A new Mooney test.

R J Verhallen[a, 1] & J D Mollon[a]

[a] Department of Experimental Psychology, Downing Street, Cambridge CB2 3EB, United Kingdom

Abstract
Since its introduction in 1957, the Mooney test has continued to see active use in studies of visual perception, in studies using brain imaging, and in clinical research. Mooney's original version is of limited length however, and is designed to be administered by time-consuming personal interview. We have developed a new, extended version of the Mooney test suitable for online testing and suitable for use in a test–retest paradigm. The Mooney–Verhallen Test (MVT) comprises 144 trials, takes on average less than 10 minutes to complete, and has a Spearman–Brown corrected test–retest reliability of $\rho = .89$. We outline our methods for developing the stimuli and for selecting the final stimulus set, and we present results from two rounds of testing on two independent samples of 374 participants and 505 participants, respectively. The test is freely available for scientific use.

Keywords
Mooney test, closure, two-tone images, gestalt perception, face perception, individual differences

Acknowledgements
We thank Jenny Bosten for valuable advice and insights during the development of materials as well as during the testing phases, and thank John Rust and Ken Nakayama for fruitful discussions. We also thank Christina Hu for helping with part of the data collection during the first testing phase. RJV is grateful for funding received from the Prins Bernhard Cultuurfonds, the Hendrik Muller Fonds, the Grindley Fund, and the Cambridge Philosophical Society.

[1] *Corresponding author at: University of Cambridge, Department of Experimental Psychology, Downing Street, Cambridge CB2 3 EB, United Kingdom. Tel.: +44 1223 333 580.*
E-mail address: rjv31@cam.ac.uk (R.J. Verhallen).

1. Introduction

Perception of so-called Mooney images is all or none: the black and white blobs either combine into a complete face in a single percept – often within a few hundred milliseconds – or remain fully independent and abstract (Mooney, 1957). The Mooney test has seen frequent use both in neuropsychological studies and in studies of face processing, and it is referred to as the 'Mooney closure test', 'Mooney face test', or simply 'Mooney test' (Bruce & Young, 2012; Busigny et al., 2010; Kanwisher et al., 1998; Lansdell, 1968; Milner et al., 1968; Verhallen et al., 2014; Wasserstein et al., 2004). Within a normal population, there are substantial individual differences in the ability to perceive Mooney faces (Foreman, 1991; Verhallen et al, 2014).

However, Mooney's original version of the test is short (40 items), is designed to be administered by personal interview, and is not suited for test–retest estimates of reliability. Moreover, the image set is heterogeneous and shows its age (the images were created from 1950's magazine clippings). To overcome these limitations we set out to construct – from scratch – a new, online, and extended version of the Mooney test. This new test measures the ability to detect a Mooney face from among two distractors, by asking participants to click on one of the eyes of the face.

There is abundant evidence that Mooney images engage the mechanisms of face perception. The N170 component of the event-related potential, which occurs specifically in response to the image of a face (Bentin et al., 1996), is also observed in response to the presentation of Mooney faces; and when the Mooney face is consciously perceived, the amplitude of this component is increased (George et al., 2005; Jeffreys, 1989; Jemel et al., 2003; Latinus & Taylor, 2005). Furthermore, the highly face-selective fusiform face area shows increased activity upon the conscious perception of a Mooney face, as compared to the failure to perceive it (Andrews & Schluppeck, 2004; Kanwisher et al., 1998; Rossion et al., 2011).

However, do the large *individual differences* on the Mooney test arise from differences in the specific processes of face perception or do they rather reflect differences in 'closure' – a process of perceptual organisation that precedes perception of the face? It is curious, for example, that males outperform females on the Mooney test (Foreman, 1991; Verhallen et al., 2014), whereas, if a sex difference is observed in other tests of face processing, it is in favour of females (Megreya et al., 2011). Moreover, we have found (Verhallen et al, in preparation) that performance on a 3AFC version of the original Mooney test does not correlate very strongly (Spearman's $\rho = .21$) with performance on a test of face discrimination (the Glasgow Face Matching Test; Burton et al., 2010) and correlates only modestly (Spearman's $\rho = .31$) with performance on a test of face recognition (the Cambridge Face Memory Test; Duchaine & Nakayama, 2006).

Nevertheless, it is theoretically difficult to separate the processes of perceptual organisation from those of face processing in the case of Mooney images. When more conventional stimuli are presented, it is possible to envisage the two processes as sequential: perceptual organisation may be driven by low-level features such as similarity, proximity and the presence of T-junctions, and then object recognition may follow. But in the two-tone Mooney images most of the low-level, Gestalt-prompting features are absent. There are no T-junctions, for example (Moore & Cavanagh, 1998); and the individual features of a face – eyes, nose, mouth – are seldom independently apparent in a Mooney image. To detect the face, the observer must construct a specific, three-dimensional model both of the face and of the lighting. The perception is of concave and convex regions, with cast and attached shadows. The underlying processes are likely to be top-down and they must surely draw upon the observer's stored knowledge of faces, acquired over a lifetime.

With the advent and improvement of software for the manipulation of digital images, the conversion of photographs into two-tone 'Mooney' images has become rather easy. Several authors have created their own Mooney stimuli for specific experimental purposes (e.g., Jemel et al., 2003; McKeeff & Tong, 2007; Rossion et al., 2011). Our own purpose was to create a standardised, online, quick and reliable new version of the Mooney test, for use by the wider academic community. However, as with any psychological test, it is not trivial to develop a reliable and internally balanced Mooney test. We describe our method of creating the stimulus set, and report the results of two testing phases: an initial selection phase (316 trials, N = 374) to gather data to use subsequently in selecting the final stimulus set; and a testing phase using the final stimulus set (144 trials, N = 505) in order to establish test–retest reliability and to gather population statistics.

Performance on tests of face processing is known to depend on the race and the sex of the faces used as stimuli, as well as on the race and the sex of the observer looking at the faces. For example, the 'Other-Race Effect' describes the impaired recognition of faces of people belonging to a different race relative to recognition of faces of the participant's own race (Meissner & Brigham, 2001). It remains unclear to what extent the Mooney test taps face recognition ability (see above); nevertheless we limited ourselves to Caucasian faces in the development of our test, although our participant sample was not confined to Caucasians. Since the female advantage in face processing studies tends to be restricted to female faces (Megreya et al., 2011; Rehnman & Herlitz, 2007; Sommer et al., 2013), we included an equal number of female and male face stimuli in our test.

2. Development of test materials

2.1 Volunteers for face photographs

158 Caucasian volunteers (50% female; mean age 28 years, ranging from 20 to 80 years) were recruited from the Cambridge, UK area via social media, online notice boards, and electronic mailing lists. In exchange for their help, volunteers were offered the possibility of having a professional studio portrait taken of themselves, for their own, unrestricted use. Ethical permission for the study was given by the Cambridge University Psychology Research Ethics Committee.

2.2 Materials and procedure

We invited each volunteer to our photography studio, where we took their portrait from different angles and while varying the direction and brightness of the studio lamps. Volunteers did not wear any items that would occlude the face or parts thereof, such as glasses, hats, etc. The background was either black or white, and these two options were alternated across volunteers. In Adobe Photoshop, a Gaussian blur of 10 pixels was applied to all photographs to slightly reduce the amount of detail in the image. The photographs were subsequently converted to two-tone 'Mooney images' using a threshold procedure: any pixels with a luminance above the threshold value were converted to pure white, and any pixels with a luminance below were converted to pure black. A fixed threshold value of 110 – from the range 0 (white) to 255 (black) – was used for all photographs, though variation remained in the ratio of black to white areas in the resulting Mooney images, owing to the variability in studio lighting, poses, and skin tone. The latter variability was desired, since it gave rise to a diverse set of target images. For each volunteer, we selected two images and cropped them to limit the contextual information. We thus created 316 target images in total. Distractor images were created through six custom-made procedures in Adobe Photoshop: various combinations of rotation, polarity inversion, and superposition of the target image (see Figure 1A). Thus, for each target image, the six custom-made procedures yielded six distractor images, from which we selected two to accompany the target image. In this way, we created 316 three-alternative forced-choice (3AFC) test items (see Figure 1B); Figure 2 shows an additional three test items.

In a previous study (Verhallen et al., 2014) – using the original forty Mooney stimuli (Mooney, 1957) in a 3AFC paradigm – we observed a marked ceiling effect in performance, in that almost 10% of participants (total N = 397) reached the maximal score. We hypothesised that participants were sometimes able to respond correctly using cues other than the actual percept of the face, for example the extent to which the shapes in the images were potentially organic and therefore likely to be part of a face. Alternatively, the participant might try to rule out which two images were not organic or face-like. Thus, participants could respond 'correctly' without perceiving the face. In order to avoid this in our new test, we asked participants to respond not just by clicking on the panel (out of three) that showed a face, but rather by clicking on either of the eyes of the face. We quantified responses by dividing the target image into a grid of 6 by 9 squares: horizontally, the eyes fell always within columns 2 to 5 inclusive, and vertically, the eyes fell only in one of the four rows C, D, E, or F (see Figure). We used the original photograph to determine the correct eye region for each item. Participants were not informed about this method of quantifying response, or about the fact that there were only four possible correct regions, but were merely instructed to find the face and to click on either of the eyes.

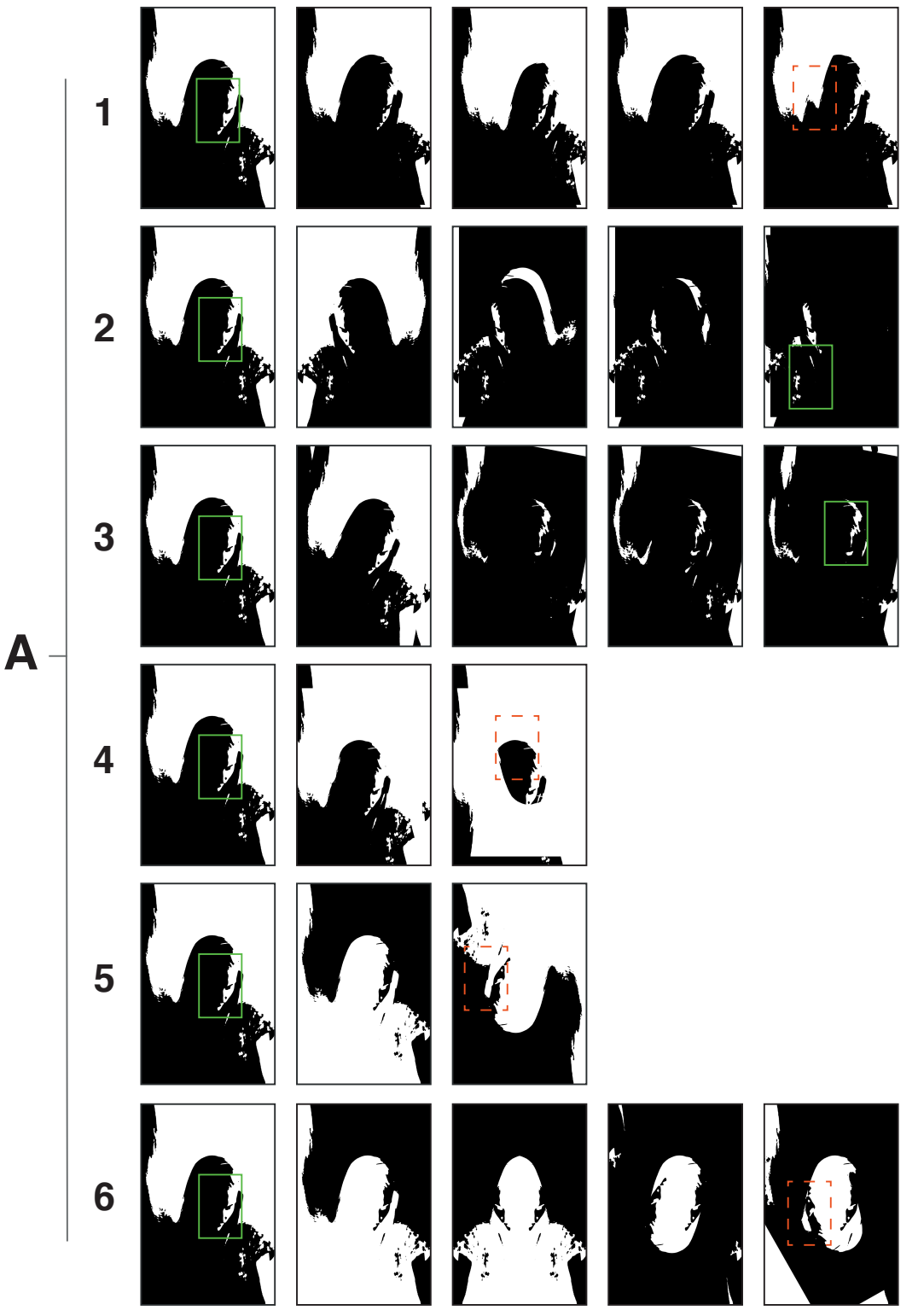
A
1
2
3
4
5
6
B

Figure 1 (see previous page). *Panels A1–A6:* the individual steps (from left to right) for each of our six custom-made procedures to create distractor images. All distractor images were created from the original, non-cropped target image (furthest left image of each row) in order to have more information to work with. The six custom-made procedures yielded six distractor images for each target image: green, solid rectangles indicate the final crop of the two distractor images selected for the 3AFC stimulus (rows 2 & 3; see panel B for the final stimulus triplet); red, dashed rectangles indicate the final crop of the distractor images not selected (rows 1, 4, 5 and 6). The sequence of manipulations for each procedure was as follows: *row 1:* overlay a copy of the original image (using only the black parts of the original image), and translate it rightward by 115 pixels and upward by 130 pixels (all original images were 4,032 pixels wide and 6,048 pixels high); overlay another copy of the original image (using only the white parts), and translate it rightward by 488 pixels and downward by 110 pixels; overlay another copy of the original image (using only the black parts), and translate it rightward by 488 pixels and upward by 410 pixels; overlay another copy of the original image (using only the black parts), and translate it rightward by 603 pixels and upward by 540 pixels — *row 2:* flip original image horizontally; overlay a (horizontally-flipped) copy of the original image using 'subtraction' (any white areas overlapping a white area become black; any black areas overlapping a white area become white; any black areas overlapping a black area remain black), and translate it rightward by 213 pixels and upward by 316 pixels; overlay another copy of the original image (though not flipped horizontally) using 'subtraction'; translate this last copy leftward by 331 pixels and downward by 263 pixels — *row 3:* overlay a copy of the original image and rotate it clockwise by 11°; overlay another copy of the original image (using only the black parts) and rotate it clockwise by 2°; overlay another copy of the original image and delete all white areas adjoining the white area at the top of the original image; flip this second copy vertically, and translate it rightward by 353 pixels and upward by 3,145 pixels — *row 4:* overlay a copy of the original image (using only the white parts) and translate it leftward by 253 pixels and downward by 726 pixels; overlay another copy of the original image (using only the white parts), flip it vertically, and translate it rightward by 149 pixels and upward by 270 pixels — *row 5:* invert image polarity; rotate the image 180° — *row 6:* invert image polarity; overlay a copy of the (polarity inverted) original image (using only the black parts), and flip it horizontally; flip the overlaid copy vertically; rotate the overlaid copy counter-clockwise by 29°.

We manually cropped both the final distractor images (furthest right image of each row) as well as the target image, as indicated by the overlaid rectangles. This served to limit contextual information and to isolate a suitable area of the image: the face area for the target images, and an area of ample variation in the black and white patches for the distractor images. The cropped images were – owing to being cropped by hand – of different dimensions. However, simply resizing the distractor images to the same dimensions as their corresponding target image could give rise to differences in scaling and amount of detail. In order to prevent these differences from being informative, the distractor images were cropped (instead of resized) once more, this time to the dimensions of their target image.

Panel B: one of our 3AFC test items, featuring the target image on the left, and its two accompanying distractors in the middle (the distractor image from panel A, row 3) and on the right (the distractor image from panel A, row 2).

Figure 2. Three additional examples of test items.

For each trial, we recorded the coordinates (in pixels) of where the participants clicked with the mouse, and thus we could determine whether the pixel that the participant had clicked was located both within the correct columns (2–5) and in the correct row (C, D, E, or F, depending on the item). If the participant's click was indeed located both in the correct columns and in the correct row for that particular item, the response was recorded as correct; if the participant clicked on the target image but not within the correct 'eye-region,' or clicked on one of the distractor images, the response was labelled as incorrect.

We set out to balance our first stimulus set (316 items) on four variables: 1. an equal number of stimuli for each of the three possible positions of the target image (the positions being the left, the middle, or the right panel); 2. for each of these three target image positions an equal number of stimuli for each of the four possible eye regions (rows C, D, E, or F); 3. for each of the aforementioned possibilities an equal number of stimuli for both sexes of the volunteer depicted (female or male); and 4. for each of the aforementioned possibilities an equal number of stimuli for the type of background used in the photograph (black or white). The above criteria were not perfectly fulfilled for our set of 316 items: an exact division of the number of photographed volunteers (158) by the number of options ($3 \times 4 \times 2 \times 2 = 48$) was not possible. However, since the main goal of the first testing phase was stimulus selection, we preferred to use a set that was not perfectly balanced as opposed to not using all possible images in this initial testing phase.

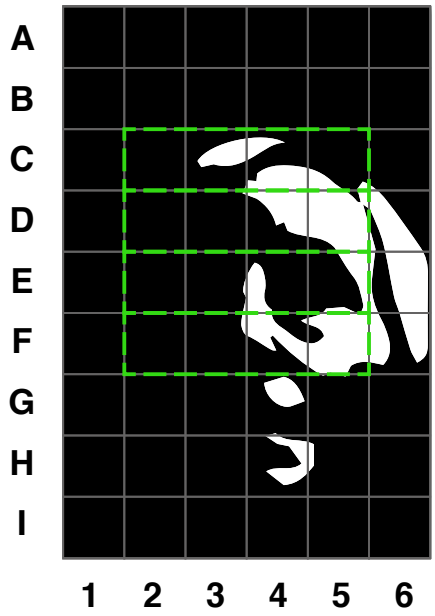

Figure 3. One of the target images with – overlaid as solid grey lines – the 6 by 9 squares grid that is used to quantify response. The dashed, green, thicker lines mark the four rectangular regions in which the eyes could be located: rows C, D, E, or F (in this case, the correct region is row E), always extending from column 2 to column 5 inclusive. Participants were not told that there were limitations on the possible location of the eyes.

2.3 Selection of final stimulus set

Using the results from the first testing phase (see §3.4) we were able to narrow down our selection of 3AFC items from 316 to 144 (the final stimulus set), thereby reducing the overall testing time while retaining the most informative stimuli. We made our selection by fitting a two-component model from Item Response Theory (Embretson & Reise, 2000; Nunnally & Bernstein, 1994) to the performance data, keeping the third component (the guessing parameter) fixed at zero, since all items are identical in arrangement (3AFC) and response (clicking on either of the eyes). The two components of the IRT model that we calculated for each item were the so-called *difficulty* and *discrimination* indices. The former ($\frac{\Phi^{-1}(p)}{r}$) is calculated by dividing the z-score (under the cumulative probability curve) of the proportion of participants who scored correctly ($\Phi^{-1}(p)$) by the correlation between the score – of all participants – on this particular item and the overall score (r). Item *discrimination* ($\frac{r}{\sqrt{1-r^2}}$) is calculated by dividing the correlation between the score on that particular item and the overall score (r; also referred to as the point-biserial correlation) by the square root of the variance that is *not* explained by this correlation ($\sqrt{1-r^2}$). The *discrimination* index is a measure of how well the item (and performance on that item) can discriminate between participants of differing ability. This index was the primary criterion in selecting our new stimulus set – for every item the value of the discrimination index should be as close to 1 as possible. We sought a large range on the difficulty index, hence we did not restrict it. The second criterion was the requirement to balance the stimulus set on the variables described before (§2.2).

Unfortunately, simply selecting the set of 144 items with discrimination index values closest to 1 did not satisfy our balancing criteria (described above in §2.2). We were thus forced to swap some items with discrimination indices closest to 1 for items whose value was less close to 1. In order to reduce the number of swaps, we tried to limit the number of variables that needed balancing during item selection: we decided to disregard for the moment the position of the target image among the three panels (whether the target image is in the left, middle, or right panel). Instead, *after* item selection, we re-shuffled the positioning of the target and distractor images within each 3AFC item – but not *across* items, since the difficulty of stimuli is likely to depend both on the target image as well as on its accompanying distractors – until again our set contained an equal number of items per target image position (left, middle, or right panel). As

for the other three variables: we left unchanged the target images themselves and thus kept the eye region of the images fixed; sex of the volunteer depicted in the target image was inherently fixed; and the type of background was disregarded completely because it proved of non-significant influence (see §3.4). With these conditions specified, we selected a set of 144 items for which the discrimination index value for all individual items lay as close as possible to 1, while the set as a whole was balanced on variables 1 to 3 described above.

Since a Mooney face once perceived seems to be easily found upon repeat presentation, a test–retest paradigm using identical image sets would not be informative. We thus split the new selection into two different parts of equal length ('A' and 'B', each comprising 72 items) to allow for test–retest. Test parts A and B were each balanced on the same variables as the overall stimulus set of 144 items.

3. First phase of testing: 316 items

3.1 Participants

374 participants (57% female) from varying ethnic groups (though predominantly white: 86%), whose ages ranged from 18 to 68 (M = 26 years), were recruited via word-of-mouth, social media, online notice boards, and electronic mailing lists. Ethical permission for the study was given by the Cambridge University Psychology Research Ethics Committee.

3.2 Materials

The stimulus set of 316 items was used in the first phase of testing (see §2.2).

3.3 Procedure

Of our sample, 330 participants (59% female) completed our test online, while the other 44 participants (45% female) completed the test in the lab. The procedure for both was identical: we asked participants to give basic demographic information and subjectively to rate their face recognition ability in response to the question "On a scale of 1 to 10 (with 1 being really bad, and 10 being really good), where would you place yourself in terms of recognising faces?" The subsequent instruction screen informed participants that their goal would be to identify the face and asked them to respond by clicking on either of the eyes of the face. Trials did not have a time limit, in order to make sure that we would have response data for every trial, although participants were instructed to respond as quickly as possible. The stimulus remained on the screen until participants responded. An inter-stimulus interval consisting of a blank screen, but without fixation cross, was presented for 500 ms between trials. A practice trial with feedback preceded the first of four blocks of 79 trials; no feedback was given on test trials. Between blocks, participants could take a break of indefinite length. For the online sample, all stimuli were downloaded to the participant's computer before the test trials began, to ensure that participants did not experience a lag between or during trials.

3.4 Results

Before analysing the data, we visually inspected the responses to each item. Since we recorded the exact x and y coordinates of participants' mouse clicks, we could overlay responses on to the stimuli (see Figure), and could thereby verify the position of the eye region that we had defined during stimulus development. A lack of detail is inherent to Mooney images – the eyes might be embedded in a larger shadowed area – and our definition of the eye region (derived from the original photograph, see §2.2) did not always map well on to the response of the majority of participants (see Figure B). We thus translated – for 68 out of 316 items – the eye region up or down by between one and fifteen pixels (all target and distractor images were 192 pixels wide and

288 pixels high), and recalculated our performance measure for all participants using these new coordinates of the correct eye region (see Figure C). We use this new performance measure throughout our analyses that follow. We also translated the actual target images for future use, to avoid the need for translating the responses. The final selection of 144 items included 38 of these translated items.

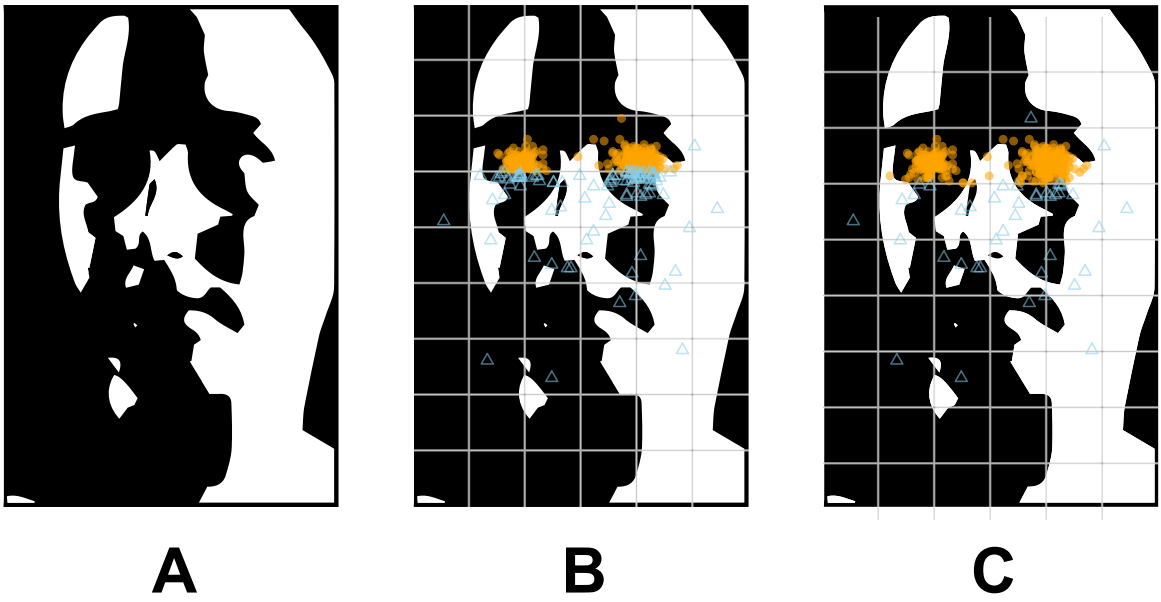

Figure 4. Visual inspection of target images. Panel A: one of our Mooney target images included in the first test phase (though not in the final set of 144 items). Panel B: the target image with the responses of 374 participants overlaid. Every symbol represents a participant: orange, solid circles are responses within the eye region that was initially labelled as correct, and these responses were thus initially deemed correct; blue, open triangles are responses outside of this eye region, and thus initially deemed as incorrect. Also overlaid is the grid of 6 by 9 squares that was used to localise the eye region during stimulus development; in this image, row C is the correct row. Panel C: the grid of 6 by 9 squares has been translated by –7 pixels (downward translation) since participants' responses indicated a slight discrepancy between labelled eye region and perceived eye region. Participants' responses have been recalculated to reflect the relocation of the correct eye region – the symbols reflect this recalculation, and the final performance measure is based on these recalculated responses.

Mean performance of the subset of 44 participants who completed our test in the lab, did not differ significantly from that of the sample of 330 participants who completed the test online (Mann–Whitney U = 7646.5, p = .57). We therefore combined the two samples, and use all 374 participants in all subsequent analyses.

In our first testing phase, for our stimulus set of 316 items, the overall performance was wide but negatively skewed: the mean score correct was 77.9% (SD = 16.6%), with a range from 3.5% to 97.5% (see Figure A). Since the lower end of the performance range is surprising (3.5% correct), we investigated an alternative performance measure: a score based on the correct clicking of the target image only, regardless of clicking within the correct eye region. The comparison of this measure with our eye-clicking performance measure showed that the majority of participants in the extremely low range of the latter performed reasonably well on the former. Additionally, feedback from a number of participants suggested that some participants had forgotten the instructions to click on either of the eyes to respond, and instead had clicked merely on the image that contained the face. However, we could not reliably separate the participants who had forgotten the instructions from those whose performance might actually be at the lower end; and since these data were used for trial selection only, and since participants who had forgotten the instructions were probably consistent in their forgetting, we did not exclude any participants.

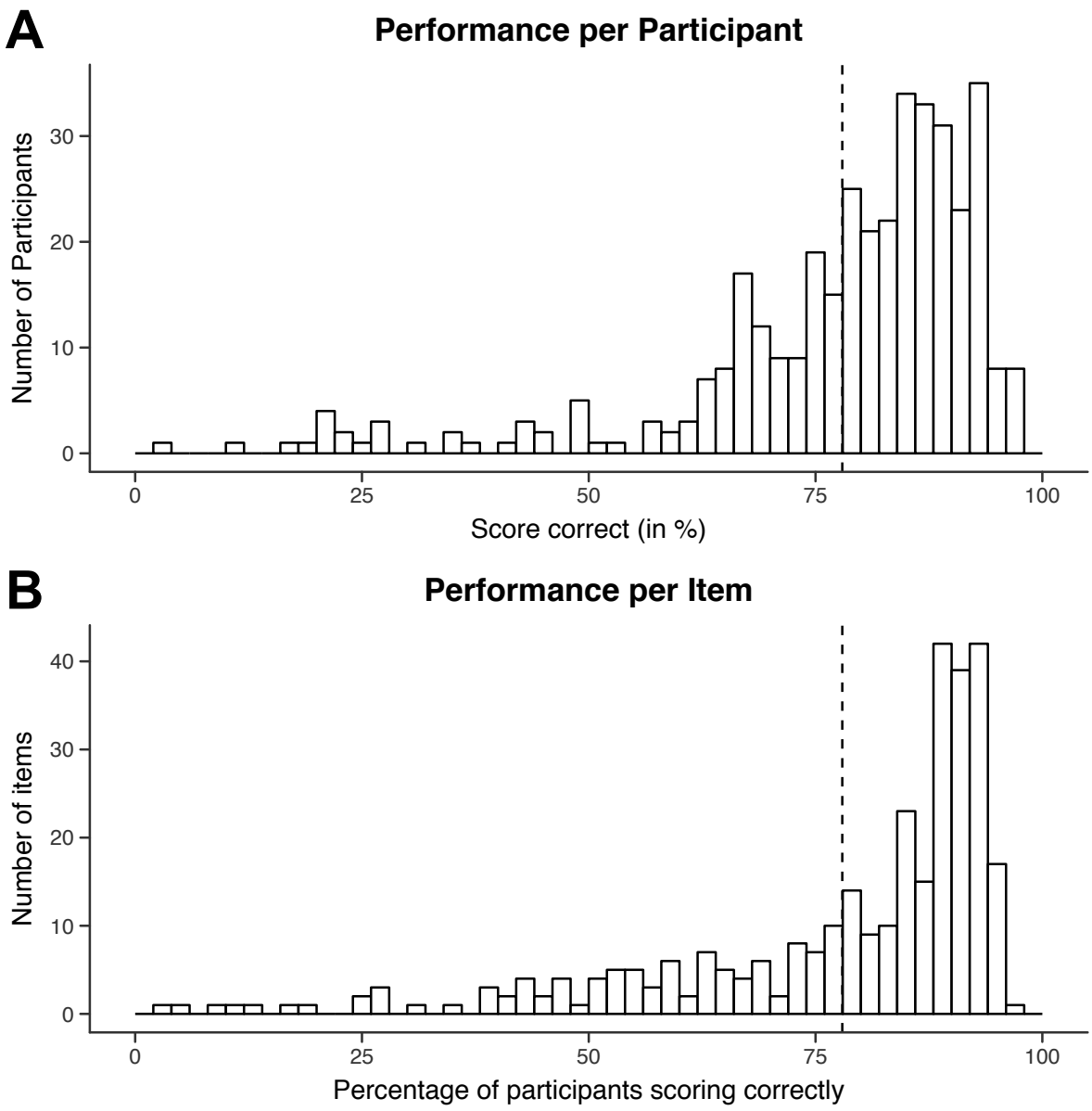

Figure 5. Panel A: the distribution of scores (in percentage) for participants. The dashed vertical line indicates the mean. Panel B: the distribution of scores (in percentage of participants scoring correctly) for items. The dashed vertical line indicates the mean.

We observed a significant sex difference favouring males (Mann–Whitney U = 12,279, $p = 2.87\times10^{-6}$; $\bar{x}_{\text{females}} = 75.1\%$, $\bar{x}_{\text{males}} = 81.7\%$ – a difference of .40 SD), confirming previous findings using the classical Mooney stimuli (Foreman, 1991; Verhallen et al., 2014). We did not observe a significant correlation between performance and age (Spearman's $\rho = .04$, $p = 0.47$), even when sex was regressed out from both variables (Spearman's $\rho = .01$, $p = 0.84$). Participants' subjectively rated ability for 'recognising faces' (M = 6.8, SD = 1.6, range 1–10) did not correlate significantly with their Mooney performance (Spearman's $\rho = .07$, $p = .17$). We did not observe a significant difference of performance between ethnic groups (Kruskal–Wallis $\chi^2 = 1.81$, $p = .77$), even when we grouped all non-white participants (since the target images depicted Caucasian volunteers only) and compared them to the group of white participants (Mann–Whitney U = 8,169.5, $p = .78$). However, group sizes in our sample are very disparate ($N_{\text{asian}} = 24$, $N_{\text{black}} = 4$, $N_{\text{mixed}} = 19$, $N_{\text{other}} = 5$, $N_{\text{white}} = 322$); and – if the Other Race Effect is an effect of training – then

ethnicity of the observer would not be the best measure of exposure to faces of a specific ethnic group, but instead country of birth and country of residence should be used.

The distribution of performance per item (i.e. the percentage of participants scoring correctly) shows a large cluster around 80–95% (see Figure 5B), although no item was solved by every participant. We observed no significant difference in performance for stimuli featuring black backgrounds as compared to performance for stimuli featuring white backgrounds (Wilcoxon signed-rank W = 75,339, $p = .068$). We did observe a significant difference in participants' performance for stimuli depicting a female volunteer as compared to performance for stimuli depicting a male volunteer (W = 60,476.5, $p = 2.05\times10^{-48}$): participants scored on average 4.7% (or .28 SD) higher for images depicting females (see Table 1). However, there does not appear to be a strong interaction between sex of participant and sex of the target face: the advantage for female faces over male faces is only slightly greater for female participants than for male participants (5.2% vs. 4.2%, respectively).

Sex of volunteer in image:	Female volunteer	Male volunteer
Mean performance in % (SD in %)	**80.3 (16.8)**	**75.6 (16.7)**
Female participants (N = 214)	77.7 (17.9)	72.5 (17.6)
Male participants (N = 160)	83.8 (14.5)	79.6 (14.6)

Table 1. Performance (in percentage) presented separately for the two sexes of the volunteer depicted in the target image ("Female volunteer" vs. "Male volunteer"), and broken down by sex of the participant.

We observed a significant difference in performance across target image location (Friedman $\chi^2 = 343.06$, $p = 3.20\times10^{-75}$), whereby performance was significantly higher for target images in the middle panel ($\bar{x} = 82.3\%$), intermediate for those in the left panel ($\bar{x} = 77.4\%$), and lowest for those in the right panel ($\bar{x} = 73.7\%$). Another significant difference was found for participants' performance across the four different eye regions (Friedman $\chi^2 = 735.23$, $p = 4.82\times10^{-159}$): performance was highest for row E ($\bar{x} = 84.8\%$), lower for row D ($\bar{x} = 82.3\%$), lower still for row C ($\bar{x} = 76.8\%$), and lowest for row F ($\bar{x} = 67.6\%$) – all group differences were significant. However, in the development phase of the experiment, there was no independent means of quantifying the difficulty of a target image or a 3AFC item, hence difficulty could not be controlled for. The observed differences in performance could thus be inherent to the stimulus, rather than a product of the sex of the volunteer or the location of the eyes.

Since there were two Mooney target images for each photographed volunteer, we investigated a potential priming effect. For each participant, and for every volunteer pair for which the participant had correctly responded to the first image, we computed mean performance on the second image of the pair and – as a comparison – mean performance on all other volunteer pairs. Across participants, their mean performance on a volunteer pair's second image ($\bar{x} = 80.9\%$) was significantly higher than their mean performance for all other images ($\bar{x} = 77.9\%$; Mann–Whitney U = 80,548, $p = .0006$). To verify that this result was not an effect of image similarity within volunteer pairs – i.e. if images of the same volunteer tend to be similar in difficulty, then selecting only those pairs for which the participant scored correctly on the first image would yield a skewed measure – we correlated performance between the two images of volunteer pairs. We did not observe a significant correlation (Spearman's $\rho = .09$, $p = 0.26$).

4. Second phase of testing: 144 items

4.1 Participants

505 participants (61% female) from varying ethnic groups (again predominantly white: 84%), whose ages ranged from 18 to 70 (M = 27 years), were recruited via word-of-mouth, social media, online notice boards, and electronic mailing lists. There was no overlap of participants between the samples of the two testing phases, after we excluded 43 participants from the second phase who had previously taken part in the first phase (their overall performance was significantly higher by 6.4%, or .46 SD, than performance of the 505 remaining participants who had not taken the test before: Mann–Whitney U = 13,681, $p = .0046$). Ethical permission for the study was given by the Cambridge University Psychology Research Ethics Committee.

4.2 Materials

The final stimulus set of 144 items was used in a test–retest paradigm (see §2.3 for the selection procedure).

4.3 Procedure

The second testing phase was entirely conducted online. The procedure was very similar to that of the first phase: participants were asked to supply basic demographic information and to rate subjectively their face recognition ability, after which they were shown the same instruction screen as in phase one. Again, stimuli remained on screen until participants responded and an inter-stimulus interval consisting of a blank screen (with no fixation cross) was presented for 500 ms between trials.

However, we now also asked participants to indicate the "country where [they] grew up (or where [they] spent most time until age 18)" using a drop-down menu that listed all countries in the world (ISO 3166 standard; retrieved 19 June 2014 – www.iso.org/obp/). Additionally, we asked participants for their handedness ("left", "right", or "both") in response to the question "Which hand do you write with?" We also asked participants to indicate whether they had "taken (a version of) this test before", and if yes, how long ago they had taken it ("Less than 1 week", "Less than 1 month", "Less than 3 months", "Less than 6 months", "Less than 1 year", "More than 1 year").

In contrast to the first test phase, the second phase of testing followed a test–retest paradigm: participants completed parts A and B in a randomly assigned order, with a minimum interval of three days. Both parts consisted of two blocks of 36 trials each; participants could take breaks of indefinite length between blocks. Results from phase one suggested that – as the test progressed – some participants forgot the instructions to click on the eyes of the face and instead merely clicked on the target image containing the face (see §3.4). Hence, in phase two, a different practice trial with feedback preceded *each* block of 36 trials, instead of only one practice trial preceding the entire test. Again, no feedback was given on test trials.

4.4 Results

The overall performance was again wide, and negatively skewed: the mean score correct was 77.6% (SD = 14.1%), with a range from 36.8% to 98.6% (see Figure 6). No participant hit ceiling, and no single item was solved by all participants. We continued to observe a significant sex difference favouring males (Mann–Whitney U = 22,287.5, $p = 5.64 \times 10^{-7}$; $\bar{x}_{females} = 75.3\%$, $\bar{x}_{males} = 81.2\%$ – a difference of .42 SD; see Figure 6). We also continued to observe no significant difference between ethnic groups (Kruskal–Wallis $\chi^2 = 4.44$, $p = .49$), though group sizes remained disparate ($N_{eastAsian} = 19$, $N_{southAsian} = 16$, $N_{black} = 2$, $N_{mixed} = 28$, $N_{other} = 14$, $N_{white} = 426$). Participants originated from 52 different countries; the largest group of participants (49%)

originated from the United Kingdom. As was the case for ethnicity, we did not observe a significant difference in performance, when countries were pooled into five distinct groups: African, Arabic, Asian, Caucasian, and South American (Kruskal–Wallis $\chi^2 = 3.02$, $p = .55$). This time we did observe a significant, though modest, correlation of participants' subjectively rated ability with their performance (Spearman's $\rho = .12$, $p = .006$).

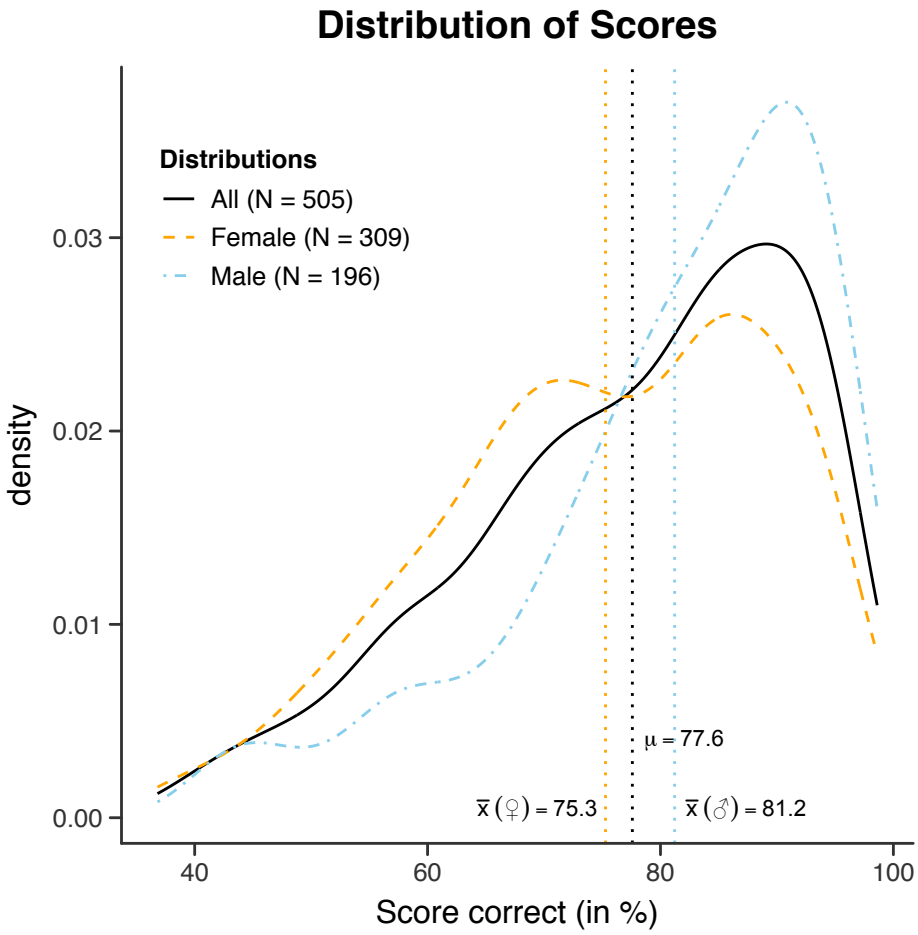

Figure 6. A density plot of scores (in percentage) for the entire population (black, solid line) and separately for each of the sexes: an orange, dashed line for females; and a blue, dot-dashed line for males. The three vertical, dotted lines indicate – from left to right – the mean for female participants (75.3%), the population mean (77.6%), and the mean for male participants (81.2%).

Our 505 participants completed all 144 trials (i.e. the sum of both sessions) in on average 9.85 minutes (SD = 4.25 minutes) with a range of 3.8 to 37 minutes; two outliers were removed from these statistics, because these two participants had taken 4 hours and 16 hours, respectively, to complete the test. There was a modest but significant, negative correlation between performance and average time taken per trial, with the influence of sex removed from both variables by means of linear regression (Spearman's $\rho = -.14$, $p = .002$). On the trials where participants responded correctly, they took on average 3,841 milliseconds to respond (i.e. to reach closure), although this measure includes the time it takes to move the mouse and click on the eyes. In contrast, participants took on average 6,258 milliseconds to respond incorrectly.

As in the first test phase, we continued to observe a significant difference in performance across target image location (Friedman $\chi^2 = 333.89$, $p = 3.16 \times 10^{-73}$), although no longer for all

combinations: the difference between the right ($\bar{x}$ = 71.8%) and left ($\bar{x}$ = 79.5%) panels, as well as that between the right and middle ($\bar{x}$ = 80.5%) panels, was again significant, but the difference between the left and middle panels was not. We also continued to observe a significant difference for participants' performance across the four different eye regions (Friedman $\chi^2 = 910$, $p = 6.00\times10^{-197}$), but now this was true only when we compared the outer two rows (C and F) to the middle two rows (D and E); there was no significant difference of performance between rows C ($\bar{x}$ = 70.8%) and F ($\bar{x}$ = 70.9%), nor between rows D ($\bar{x}$ = 83.8%) and E ($\bar{x}$ = 84.9%). This difference in performance between the outer two and middle two rows might be due to participants' tendency, when guessing, to click in the centre of the image. The effect of target image location, as well as the effect of eye region, held when analysed for females and males independently. We again observed a significant difference in participants' performance for target images depicting a female volunteer as compared to performance for target images depicting a male volunteer (Wilcoxon signed-rank W = 79,196, $p = 5.01\times10^{-22}$; see Table 2). Here the advantage for female faces relative to male faces does seem to be somewhat larger for female participants as compared to male participants (4.0% vs. 1.6%, respectively).

Sex of volunteer in image:	Female volunteer	Male volunteer
Mean performance in % (SD in %)	**77.9 (13.7)**	**74.8 (15.8)**
Female participants (N = 309)	76.2 (13.6)	72.2 (16.0)
Male participants (N = 196)	80.5 (13.6)	78.9 (14.8)

Table 2. Performance (in percentage) presented separately for the two sexes of the volunteer depicted in the target image ("Female volunteer" vs. "Male volunteer"), and broken down by sex of the participant.

In our test–retest paradigm, participants were randomly assigned an order in which they completed parts A and B: either first A and then B (referred to as "AB"), or first B and then A ("BA"). A minimum of three days separated the two sessions; on average, participants took 6.6 days between sessions. Since there were a number of participants who started our test but did not complete it (for this reason they are not part of the sample of 505 participants we report here), there was a slight discrepancy in sample size for the two different orders: 260 participants completed part A then part B, whereas 245 participants completed part B then part A. We investigated whether this discrepancy was due to a differential difficulty of the two parts, and found an interaction between difficulty and test order: we observed a small but significant difference in performance between part A and part B for participants' first session (Mann–Whitney U = 36,150, $p = .009$; $\bar{x}_{\text{part A – session 1}} = 75.3\%$; $\bar{x}_{\text{part B – session 1}} = 71.2\%$), but not for participants' second session (U = 31,039, $p = .62$; $\bar{x}_{\text{part A – session 2}} = 81.7\%$; $\bar{x}_{\text{part B – session 2}} = 82.0\%$). However, when we regressed out the influence of test order, we did not observe a significant difference of overall performance between parts A and B (Wilcoxon signed-rank W = 127,508, $p = 1$; $\bar{x}_{\text{part A}} = 78.4\%$, $SD_{\text{part A}} = 14.8\%$; $\bar{x}_{\text{part B}} = 76.8\%$, $SD_{\text{part A}} = 16.2\%$). We did observe a marginally significant difference between parts in the average time taken to complete each part, again with the influence of test order removed from both measures (W = 71,183, $p = .01$; $\bar{x}_{\text{part A}} =$ 4.74 minutes, $SD_{\text{part A}} = 2.31$; $\bar{x}_{\text{part B}} = 5.10$ minutes, $SD_{\text{part B}} = 2.95$).

In order to see whether there was a learning effect over sessions, we investigated the difference in performance between the two test sessions, combining results from the two test orders: we observed that – regardless of participants' test order – performance was significantly higher for the second session ($\bar{x}$ = 81.9%) as compared to the first session ($\bar{x}$ = 73.3%; W = 85,932, $p =$

2.83×10^{-19}; see Figure 7). We also observed a significant difference in time taken between sessions: participants took on average less time during their second session ($\bar{x}$ = 4.07 minutes) as compared to their first ($\bar{x}$ = 5.70 minutes; W = 118,135, $p = 3.82\times10^{-65}$). However, parts A and B yielded virtually identical results when running the analyses from previous paragraphs – sex differences, age differences, influence of target image location and eye region – for the two parts independently.

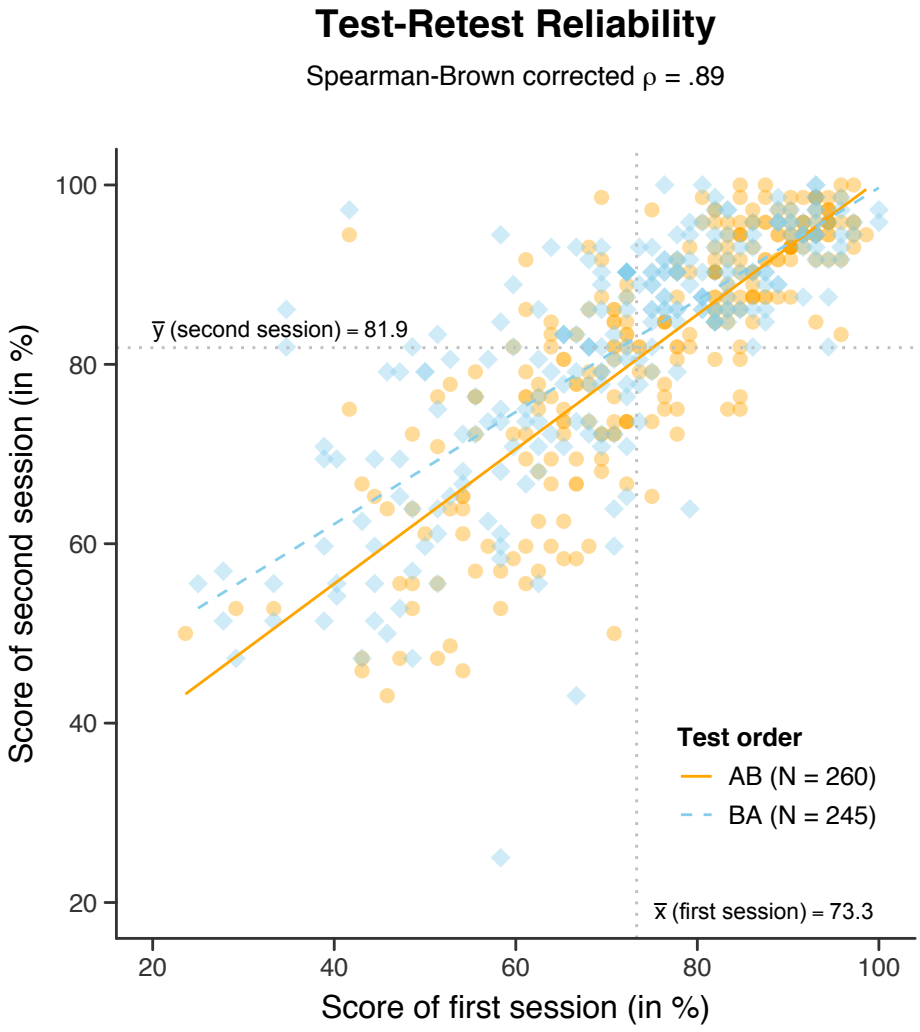

Figure 7. Scatter plot showing the test–retest reliability. Scores (in percentages) are plotted separately for participants who completed part A then B ("AB"; orange dots), and for participants who completed part B then A ("BA"; blue diamonds). The saturation reflects the number of participants with that particular score. The diagonal lines show the linear model fitted to the data of the two groups: an orange, solid line for "AB"; and a blue, dashed line for "BA". The vertical and horizontal dotted lines indicate the mean scores of the two sessions (regardless of test order): 73.3% for participants' first session (vertical line), and 81.9% for their second session (horizontal line). The Spearman–Brown corrected test–retest correlation was .89 for the entire test of 144 trials.

For any new test, there are three measures of reliability worth calculating: internal reliability (the extent to which performance on individual items correlates with overall performance), parallel-forms reliability (to see whether our parts A and B are indeed equivalent), and test–retest reliability (to see whether our test yields similar results at different points in time). The final set of 144 items has an internal reliability estimate of Guttman's λ_6 = .97. This estimate considers the amount of variance in each item that can be accounted for by the linear regression of all other

items; in other words, how similar are the abilities that each item taps. When we calculate the internal reliability estimates for the parts A and B independently they are Guttman's $\lambda_6 = .94$, and $\lambda_6 = .95$, respectively. These estimates are almost identical to the estimate of parts A and B combined (see above) – it is thus likely that data from a single part will yield very similar results to those from the test as a whole. Indeed, if a very brief testing time is critical, it might prove sufficient to administer only one part – in that case, although our results do not point towards a superiority of one part over the other, we advise the use of part A in the interest of consistency across studies.

To obtain the parallel-forms reliability, we correlated participants' score on part A with their score on part B, after having removed – by means of linear regression – any variance due to participant sex and due to test order. The resulting correlation is Spearman's $\rho = .80$ ($p = 2.27\times10^{-111}$). Our two parallel forms (part A and part B) are thus very similar. Finally, to obtain the test–retest reliability, we correlated participants' score on their first session with their score on their second session, but only after we had used linear regression to remove the variance due to participant sex and due to test order. The resulting test–retest correlation is Spearman's $\rho = .80$ ($p = 6.62\times10^{-116}$; see Figure 7); the result of a Spearman–Brown correction is a test–retest correlation of $\frac{2 \times .80}{1 + (2-1) \times .80} = .89$ for the final test as a whole (144 items). Interestingly, we observed a slightly elevated test–retest correlation for those who completed part A then B (uncorrected $\rho = .83$) as compared to those who completed part B then A (uncorrected $\rho = .79$). We thus suggest – if indeed both parts A and B are administered – that the test order be "AB."

Figure 8 gives the cumulative distribution curve of our 505 participants for the final stimulus set, to allow these data to be used as normative sample for comparison of future studies.

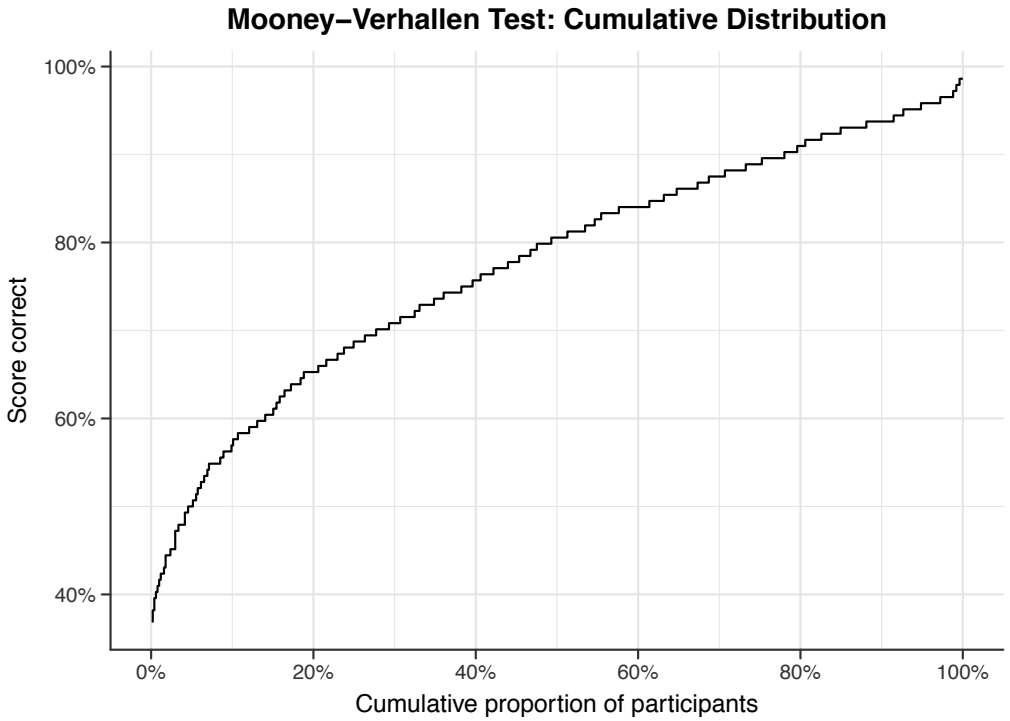

Figure 8. The cumulative distribution of performance for the final set of 144 items, from data of 505 participants gathered in the second phase of testing.

5. Discussion

We have developed a new, online version of the Mooney test that is suitable for test–retest paradigms. The Mooney–Verhallen Test has an internal reliability estimate of Guttman's $\lambda_6 = .97$, and a Spearman–Brown corrected test–retest correlation of Spearman's $\rho = .89$. Participants take on average 9.7 minutes to complete all 144 trials, and performance on our test shows marked individual differences in the perception of Mooney faces. Data from the initial 316-item version did not show a significant difference in mean performance between an online and a lab-based sample, a result consistent with other comparisons of online and lab-based administration of tests of face processing (Germine et al., 2012). A possible further avenue to explore would be the development of an adaptive version of the Mooney–Verhallen Test, in order to fully harness the benefits of Item Response Theory. An adaptive version has the potential to further reduce the required length of testing, and to increase discriminability across the entire gamut of Mooney face perception ability.

For both the initial stimulus set (316 items) and the final stimulus set (144 items), we continued to observe a significant sex difference favouring males (6.6% and 5.9%, respectively; or .40 and .42 standard deviation), confirming earlier results using the original Mooney stimuli (Foreman, 1991; Verhallen et al., 2014). Although previous studies investigating sex differences in face processing have reported mixed results, they largely point to a female superiority (Megreya et al., 2011; Sommer et al., 2013). The observed sex differences for the Mooney test thus could reflect other, non-face-perception processes that might be at play. Comparison of performance on the Mooney test with performance on other tests of face processing and on tests of visual processing, as well as the manipulation of variables that could affect performance (viewing distance, degradation of images), could shed further light on the exact processes underlying the perception of Mooney images.

In our second test phase, we did not observe a significant correlation of performance with either age or handedness. The latter finding replicates that of Vigen and colleagues (1982), who also did not observe a significant difference of performance between handedness groups (total N = 100 college students). However, they did observe a significant correlation of performance with age, in that performance deteriorates with age, though for females only (Vigen et al., 1982). Although we currently did not observe a correlation of performance with age, for our new Mooney test any analyses of age should be interpreted with caution: the performance measure depends partially on the participant's dexterity with using a mouse or track pad to click within the correct eye region, a skill on which participants of advanced ages might be impaired.

We also observed, in our second test phase, a significant difference in performance between sessions one and two: an 8.6% increase of *overall* performance, regardless of the order in which participants completed the two parts. The average interval between sessions was 6.6 days, and the improvement could be an example of 'reminiscence,' which is observed both for motor skills (Buxton, 1943; Hovland, 1951) and for perceptual skills (Karni & Sagi, 1993) when an interval follows training. That performance is aided by reminiscence is suggested by our preliminary finding that mean performance in our second phase of testing was significantly higher for participants who had previously taken part in our first phase of testing (where the interval was more than six months). However, our repeat participants might have been a self-selecting group – those who performed well might have enjoyed the test more, motivating them to take it again during the second round of testing – and our sample of repeat participants was relatively small (N = 43).

Previous research has shown that priming of Mooney faces using non-degraded photographs has an influence on the participant's subsequent judgement of the familiarity of the person depicted in the Mooney image (Jemel et al., 2003). Since our original stimulus set of 316 items contained two images for each volunteer, we investigated a potential priming effect for the ability to *perceive* a Mooney face. Average performance for the second image of a volunteer – for those volunteer images of which the participant had previously correctly identified the first image – was significantly, though modestly, higher than average overall performance. In combination with the absence of a significant correlation of performance *between* volunteer images (i.e. a participant's performance on one volunteer image does not predict his or her performance on the other volunteer image), this seems to suggest that perception of the volunteer's first image somehow influenced (aided?) the subsequent perception of the volunteer's second image. Although the Mooney images are two-dimensional, participants may construct an internal 3D model (Moore & Cavanagh, 1998) of the perceived face, which could facilitate perception of a subsequently presented Mooney image of that same face.

In both rounds of testing we observed a significant difference in performance across target image location, in that performance for target images presented in the right panel was always significantly lower as compared to the middle and left panel. During the creation of the final stimulus set of 144 items, we reshuffled the 3AFC items in order to balance our set again, thereby changing the position of the target image; yet the difference is present in both test phases. Although eye movements were unconstrained, the observed difference in performance could be due to preferential processing of one hemifield as opposed to the other. Indeed, a left hemifield superiority for processing of faces has previously been reported (Bradshaw & Nettleton, 1983), including for Mooney faces (Parkin & Williamson, 1987). Furthermore, neuropsychological, electrophysiological, and fMRI studies suggest a right-hemisphere specialisation for the processing of faces (Bentin et al., 1996; Newcombe et al., 1989; McCarthy et al., 1997), an effect that is also found for Mooney faces specifically (George et al., 2005; Newcombe & Russell, 1969; Rossion et al., 2011). Future studies could conduct a more rigorous investigation of left field superiority for Mooney images, either by means of restricting the eye movements (having the stimulus disappear whenever the eyes stray too far from a central fixation point), or by briefly flashing the target image in only one hemifield (and subsequently comparing performance across the two hemifields).

The Mooney test remains of interest and continues to be used in visual perception research (especially on face processing), in clinical research, and in studies using brain imaging (Carbon et al., 2013; Grützner et al., 2013; Rivolta et al., 2014; Rossion et al., 2011; Towler et al., 2014). In combination with other measures, and across different populations – both clinical and non-clinical – the quick, reliable and standardised new Mooney–Verhallen Test might give further insights into the still mysterious nature of closure and its relationship to different genotypes, phenotypes and behaviours.

6. References

Andrews, T. J., & Schluppeck, D. (2004). Neural responses to Mooney images reveal a modular representation of faces in human visual cortex. *NeuroImage*, *21*(1), 91–98. http://doi.org/10.1016/j.neuroimage.2003.08.023

Bentin, S., Allison, T., Puce, A., Perez, E., & McCarthy, G. (1996). Electrophysiological Studies of Face Perception in Humans. *Journal of Cognitive Neuroscience*, *8*(6), 551–565.

Bradshaw, J. L., & Nettleton, N. C. (1983). Human cerebral asymmetry. Englewood Cliffs, N.J.: Prentice Hall.

Bruce, V., & Young, A. W. (2012). Face perception. London: Psychology Press.

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods*, *42*(1), 286–291. http://doi.org/10.3758/BRM.42.1.286

Busigny, T., Joubert, S., Felician, O., Ceccaldi, M., & Rossion, B. (2010). Holistic perception of the individual face is specific and necessary: evidence from an extensive case study of acquired prosopagnosia. *Neuropsychologia*, *48*(14), 4057–4092. http://doi.org/10.1016/j.neuropsychologia.2010.09.017

Buxton, C. E. (1943). The Status of Research in Reminiscence. *Psychological Bulletin*, *40*(5), 313–340. http://doi.org/10.1037/h0053614

Carbon, C.-C., Grüter, M., & Grüter, T. (2013). Age-dependent face detection and face categorization performance. *PloS One*, *8*(10), e79164. http://doi.org/10.1371/journal.pone.0079164

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, *44*(4), 576–585. http://doi.org/10.1016/j.neuropsychologia.2005.07.001

Embretson, S. E., & Reise, S. P. (2000). Item Response Theory for Psychologists. New Jersey: Lawrence Erlbaum Associates, Inc.

Foreman, N. (1991). Correlates of Performance on the Gollin and Mooney Tests of Visual Closure. *The Journal of General Psychology*, *118*(1), 13–20.

George, N., Jemel, B., Fiori, N., Chaby, L., & Renault, B. (2005). Electrophysiological correlates of facial decision: insights from upright and upside-down Mooney-face perception. *Cognitive Brain Research*, *24*(3), 663–673. http://doi.org/10.1016/j.cogbrainres.2005.03.017

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, *19*(5), 847–857. http://doi.org/10.3758/s13423-012-0296-9

Grützner, C., Wibral, M., Sun, L., Rivolta, D., Singer, W., Maurer, K., & Uhlhaas, P. J. (2013). Deficits in high- (>60 Hz) gamma-band oscillations during visual processing in schizophrenia. *Frontiers in Human Neuroscience*, *7*, 88. http://doi.org/10.3389/fnhum.2013.00088

Hovland, C. I. (1951). Human Learning and Retention. In S. S. Stevens, *Handbook of Experimental Psychology* (pp. 613–689). New York.

Jeffreys, D. A. (1989). A face-responsive potential recorded from the human scalp. *Experimental Brain Research*, *78*(1), 193–202.

Jemel, B., Pisani, M., Calabria, M., Crommelinck, M., & Bruyer, R. (2003). Is the N170 for faces cognitively penetrable? Evidence from repetition priming of Mooney faces of familiar and unfamiliar persons. *Cognitive Brain Research*, *17*(2), 431–446. http://doi.org/10.1016/S0926-6410(03)00145-9

Kanwisher, N., Tong, F., & Nakayama, K. (1998). The effect of face inversion on the human fusiform face area. *Cognition*, *68*(1), 1–11. http://doi.org/10.1016/S0010-0277(98)00035-3

Karni, A., & Sagi, D. (1993). The time course of learning a visual skill. *Nature*, *365*(6443), 250–252. http://doi.org/10.1038/365250a0

Lansdell, H. (1968). Effect of extent of temporal lobe ablations on two lateralized deficits. *Physiology & Behavior*, *3*(2), 271–273.

Latinus, M., & Taylor, M. J. (2005). Holistic processing of faces: Learning effects with Mooney faces. *Journal of Cognitive Neuroscience*, *17*(8), 1316–1327. http://doi.org/10.1162/0898929055002490

McCarthy, G., Puce, A., Gore, J. C., & Allison, T. (1997). Face-Specific Processing in the Human Fusiform Gyrus. *Journal of Cognitive Neuroscience*, *9*(5), 605–610. http://doi.org/10.1162/jocn.1997.9.5.605

McKeeff, T. J., & Tong, F. (2007). The timing of perceptual decisions for ambiguous face stimuli in the human ventral visual cortex. *Cerebral Cortex*, *17*(3), 669–678. http://doi.org/10.1093/cercor/bhk015

Megreya, A. M., Bindemann, M., & Havard, C. (2011). Sex differences in unfamiliar face identification: evidence from matching tasks. *Acta Psychologica*, *137*(1), 83–89. http://doi.org/10.1016/j.actpsy.2011.03.003

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3–35. http://doi.org/10.1037/1076-8971.7.1.3

Milner, B., Corkin, S., & Teuber, H.-L. (1968). Further analysis of the hippocampal amnesic syndrome: 14-year follow-up study of HM. *Neuropsychologia*, *6*(3), 215–234.

Mooney, C. M. (1957). Age in the development of closure ability in children. *Canadian Journal of Psychology*, *11*(4), 219–226.

Moore, C., & Cavanagh, P. (1998). Recovery of 3D volume from 2-tone images of novel objects. *Cognition*, *67*(1-2), 45–71.

Newcombe, F., & Russell, W. R. (1969). Dissociated visual perceptual and spatial deficits in focal lesions of the right hemisphere. *Journal of Neurology, Neurosurgery & Psychiatry*, *32*(2), 73–81. http://doi.org/10.1136/jnnp.32.2.73

Newcombe, F., De Haan, E. H., Ross, J., & Young, A. W. (1989). Face Processing, Laterality and Contrast Sensitivity. *Neuropsychologia*, *27*(4), 523–538.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). New York: McGraw-Hill, Inc.

Parkin, A. J., & Williamson, P. (1987). Cerebral Lateralisation at Different Stages of Facial Processing. *Cortex*, *23*(1), 99–110. http://doi.org/doi:10.1016/S0010-9452(87)80022-9

Rehnman, J., & Herlitz, A. (2007). Women remember more faces than men do. *Acta Psychologica*, *124*(3), 344–355. http://doi.org/10.1016/j.actpsy.2006.04.004

Rivolta, D., Castellanos, N. P., Stawowsky, C., Helbling, S., Wibral, M., Grutzner, C., et al. (2014). Source-Reconstruction of Event-Related Fields Reveals Hyperfunction and Hypofunction of Cortical Circuits in Antipsychotic-Naive, First-Episode Schizophrenia Patients during Mooney Face Processing. *Journal of Neuroscience*, *34*(17), 5909–5917. http://doi.org/10.1523/JNEUROSCI.3752-13.2014

Rossion, B., Dricot, L., Goebel, R., & Busigny, T. (2011). Holistic Face Categorization in Higher Order Visual Areas of the Normal and Prosopagnosic Brain: Toward a Non-Hierarchical View of Face Perception. *Frontiers in Human Neuroscience*, *4*. http://doi.org/10.3389/fnhum.2010.00225

Sommer, W., Hildebrandt, A., Kunina-Habenicht, O., Schacht, A., & Wilhelm, O. (2013). Sex differences in face cognition. *Acta Psychologica*, *142*(1), 62–73. http://doi.org/10.1016/j.actpsy.2012.11.001

Towler, J., Gosling, A., Duchaine, B., & Eimer, M. (2014). Normal perception of Mooney faces in developmental prosopagnosia: Evidence from the N170 component and rapid neural adaptation. *Journal of Neuropsychology*, 1–18. http://doi.org/10.1111/jnp.12054

Verhallen, R. J., Bosten, J. M., Goodbourn, P. T., Bargary, G., Lawrance-Owen, A. J., & Mollon, J. D. (2014). An online version of the Mooney Face Test: Phenotypic and genetic associations. *Neuropsychologia*, *63*, 19–25. http://doi.org/10.1016/j.neuropsychologia.2014.08.011

Verhallen, R. J., Bosten, J. M., Goodbourn, P. T., Bargary, G., Lawrance-Owen, A. J., & Mollon, J. D. (2015). Limited overlap between different measures of face processing. *In preparation.*

Vigen, M. P., Goebel, R. A., & Embree, L. J. (1982). Adults' performance on a measure of visual closure. *Perceptual and Motor Skills*, *55*(3), 943–952. http://doi.org/10.2466/pms.1982.55.3.943

Wasserstein, J., Barr, W. B., Zappulla, R., & Rock, D. (2004). Facial closure: interrelationship with facial discrimination, other closure tests, and subjective contour illusions. *Neuropsychologia*, *42*(2), 158–163. http://doi.org/10.1016/j.neuropsychologia.2003.07.003