT 3

Summary statistics for sensitivity (*A*') and bias (*B*") measures in Experiment 3 are shown in Table 5. ANOVA for sensitivity scores shows a significant main effect of

Familiarity, [*F* (1,87) = 40.5, *p* < 0.01], but no significant main effect of Image Type and no interaction (Fs < 1). Bias data also show a main effect of Familiarity [*F* (1,87) = 18.6, *p* < 0.01], but no significant main effect of Image Type (*F* < 1), and a non-significant interaction [*F* (1,87) = 3.43, *p* = 0.069].

	Unfamiliar		**Familiar**	
	4-Photo Array	*Average*	*4-Photo Array*	*Average*
A'	.866 (*.088*)	.850 (*.084*)	.934 (*.087*)	.934 (*.088*)
B"	.047 (*.449*)	.172 (*.488*)	-.162 (*.424*)	-.221 (*.506*)

Table 5. Signal Detection Measures for the face matching task in Experiment 3 (Standard Deviations in parenthesis)

REPLICATION EXPERIMENT

Prior to Experiment 2, we conducted a very similar study that did not include the Study Time factor. The results of these two studies were consistent, but the interest of the earlier study was diminished in light of the results of Experiment 2, and so was removed from the manuscript during the review process. For the benefit of the reader, we include details of this earlier experiment below.

Method

Participants

Thirty-four undergraduates (23 Female) from the University of New South Wales participated in the study. The average age of the sample was 19.2 years (sd = 2.8).

Stimuli, Design and Procedure

Stimuli and procedural details were identical to Experiment 2, except that in this study we did not manipulate study duration. Instead, the task was self-paced and participants were asked to be as accurate as possible. As in Experiment 2, Trial Type (Match, Mismatch) and Array Size (1image, 2images, 3images, 4 images) were manipulated within-subjects.

Results

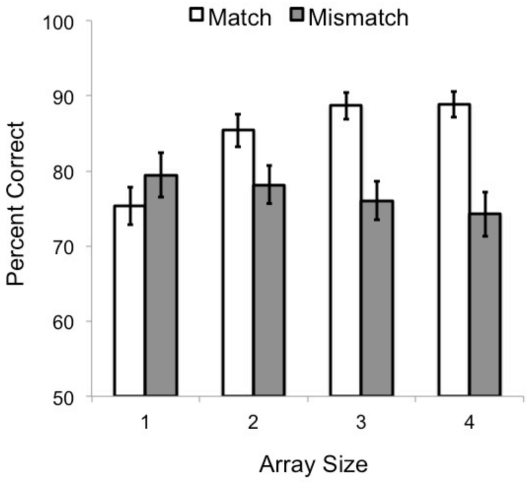

Figure 1. Mean percent correct for face matching in Experiment 2 (± standard error).

Accuracy data for the Replication Experiment are shown in Figure 1. For accuracy data, a two-way ANOVA revealed main effects of both Trial Type [*F* (1,33) = 5.91, *p* < 0.05], and Array Size [*F* (3,99) = 5.44, *p* < 0.05], and the interaction between these factors was also significant [*F* (3,99) = 10.75, *p* < 0.05]. Analysis of Simple Main Effects showed that the effect of Array Size was significant for Match trials [*F* (3,99) = 20.52, *p* < 0.05], but was non-significant for Mismatch trials [*F* (3,99) = 1.65, *p* > 0.05]. In addition, the effect of Trial Type was non-significant for single-photo arrays (*F* < 1), but significant for two-photo [*F* (1,33) = 4.34, *p* < 0.05], three-photo [*F* (1,33) = 14.4, p < 0.05], and four-photo arrays [*F* (1,33) = 16.5, p < 0.05].

As in Experiment 2, we carried out planned comparisons to explore the effect of Array Size. Overall accuracy was 77.4% (SD = 8.3) for one-photo arrays, 81.8% (SD = 9.3) for two-photo arrays, 82.4% (SD = 8.6) for three-photo arrays, and 81.6% (SD = 9.3) for four-photo arrays. Planned comparisons revealed a significant difference between one-photo and two-photo arrays [*t* (33) = 2.98, *p* < 0.05, *Cohen's d* = 0.497], but no differences between two-photo and three-photo arrays (*t* < 1, *d* = 0.067) or between three-photo and four-photo arrays (*t* < 1, *d* = -0.089).

Signal Detection Statistics

Summary statistics for sensitivity (*A'*) and bias (*B*") measures in Experiment 2 are shown in Table 6. Sensitivity data were consistent with accuracy data, showing a significant main effect of Array Size [F (3,99) = 4.74; p < 0.05]. Planned comparisons show significant improvements between one-photo and two-photo arrays [*t* (33) = 2.32; *p* < 0.05, *Cohen's d* = 0.411], but no differences between two-photo and three-photo arrays (*t* < 1, *d* = 0.089) or between three-photo and four-photo arrays (*t* < 1, *d* = 0.062). Bias data also show a significant main effect of Array Size [F (3,99) = 11.7; p < 0.05], with participants more likely to respond 'same' for larger array sizes. Again, planned comparisons show a significant difference between one-photo and two-photo arrays [*t* (33) = 3.79; *p* < 0.05, *d* = 0.661], but no differences between two-photo and three-photo arrays [*t* (33) = 1.27; *p* > 0.05, *d* = 0.211] or between three-photo and four-photo arrays [*t* < 1, *d* = -0.025].

	Array Size			
	1	*2*	*3*	*4*
A'	.858 (*.070*)	.887 (*.071*)	.893 (*.064*)	.889 (*.065*)
B"	.149 (*.507*)	-.186 (*.507*)	-.289 (*.469*)	-.278 (*.428*)

Table 6. Signal Detection Measures for the face matching task in Experiment 2 (Standard Deviations in parenthesis)

Confidence Ratings

Confidence data for the Replication Experiment are shown in Table 7. For correct responses, a two-way ANOVA revealed a significant main effect of Array Size [*F* (3, 99) = 5.01, *p* < 0.05], and a significant main effect of Trial Type [*F* (1, 33) = 18.7, *p* < 0.05]. The interaction between these factors was significant [*F* (3, 99) = 5.10, *p* < 0.05]. Simple Main Effects revealed a significant effect of Array Size for match trials [*F* (3, 99) = 7.75, *p* < 0.05], but not mismatch trials [*F* (3, 99) = 2.37, *p* > 0.05]. The effect of Trial Type was non-significant for single-photo arrays [*F* (1, 33) = 2.79, *p* < 0.05], but was significant for two-photo arrays [*F* (1, 33) = 15.3, *p* < 0.05], three-photo arrays [*F* (1, 33) = 13.6, *p* < 0.05], and four-photo arrays [*F* (1, 33) = 26.6, *p* < 0.05], with higher confidence ratings for match trials.

As in Experiment 2, we carried out planned comparison t-tests between successive Array Sizes to determine if confidence increased as a function of Array Size. For match trials, mean confidence was higher for single photos than for two-photo arrays [t (33) = 2.77, $p < 0.05$, $d = 0.346$], but there were no differences between either two-photo and three-photo arrays [t (33) = 1.59, $p > 0.05$, $d = 0.167$], or three-photo and four-photo arrays ($t < 1$, $d = 0.020$). For mismatch trials, differences between single photos and two-photo arrays ($t < 1$, d = 0.065), and between two-photo and three-photo arrays [t (33) = 1.51, $p > 0.05$, $d = 0.142$] were non-significant, however confidence on three-photo arrays was significantly *higher* that for four-photo arrays [t (33) = 2.68, $p < 0.05$, $d = 0.216$]. For incorrect trials, there were a large proportion of missing trials (i.e. for cells where participants did not make incorrect responses), which precluded reliable analysis.

Response Latency

Response latency data for correct trials in the Replication Experiment are shown in Table 7. A two-way ANOVA revealed a significant main effect of Trial Type with faster responses for 'match' decisions than for 'mismatch' decisions [F (1, 33) = 11.8, $p < 0.05$], and a significant main effect of Array Size [F (3, 99) = 19.6, $p < 0.05$], with longer response latencies for larger arrays. The interaction between these factors was also significant [F (3, 99) = 5.07, $p < 0.05$].

Simple Main Effects revealed a non-significant effect of Array Size for match trials [F (3, 99) = 2.00, $p > 0.05$], and a significant effect for mismatch trials [F (3, 99) = 22.3, $p < 0.05$]. The effect of Trial Type on response latency was non-significant for single-photos ($F < 1$), but was significant for two-photo arrays [F (1, 33) = 5.82, $p < 0.05$], three-photo arrays [F (1, 33) = 8.46, $p < 0.05$], and four-photo arrays [F (1, 33) = 23.7, $p < 0.05$], with participants taking longer to respond in mismatch trials.

David White[1], A. Mike Burton[2], Rob Jenkins[3] & Richard I. Kemp[1]

[1] School of Psychology, The University of New South Wales, Australia (david.white@unsw.edu.au)

[2] School of Psychology, University of Aberdeen, UK

[3] Department of Psychology, University of York, UK