Assessment
1–11
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/10731911221087746
journals.sagepub.com/home/asm
SAGE

Evaluating the Reliability and Validity of the Famous Faces Doppelgangers Test, a Novel Measure of Familiar Face Recognition

Emilia Pozo[1], Laura T. Germine[1,2], Luke Scheuer[1], and Roger W. Strong[1,2]

Abstract

Face recognition assessments that use images of celebrities require not only face recognition ability but also pop-culture knowledge and successful recall of identifying information. Here, we introduce a task designed to measure face recognition more specifically: the Famous Faces Doppelgangers Test (FFDT). Participants ($N = 57{,}407$) identified 40 celebrities paired with lookalike doppelgangers, allowing face recognition ability to be assessed without requiring information recall. In addition, participants reported whether they were familiar with each celebrity, allowing poor face recognition ability to be differentiated from low pop-culture knowledge. FFDT performance was reliable ($r_{xx} = .80$), similar across participants of different racial and ethnic groups, and more highly correlated with memory for faces ($r = .50$) and self-reported face recognition ability ($r = .48$) than processing speed ability ($r = .10$). Thus, the FFDT is a reliable, valid, and specific measure of the ability to identify familiar faces, making it a promising new tool for assessing face recognition ability.

Keywords

face recognition, visual recognition, test development, face perception, test validity, familiar face recognition

The ability to recognize faces varies greatly between individuals (Wilmer et al., 2012), ranging from those with extraordinarily good recognition ability (Russell et al., 2009) to those with impairment (Corrow et al., 2016). For instance, prosopagnosia is the impairment of recognizing and identifying faces, despite otherwise normal visual processing ability and intellect (Corrow et al., 2016). Often associated with disrupted connectivity between a distributed network of face processing regions in the right cerebral hemisphere (Fox et al., 2008; Thomas et al., 2009), prosopagnosia affects approximately 2% of the population (Kennerknecht et al., 2006, 2008) with a variety of face processing impairments (Susilo & Duchaine, 2013). These impairments can cause many challenges in everyday life such as confusing characters on TV, walking past family as if they were strangers, or not recognizing oneself in the mirror (Corrow et al., 2016). The frustration and helplessness of those experiencing recognition deficits (Yardley et al., 2008) are driving forces behind creating reliable and valid tools to assess face recognition ability, particularly given the possibility of using interventions to mitigate impairment (DeGutis et al., 2014).

Novel Versus Familiar Face Recognition

Face recognition tests can generally be sorted into one of two categories: those that assess recognition of novel faces versus those that assess recognition of familiar faces. Novel face recognition tests ask participants to remember and later recognize faces they are viewing for the first time (Benton et al., 1983; Warrington, 1984). These tests simulate the experience of meeting someone for the first time, and using short-term memory to recognize that person upon encountering them again soon after. Although these tests have the advantage of ensuring no previous exposure to the face stimuli across participants, they also have limitations. For example, participants can use nonfacial information (i.e.,

[1]McLean Hospital, Belmont, MA, USA
[2]Harvard Medical School, Boston, MA, USA

Corresponding Author:
Roger W. Strong, Institute for Technology in Psychiatry, McLean Hospital, 1010 Pleasant St, Belmont, MA 02478-9106, USA.
Email: rstrong@mclean.harvard.edu

hair, clothing, posture) to perform well on novel face recognition tests (Duchaine & Weidenfeld, 2003), likely contributing to tests that include nonfacial information being ineffective at classifying prosopagnosics as impaired (Duchaine & Nakayama, 2004, 2006). Notably, the more recent Cambridge Face Memory Test (CFMT; Duchaine & Nakayama, 2006) addressed this feature-matching limitation by using a memory paradigm with multiple faces and face stimuli strictly limited to facial information and has been successful classifying both those with extraordinary recognition abilities (Russell et al., 2009) and those with impairments (Duchaine & Nakayama, 2006). Still, given the length and number of stimuli used in the test, working memory capacity is a possible confounding variable in performance on the CFMT; how well participants are able to remember the many face images may affect their score, regardless of their face recognition abilities.

In contrast to novel face recognition tests, familiar face recognition tests require participants to recognize faces they have previously encountered in their everyday lives. These tests simulate the common experience of using long-term memory to recognize a previously seen individual. To test familiar face recognition using the same stimuli across participants of different social circles, familiar face recognition tests often use images of celebrities (Arizpe et al., 2019; Duchaine et al., 2007; Wilmer et al., 2012). For example, The Famous Faces Memory Test (Duchaine et al., 2007) presents the images of celebrities with limited nonfacial information and asks participants to identify the celebrity by typing the celebrity's name or providing other identifying information. The Famous Faces Memory Test has successfully measured face recognition abilities and accurately indicated impairment for those with prosopagnosia in multiple studies (Duchaine, 2000; Duchaine et al., 2007). Still, performance on this and similar familiar face recognition tests depends on more than just face recognition—performance is confounded by pop-culture knowledge and the ability to recall identifying information about recognized celebrities. In addition, many previous tests of familiar face recognition have lacked racial diversity in their stimuli, which is problematic given the effects that the race and ethnicity of faces used as stimuli have on recognition performance (Bothwell et al., 1989; Bowles et al., 2009; McKone et al., 2012).

Creating a Famous Faces Doppelgangers Test (FFDT)

Here we introduce a familiar face recognition task designed to assess face recognition independently of confounding factors present in previous studies. The FFDT assesses facial recognition ability without requiring participants to recall identifying information about celebrities; this is done by asking participants to simply identify a celebrity paired with a lookalike doppelganger. Recognition ability is differentiated from the confounder of pop-culture knowledge by giving participants the option to report that they did not know the target celebrity before making their selection. The use of doppelgangers assesses participants' ability to discriminate between two familiar faces, simulating the experience of seeing someone that looks familiar and identifying whether or not they are the correct person. In addition, the stimuli used in the test include a racially diverse selection of celebrities, as the race and ethnicity of stimuli have been found to have an effect on face recognition performance (Bothwell et al., 1989; Bowles et al., 2009; McKone et al., 2012).

Method

All study procedures were approved by the Harvard University Committee on the Use of Human Subjects. Below we report how we determined our sample size, all data exclusions, and all measures in the study.

Participants

Participants were recruited via the online neuropsychology research platform TestMyBrain.org (Germine et al., 2012). To our knowledge, all participants were directed to TestMyBrain.org from Google searches, web pages mentioning the website, or previous participants sharing a link to the website (e.g., via social media posts). Between August 2019 and April 2021, 57,407 participants between the ages of 12 and 99 years (M = 30.3, SD = 14.6; 58.9% female, 37.9% male, 3.2% nonbinary or genderqueer; see Table 2 for a summary of race/ethnicity) completed the FFDT; participants reporting an age outside of 12 to 99 were excluded from analyses. Of those 57,407 participants, 16,344 additionally completed the CFMT (Duchaine & Nakayama, 2006; Germine et al., 2011, 2012), 16,250 additionally completed a seven-item shortened version of the Cambridge Face Memory Questionnaire (CFMQ; Arizpe et al., 2019), and 7,662 additionally completed the TestMyBrain Digit Symbol Matching Test (Hartshorne & Germine, 2015). All participants who completed the seven-item version of the CFMQ also completed the TestMyBrain CFMT. Participants who completed multiple tests always completed the FFDT before any additional tests.

Participants were informed of their percentile test performance upon completion of their entire testing battery. We collected as many participants as possible between the FFDT's creation in August 2019 and April 2021, as the test was always available online for participants to complete during this time. Posting the test online in a remote, unsupervised setting allowed us to collect a much larger and more diverse sample than would have been possible running the test in person. Participants were able to report any combination of the race/ethnicity options listed in Figure 3 and Table 2; rather than categorizing each participant into a single category, we included participants within each subgroup they reported.

Measures

FFDT. The FFDT is composed of 40 image pairs of celebrities and their lookalike doppelgangers. Images were selected to present high quality, frontal views of faces. Doppelganger images were selected to match target images with similar facial expressions. In addition, we attempted to match the image characteristics and quality of each target-doppelganger image pair, to prevent selection of the correct answer by picking the image with higher resolution or better lighting; this resulted in most identified doppelganger images also being celebrities (31/40 doppelgangers). Doppelganger pairs were selected with the goal of including celebrities familiar to younger and older adults. Because individuals with prosopagnosia have described using features, like hair or voice, to identify people as a coping strategy (Corrow et al., 2016), images were cropped to only display the face without other identifying features. Additional editing included rotating the images to have similar head tilt, as well as attempting to match the brightness and color of the images in each pair. Out of 51 celebrity-doppelganger pairs that were initially created, 40 pairs were selected that spanned a wide difficulty range from near chance to near perfect accuracy during pilot testing.

Before beginning the test, participants were presented with instructions reading, “You will be shown a series of famous faces, alongside a doppelganger, which looks similar to the famous person. Click on the face whose name appears on top of the images. If you have no idea who the famous person is, check the box marked ‘I don’t know this person.’ (You still have to choose a face).” On each trial, images of the target celebrity and a lure doppelganger were simultaneously presented alongside one another (Figure 1). Above the images was text asking the participant to select the target celebrity (e.g., “Which face is Michael B. Jordan?”). Before making their selection, participants had the option to disclose that they did not know the target celebrity by clicking a checkbox below the images; participants were required to make a selection even when they reported not knowing the target celebrity. Participants were given accurate feedback following each trial. If a participant took longer than 50 seconds to respond, a message appeared stating “You are taking too long to respond. Click here to retry.” Across all participants, this timeout message was presented for only 0.1% of trials. Median response time was 3.0 seconds.

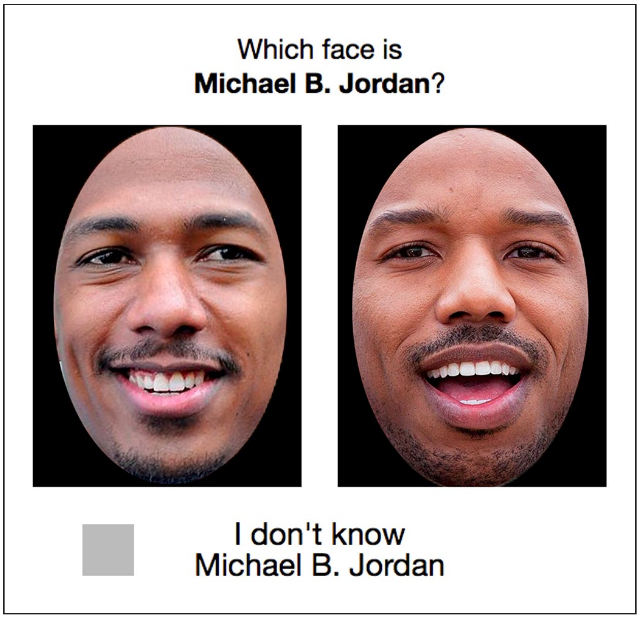

Figure 1. Example Stimulus.
The target celebrity (Michael B. Jordan) is presented on the right, while a lure doppelganger (here also a celebrity, Nick Cannon) is presented on the left. Participants were instructed to check the box below the images if they were not familiar with the target celebrity, but still were required to make a response. Both images are licensed under the Creative Commons Attribution 2.0 Generic license and have been cropped. Left image by Nick Step, right image by Joan Hernandez Mir.

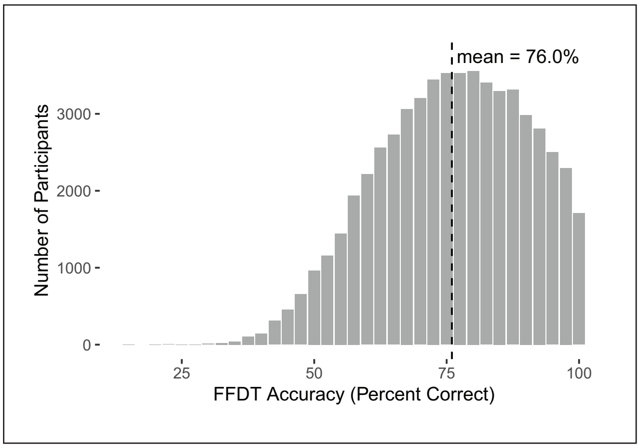

Figure 2. Histogram of Famous Faces Doppelganger Test Performance.

The target celebrities included in the FFDT were: Michael B. Jordan, Natalie Portman, Christian Bale, Rihanna, Jessica Alba, Katy Perry, Amy Adams, Chris Rock, Oprah Winfrey, Joseph Gordan Levitt, William Windsor, Stephen Colbert, Margot Robbie, Salma Hayek, Leonardo DiCaprio, Jennifer Garner, Scarlett Johansson, Mila Kunis, Shakira, Demi Moore, Neil Patrick Harris, Brittany Murphy, Dax Shepard, Morgan Freeman, Jeffrey Dean Morgan, Alicia Silverstone, Emma Watson, America Ferrara, Cristiano Ronaldo, John Mayer, Jennifer Lawrence, Kate Middleton, Sofia Vergara, Matt Damon, Kirstie Alley, Beyonce, Julie Bowen, Barack Obama, George Clooney, Jake Gyllenhaal, and their doppelgangers.

Of the 40 target celebrities, 31 were paired with doppelgangers who were also celebrities (e.g., target celebrity Michael B. Jordan was paired with doppelganger Nick Cannon), while nine were paired with noncelebrity doppelgangers. In an effort to make the stimulus set racially and ethnically diverse, 50% (20 celebrity pairs) of the stimuli used were people of color. We used human visual inspection across two raters (authors E.P. and L.G.) to approximately match the image quality within each target celebrity and doppelganger image pair. Performance on trials where participants reported that they were unfamiliar with the celebrity (51.4%) was very near chance level (50%), suggesting that images were matched well enough to make response selection based upon image quality an ineffective strategy.

We computed FFDT scores in two different ways. First, "full scores" were computed as the percentage of the 40 trials where the target celebrity was correctly identified; checking the "I don't know this celebrity" button did not affect this score. In addition, "familiar-celebrity scores" were calculated as the percentage correctly identified celebrities only on trials where participants were familiar with the target celebrity; familiar-celebrity scores were calculated only for the 99.1% of participants (n = 56,864) who were familiar with at least 10 target celebrities . Before completing the 40 test trials, participants completed two unscored practice trials, where they were asked to simply differentiate a dog from a cat.

CFMT. The CFMT (Duchaine & Nakayama, 2006) is composed of 72 test trials with six target faces and other distractor faces. There are three increasingly difficult blocks of trials. In the first block, participants are presented with three faces where one is the target. In the second block, participants are presented with the target faces at varying angles. In the final block, participants are presented with the target faces with visual noise obstructing the faces. Scores were calculated as the percentage of trials answered correctly. The CFMT has been found to accurately classify prosopagnosics (Duchaine & Nakayama, 2006) and super-recognizers (Russell et al., 2009). The version of the CFMT used in the present study used computer-generated faces.

CFMQ Seven-Item Version. We used a seven-item shortened version of the CFMQ (Arizpe et al., 2019)—henceforth abbreviated as CFMQ-7—to obtain participants' self-reported facial recognition ability and memory ability. The statements presented in the questionnaire were composed of four that directly related to facial recognition ability (the first four in the list below), and three that assessed general memory ability or pop-culture knowledge (the final three in the list below). The statements were as follows: I can recognize famous celebrities in photos or on TV, I find it hard to keep track of characters in TV shows or movies, I have trouble recognizing familiar people out of context, I can conjure up a vivid visual image of a familiar person's face, I can remember a seven-digit telephone number long enough to write it down, I can recognize famous landmarks in photos or on TV, and I pay attention to pop culture.

For each statement, participants selected one of five response options: never or almost never, not usually, sometimes, frequently, and always or almost always. These responses were recoded numerically between 1 (never or almost never) and 5 (always or almost always). Two items (I find it hard to keep track of characters in TV shows or movies, I have trouble recognizing familiar people out of context) were reverse scored. A total CFMQ-7 score was created by summing the scores of the seven individual items.

Ninety-nine participants (0.6%) failed to respond to at least one of the seven CFMQ-7 items. Of these participants, 44 who failed to respond to over half the items were excluded from analyses involving the CFMQ-7. For the 55 participants who did not respond to one to three items, missing responses were imputed as the mean score of the items participants did complete.

TestMyBrain Digit Symbol Matching. The TestMyBrain Digit Symbol Matching Test is a measure of processing speed adapted from the Wechsler Adult Intelligence Scale, Third Edition (WAIS-III; Wechsler, 1997). Participants were presented with six symbols, each of which was paired with a single number between 1 and 3 (i.e., two symbols were paired with each number); these pairings remained visible throughout the duration of the test. Individual probe symbols were sequentially presented above these pairings, to which participants responded by selecting the corresponding number as quickly as possible. Scores were recorded as the total number of correct responses made in 90 s.

Analyses

For correlations between measures, we report both the observed correlation (r) and the correlation adjusted for the reliability of the two measures being correlated (r_{adj}; Osborne, 2002).

Results

FFDT Performance and Reliability

When calculating full scores across all trials, participants correctly identified an average of 76.0% of faces (SD = 14.1%) on the FFDT (Figure 2); mean performance was very close to the midpoint between maximum possible performance (100%) and chance-level performance (50%). Performance increased with age until peaking for

Table 1. FFDT Performance and Reliability by Age.

	All trials				Familiar-celebrity trials			
Age	*N*	*M* (%)	*SD* (%)	Reliability	% Trials familiar	*M* (%)	*SD* (%)	Reliability
12–19	15,664	71.1	13.1	.74	75.6	77.4	14.2	.71
20–29	18,415	77.5	13.6	.80	80.5	83.1	13.2	.75
30–39	10,296	81.9	13.6	.84	86.9	85.8	12.6	.79
40–49	5,968	79.0	13.7	.82	84.4	83.5	12.8	.76
50–59	4,004	73.8	13.6	.79	78.8	79.7	13.1	.72
60–69	2,015	70.2	13.0	.74	74.4	76.7	13.1	.64
70+	1,045	66.5	13.2	.72	68.9	73.9	13.5	.60

Note. FFDT = Famous Faces Doppelgangers Test.

Table 2. FFDT Performance and Reliability by Race/Ethnicity.

	All trials				Familiar-celebrity trials			
Race or ethnicity	*N*	*M* (%)	*SD* (%)	Reliability	% Trials familiar	*M*	*SD* (%)	Reliability
African/Black	2,726	76.5	13.5	.78	80.8	81.9%	13.4	.74
American Indian/Alaskan Native	1,162	74.6	14.5	.79	81.5	79.2%	14.8	.77
Asian	6,023	75.2	14.2	.79	76.4	81.9%	13.9	.74
European/White	36,073	76.3	14.0	.80	80.2	81.8%	13.5	.74
Native Hawaiian/Pacific Islander	505	77.2	14.7	.81	85.2	81.5%	14.9	.79
Hispanic/Latino	4,340	78.1	13.6	.80	82.4	83.2%	13.3	.76
Missing	7,787	75.2	14.4	.80	80.3	80.6%	14.4	.77

Note. FFDT = Famous Faces Doppelgangers Test.

participants in their 30s, after which performance decreased with increasing age (Figure 3A), consistent with previous studies of face recognition (Germine et al., 2011). Performance was very similar across subgroups of participants reporting different races and ethnicities (*SD* = 1.3%), with the highest-scoring subgroup scoring only 3.5% better than the lowest scoring subgroup (Figure 3B). For trials where participants reported that they were unfamiliar with the target celebrity, performance was near chance level (51.4%). When calculating familiar-celebrity scores, mean accuracy was 81.5% (*SD* = 13.7%), with similar performance patterns between participants of different demographic subgroups as when scoring all trials; see Tables 1 and 2 for the percentage of familiar celebrities by each age and demographic group, as well as the performance on familiar-celebrity trials only. Note that the final three columns of Tables 1 and 2 were calculated using only the subset of participants who reported being familiar with at least 25% of target celebrities.

The FFDT had a Spearman–Brown corrected split-half reliability (r_{xx}) of .80 when scoring all trials (r_{xx} = .75 when restricting analyses to familiar-celebrity trials only and participants who were familiar with at least 25% of celebrities). Similar to test performance, test reliability peaked for participants in their 30s (Figure 3C) and was largely consistent across participants reporting different races and ethnicities (Figure 3D).

We have created a web-based tool (https://rstrong.shinyapps.io/ffdt/) for viewing the distribution of FFDT scores for different age bins, both for scores across all trials and for familiar-celebrity trials only. Although we refrain from suggesting specific cutoff scores for flagging suspected cases of face recognition impairment, this tool can be used to determine such cutoff scores (e.g., 2 *SD* below mean performance), as well as identifying ceiling or floor effects for different age bins (e.g., for 30–39 year olds the distribution of scores is skewed left, indicating a ceiling effect for this age group).

Convergent and Discriminant Validity

CFMT. The CFMT measures face processing and accurately classifies both prosopagnosics (Duchaine & Nakayama, 2006) and super-recognizers (Russell et al., 2009); thus, we expected CFMT performance to correlate with performance on the FFDT. Participants' average accuracy on the CFMT was 85.6% (*SD* = 11.9%). The CFMT had a Spearman–Brown corrected split-half reliability of .92 in our sample. We found a large correlation between performance on the CFMT and performance on the FFDT, both when using all

FFDT trials, $r(16{,}342) = .50$, 95% CI = [.49, .51], $r_{\text{adj}} = .59$, 95% CI = [.57, .60], $p < .0001$ (Figure 4A), and when restricting analyses to familiar-celebrity trials only in participants who were familiar with at least 25% of celebrities, $r(16{,}219) = .53$, 95% CI = [.52, .54], $r_{\text{adj}} = .64$, 95% CI = [.63, .66], $p < .0001$, demonstrating that the FFDT assesses face processing and captures recognition abilities.

To test whether low-range performance on the FFDT was predictive of poor performance on the CFMT, we performed receiver operating characteristics (ROC) analyses (Fawcett, 2006) predicting bottom 5% performance on the CFMT; we did this using familiar celebrity trials in the subset of participants who were familiar with at least 25% of the target celebrities, and who completed both the FFDT and CFMT. We used two different methods for choosing optimal cutoffs using ROC analyses: (a) balanced accuracy, which equally weights sensitivity and specificity, and (b) F1 score, which equally weights sensitivity and positive predictive value (PPV; the percentage of flagged positives that are true positives, which has a chance level of .05 in our analyses). The balanced accuracy approach would be preferred in contexts where follow-up testing is easy (as a relatively high number of false positives will occur), whereas the F1 approach would be preferred when follow-up testing is more difficult.

Using raw scores (unadjusted for age) to predict scoring in the bottom 5% of participants on the CFMT, an ROC analysis optimized on balanced accuracy suggested an FFDT cutoff score of 80.6% (sensitivity = .78, specificity = .75, PPV = .13), whereas an ROC analysis optimized on F1 score suggested a cutoff of 61.7% (sensitivity = .35, specificity = .96, PPV = .28). Using age adjusted scores (i.e., calculating standardized residual scores when using age, age^2, and age^3 as continuous predictors of performance), an ROC analysis optimized on balanced accuracy suggested an FFDT cutoff z-score of -.41 (sensitivity = .76, specificity = .75, PPV = .14), whereas an ROC analysis optimized on F1 score suggested a cutoff z-score of −1.86 (sensitivity = .35, specificity = .96, PPV = .29). Notably, FFDT performance on known-celebrity trials was better in participants who completed the CFMT (n = 16,344, M = 85.9%, SD = 12.5%) than in those who completed only the FFDT (n = 41,063, M = 79.6%, SD = 13.9%), suggesting that the subsample used for this ROC analysis is not representative of the rest of our sample. Therefore, these ROC analyses are provided as a rough metric of the ability of FFDT performance to predict poor CFMT performance, but are not intended as suggestions for cutoff scores in the general population; instead, we suggest determining cutoff scores using deviations from mean performance, which can be calculated for specific age groups using our web-based data visualization tool (https://rstrong.shinyapps.io/ffdt/).

CFMQ-7. Participants' average score on the CFMQ-7 was 26.7 (SD = 4.3). The CFMT-7 had a Cronbach's alpha of .71. CFMQ-7 scores were correlated with full performance on the FFDT, $r(16{,}204) = .48$, 95% CI = [.46, .49], $r_{\text{adj}} = .63$, 95% CI = [.62, .65], $p < .0001$ (Figure 4B); familiar-celebrity performance on the FFDT, $r(16{,}083) = .45$, 95% CI = [.44, .46], $r_{\text{adj}} = .62$, 95% CI = [.60, .64], $p < .0001$; and performance on the CFMT, $r(16{,}204) = .40$, 95% CI = [.39, .42], $r_{\text{adj}} = .50$, 95% CI = [.48, .52], $p < .0001$. In a post hoc analysis using only the four CFMQ-7 questions directly related to face recognition ability (M = 16.0, SD = 3.0, Cronbach's $\alpha = .75$), scores remained predictive of full performance on the FFDT, $r(16{,}204) = .47$, 95% CI = [.46, .48], $r_{\text{adj}} = .61$, 95% CI = [.59, .62], $p < .0001$; familiar-celebrity performance on the FFDT, $r(16{,}083) = .47$, 95% CI = [.46, .48], $r_{\text{adj}} = .63$, 95% CI = [.61, .64], $p < .0001$; and performance on the CFMT, $r(16{,}204) = .45$, 95% CI = [.43, .46], $r_{\text{adj}} = .54$, 95% CI = [.52, .55], $p < .0001$. Thus, the association of the CFMQ-7 with both the FFDT and CFMT is driven nearly entirely by the four CFMQ-7 questions related directly to face recognition ability.

TestMyBrain Digit Symbol Matching. The TestMyBrain Digit Symbol Matching test measures processing speed; thus, we expected a relatively small correlation between performance on this test and performance on the FFDT. The average number of correctly matched symbols was 67.3 (SD = 13.0), and performance was negatively associated with increasing age, $r(7{,}660) = -.34$, 95% CI = [−.32, −.36]. The TestMyBrain Digit Symbol Matching Test had a Spearman–Brown corrected split-half reliability of .99 in our sample. We found a relatively weak correlation between performance on the TestMyBrain Digit Symbol Matching Test and performance on the FFDT, both when using full scores, $r(7{,}660) = .10$, 95% CI = [.07, .12], $r_{\text{adj}} = .11$, 95% CI = [.08, .13], $p < .0001$ (Figure 4C) and familiar-celebrity scores, $r(7{,}571) = .13$, 95% CI = [.10, .15], $r_{\text{adj}} = .15$, 95% CI = [.12, .17], $p < .0001$, demonstrating that the FFDT is a specific measure of face recognition, rather than solely a measure of general cognitive ability.

Performance Validity

For those interested in using the FFDT as an embedded measure of effort, we provide a web-based tool (https://rstrong.shinyapps.io/ffdt/) for exploring how implausibly fast and repetitive (e.g., almost always selecting the face presented on the left) response choices impact performance. For example, performance in participants with a median reaction time of 750 ms or less, or who chose the same answer at least 90% of the time, is very close to chance level (n = 159, M = 51.0%, SD = 6.6%), suggesting that these may be useful criteria for flagging low effort. Notably, we did not exclude any participants based on these criteria in our own analyses, as we cannot rule out the possibility that participants with poor face recognition are those most likely to become disengaged with the test.

Discussion

Successful performance on familiar face recognition tests requires not only successful face recognition ability but also pop-culture knowledge and the recall of information needed to identify recognized celebrities. Here, we report a test designed to more specifically capture familiar face recognition ability: The FFDT. In 57,407 participants, the FFDT had reliable performance variation between participants ($r_{xx} = .80$), with similar performance between participants reporting different races and ethnicities. Critically, FFDT performance was highly correlated with both objectively assessed face memory and self-reported face recognition ability, and had a much smaller correlation with a nonsocial test of processing speed, indicating that the FFDT is a specific measure of face recognition ability (i.e., the FFDT measures face recognition ability much more strongly that it measures general cognitive performance).

FFDT Performance and Reliability

FFDT performance increased with age until peaking for participants in their 30s, after which performance decreased with increasing age (Figure 3A). Although celebrity faces used as stimuli were well-known across all age groups (ranging from 70.2% to 81.2%), performance across age groups was associated with the percentage of celebrities participants were familiar with, indicating that age-related differences in performance may partially be attributable to differences in pop-culture knowledge. However, even when restricting analyses to trials where participants were familiar with the target celebrity, these age-related performance differences persisted, suggesting that face recognition ability may peak in the mid 30s, consistent with prior findings from unfamiliar face recognition tests (Germine et al., 2011; Susilo et al., 2013). Despite age-related differences in performance, FFDT performance was reliable across all age groups (ranging from .72 to .84), indicating that the FFDT is a reliable measure of face recognition across the lifespan.

Performance was largely consistent and reliable across all racial and ethnic groups (Figure 3C, 3D). Although there were small performance differences between each group, notably white participants were not the highest scoring racial subgroup; thus the FFDT avoided a limitation sometimes present in assessments that use predominately white faces as stimuli (Dodell-Feder et al., 2019). Although the FFDT has greater racial diversity in its stimuli than most previous face recognition tests, future updates to the test could further increase the diversity of included celebrities (e.g., by including Asian celebrities). Still, the consistent test performance and reliability of participants across all assessed racial and ethnic groups indicate that the current form of the FFDT can be used to assess face recognition ability in diverse samples of participants; this is an important strength of the FFDT given previous demonstrations of the impact the race and ethnicity of face stimuli have on face recognition performance (Bothwell et al., 1989; Bowles et al., 2009; McKone et al., 2012).

Convergent and Discriminant Validity

FFDT performance was highly correlated with CFMT performance ($r = .50$; Figure 4A), demonstrating that the FFDT is a valid measure of face processing and captures recognition abilities. The CFMT is often used alongside other measures to assess face recognition ability (Corrow et al., 2018; Murray & Bate, 2020), as the test is sensitive to both to impaired (Duchaine & Nakayama, 2006) and extraordinary ability (Russell et al., 2009). Given the FFDT's correlation with the CFMT, the FFDT could potentially be used in cognitive test batteries looking to assess face recognition ability in clinical populations.

In addition to objectively assessed face memory ability, FFDT performance was highly correlated with self-reported face recognition ability and memory as measured through the CFMQ-7 ($r = .47$; Figure 4B). Past research has suggested that subjective assessment tools should be included in recognition measures to accurately assess abilities beyond what is captured by objective measures (Arizpe et al., 2019). The FFDT's high correlation with both objectively (CFMT) and subjectively (CFMQ-7) assessed face recognition ability demonstrates its validity as a measure of familiar face recognition ability.

In contrast to the FFDT's association with other measures of face recognition ability, FFDT performance was only weakly correlated with performance on the TestMyBrain Digit Symbol Matching test, a nonsocial test of processing speed ($r = .10$; Figure 4C). A strong correlation with tests assessing cognitive domains distinct from face recognition ability, such as processing speed, would have suggested that the FFDT assessed general cognitive ability (Wilmer et al., 2012, 2014). Instead, the strong association of FFDT performance with both objectively assessed and self-reported face recognition ability, in combination with a very weak association with processing speed, indicates that the FFDT is a specific measure of face recognition ability.

Relationship to Existing Face Recognition Tests

The FFDT was developed with the goal of emulating the strengths of existing face recognition tests, while simultaneously addressing those tests' weaknesses. For example, the creation of the FFDT was motivated by the Famous Faces Memory Test (Duchaine et al., 2007; Wilmer et al., 2012) and its ability to measure familiar face recognition using popular celebrities. However, by

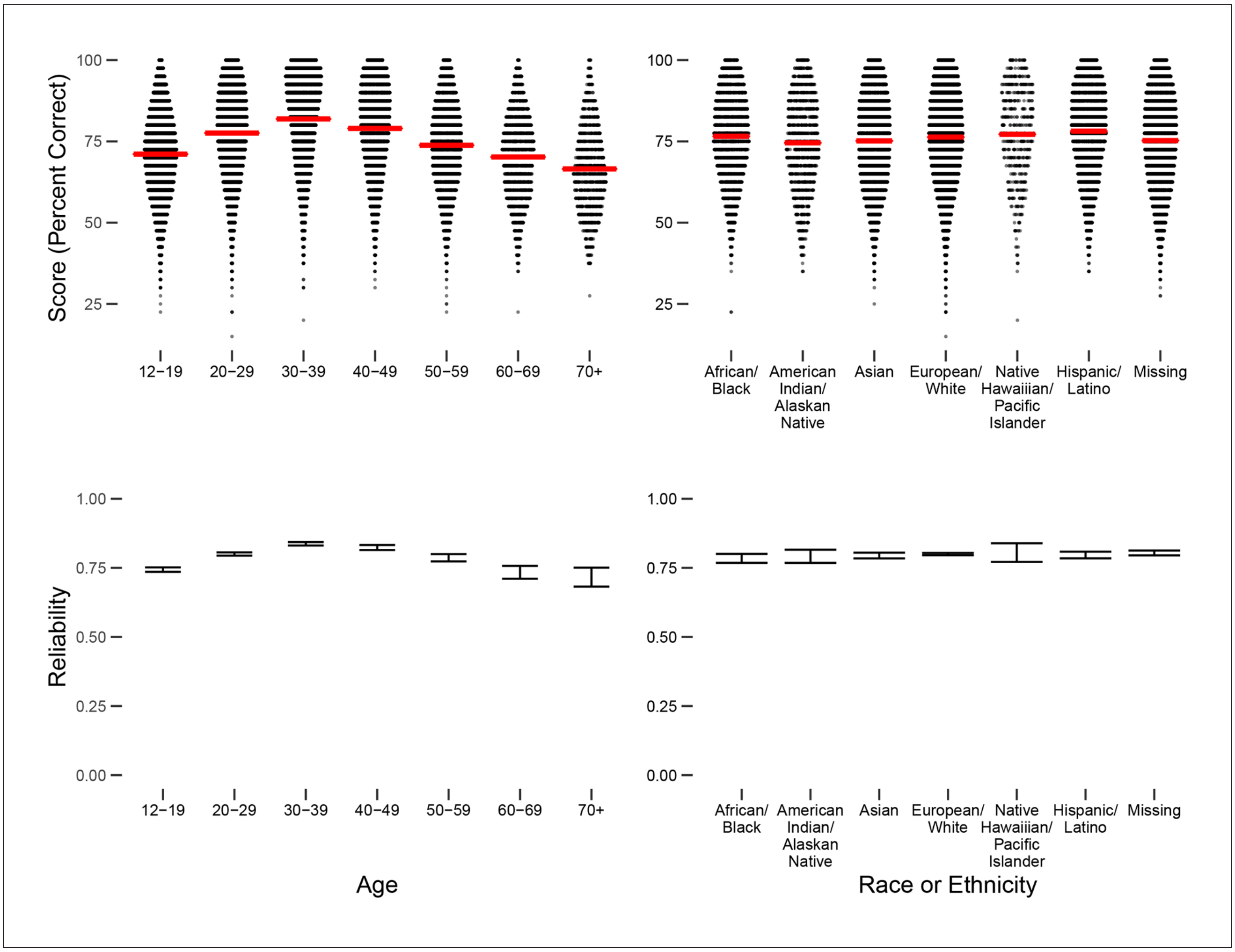

Figure 3. FFDT Performance by Demographic Subgroups: (A) FFDT Performance by Age Bin, (B) FFDT Performance by Race/Ethnicity Category, (C) Reliability by Age Bin, and (D) FFDT Reliability by Race/Ethnicity Category.

Note. For A and B, the width of the distribution corresponds to the proportion of participants who scored at each accuracy level, while the red line corresponds to mean performance. For C and D, error bars show 95% confidence intervals. For B/D, participants are included within each category they reported.

requiring participants to provide identifying information about each celebrity, successful performance on the Famous Faces Memory Test requires successful memory recall in addition to face recognition ability. The FFDT avoids this limitation by presenting participants with the target celebrity's name and using a forced-choice task where participants simply have to select the target celebrity's image. This approach also allows objective and straightforward scoring, rather than having to classify the accuracy of free-text responses or relying on participants to self-score their accuracy.

We adopted the CFMT's strategy of presenting strictly facial information in stimuli, as previous face recognition tasks failed to accurately capture facial processing impairments due to stimuli having nonfacial information and participants using feature-matching strategies (Duchaine & Nakayama, 2004, 2006; Duchaine & Weidenfeld, 2003). Although a well-validated measure of novel face recognition, the CFMT is relatively long (9–12 minutes to complete) and requires participants to remember many images; thus participants' performance may be confounded by their ability to maintain their attention to the task, as well as their working memory capacity. To avoid the confounder of working memory capacity, the FFDT always presented the target celebrity and lure doppelganger simultaneously, thus allowing participants to complete the task without needing to maintain previously viewed faces in working memory. To minimize the likelihood of participants becoming disengaged with the test, the FFDT contains only 40 trials, which on average took 5 to 6 minutes to complete. Given the

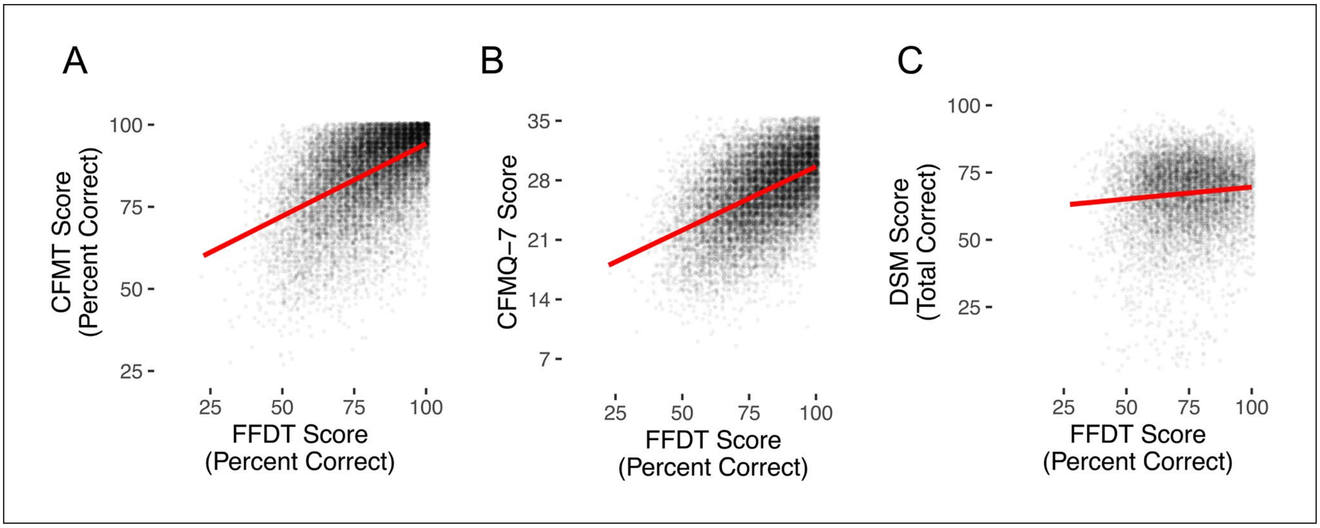

Figure 4. FFDT Convergent and Discriminant Validity: Correlation of FFDT Performance (Full Score Across all Trials) With (A) CFMT Performance (r = .50, r_{adj} = .59), (B) CFMQ 7-Item Version (CFMQ-7) Scores (r = .47, r_{adj} = .61), and (C) TestMyBrain DSM Scores (r = .10, r_{adj} = .11).

Note. Jitter has been added to plots for visualizing point density. Figure axes are scaled so that the physical slope of each red linear best-fit line is equal to the Pearson correlation between each pair of measures. FFDT = Famous Faces Doppelgangers Test; CFMT = Cambridge Face Memory Test; CFMQ = Cambridge Face Memory Questionnaire; DSM = Digit Symbol Matching.

nature of the celebrity stimuli used, we believe the FFDT is entertaining, helping boost participants' engagement and motivation; indeed, the test is one of the highest rated on the TestMyBrain.org cognitive testing platform, with an average rating of 4/5 (average CFMT rating = 3.7/5). As a relatively brief, engaging, and stimulus-diverse task of familiar face recognition, the FFDT has potential utility when used in combination with tests of novel face recognition, including the CFMT.

To our knowledge no celebrity doppelganger tests have yet been published, though at least one is in development (Meschke et al., 2018). This other doppelganger test differs from the FFDT in several important ways, including presenting uncropped images with nonfacial information, using only noncelebrity images as lures, and asking participants to identify the celebrity in each pair of images (rather than specifying the name of the celebrity to select). Although future research is necessary to determine what design parameters produce the most reliable and valid measure of familiar face recognition ability, we believe that the FFDT's use of cropped images on all trials and celebrity lures on 77.5% of trials decreases the likelihood of participants using nonfacial information or familiarity cues to complete the test. Notably, the other doppelganger test (Meschke et al., 2018) has a unique strength of asking participants to rate their familiarity with target celebrities before beginning the main part of the test, which may be useful for more clearly differentiating face recognition ability from pop-culture exposure.

Limitations and Future Directions

Although unsupervised online test administration allowed us to collect a much larger and more diverse sample than would have been possible using traditional in-person assessment, this approach has limitations. For example, we cannot be certain that all participants understood task directions, provided accurate demographic information, remained engaged throughout the task, and participated without any technical problems. Despite these concerns, our sample produced reliable data while demonstrating both convergent and discriminant validity, including expected associations between performance and age (e.g., worse performance on our processing speed task with increasing age). In addition, previous research has found similar results between remote and in-person participation using the TestMyBrain.org cognitive testing platform (Chaytor et al., 2020; Germine et al., 2012).

The FFDT uses cropped images to discourage the use of nonfacial information for identifying target celebrities, yet similar strategies may still be possible. For instance, for the 22.5% of trials where the target celebrity is paired with a noncelebrity doppelganger, picture quality differences could theoretically allow successful performance without face recognition. This does not seem to be a useful strategy, however, as participants were very near chance-level performance for trials where they reported not knowing the target celebrity, indicating that the image quality of target celebrities and lure doppelgangers was well matched.

Although giving participants the option to report that they did not know target celebrities allows us to differentiate face recognition ability from all-or-nothing pop-culture knowledge, it does not account for different levels of celebrity familiarity within people who are aware of the celebrities. A more sensitive measure of celebrity familiarity will likely be useful in future familiar face recognition tests, or participant-specific assessments where overall celebrity familiarity is matched. A related limitation of the FFDT, as with any familiar face recognition test, is the need to update the target celebrities over time to account for the current pop-culture landscape. Although our results demonstrate that the FFDT is a reliable and specific measure of familiar face recognition, more research is necessary to determine the optimal method of stimulus selection in familiar face recognition tests.

Finally, future work is necessary to determine whether the FFDT is useful for diagnosing clinical deficits in face recognition ability. The FFDT's reliability, strong correlations with both self-reported face recognition ability (CFMQ-7) and objectively assessed novel face recognition ability (CFMT, a test previously validated in clinical populations; Duchaine & Nakayama, 2006), specificity, and stimulus diversity make it a candidate for clinical use, particularly if used alongside previously validated measures like the CFMT.

Conclusion

FFDT performance was reliable, similar across participants of different racial and ethnic groups, and more highly correlated with memory for faces and self-reported face recognition ability than processing speed ability. Thus, the FFDT is a reliable and specific measure of the ability to identify familiar faces, making it a promising new tool for assessing face recognition ability.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Research reported in this publication was supported by the National Institute of Mental Health of the National Institutes of Health under award number R01MH121617 to Laura T. Germine.

ORCID iD

Roger W. Strong https://orcid.org/0000-0002-6645-8116

References

Arizpe, J. M., Saad, E., Douglas, A. O., Germine, L., Wilmer, J. B., & DeGutis, J. M. (2019). Self-reported face recognition is highly valid, but alone is not highly discriminative of prosopagnosia-level performance on objective assessments. *Behavior Research Methods*, *51*(3), 1102–1116.

Benton, A. L., Sivan, A. B., Hamsher, K., De, S., Varney, N. R., & Spreen, O. (1983). *Contribution to neuropsychological assessment*. Oxford University Press.

Bothwell, R. K., Brigham, J. C., & Malpass, R. S. (1989). Cross-racial identification. *Personality and Social Psychology Bulletin*, *15*(1), 19–25.

Bowles, D. C., McKone, E., Dawel, A., Duchaine, B., Palermo, R., Schmalzl, L., Rivolta, D., Wilson, E., & Yovel, G. (2009). Diagnosing prosopagnosia: Effects of ageing, sex, and participant–stimulus ethnic match on the Cambridge Face Memory Test and Cambridge Face Perception Test. *Cognitive Neuropsychology*, *26*(5), 423–455.

Chaytor, N. S., Barbosa-Leiker, C., Germine, L. T., Fonseca, L. M., McPherson, S. M., & Tuttle, K. R. (2020). Construct validity, ecological validity and acceptance of self-administered online neuropsychological assessment in adults. *The Clinical Neuropsychologist*, *35*(1), 148–164.

Corrow, S. L., Albonico, A., & Barton, J. J. (2018). Diagnosing prosopagnosia: The utility of visual noise in the Cambridge Face Recognition Test. *Perception*, *47*(3), 330–343.

Corrow, S. L., Dalrymple, K. A., & Barton, J. J. (2016). Prosopagnosia: Current perspectives. *Eye and Brain*, *8*, 165–175.

DeGutis, J. M., Chiu, C., Grosso, M. E., & Cohan, S. (2014). Face processing improvements in prosopagnosia: Successes and failures over the last 50 years. *Frontiers in Human Neuroscience*, *8*, Article 561.

Dodell-Feder, D., Ressler, K. J., & Germine, L. T. (2019). Social cognition or social class and culture? On the interpretation of differences in social cognitive performance. *Psychological Medicine*, *50*(1), 133–145.

Duchaine, B. C. (2000). Developmental prosopagnosia with normal configural processing. *NeuroReport*, *11*(1), 79–83.

Duchaine, B. C., Germine, L., & Nakayama, K. (2007). Family resemblance: Ten family members with prosopagnosia and within-class object agnosia. *Cognitive Neuropsychology*, *24*(4), 419–430.

Duchaine, B. C., & 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.

Duchaine, B. C., & Nakayama, K. (2004). Developmental prosopagnosia and the Benton Facial Recognition test. *Neurology*, *62*(7), 1219–1220.

Duchaine, B. C., & Weidenfeld, A. (2003). An evaluation of two commonly used tests of unfamiliar face recognition. *Neuropsychologia*, *41*(6), 713–720.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874.

Fox, C. J., Iaria, G., & Barton, J. J. (2008). Disconnection in prosopagnosia and face processing. *Cortex*, *44*(8), 996–1009.

Germine, L. T., Duchaine, B., & Nakayama, K. (2011). Where cognitive development and aging meet: Face learning ability peaks after age 30. *Cognition*, *118*(2), 201–210.

Germine, L. T., 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.

Hartshorne, J. K., & Germine, L. T. (2015). When does cognitive functioning peak? The asynchronous rise and fall of different cognitive abilities across the life span. *Psychological Science*, *26*(4), 433–443.

Kennerknecht, I., Grueter, T., Welling, B., Wentzek, S., Horst, J., Edwards, S., & Grueter, M. (2006). First report of prevalence of non-syndromic hereditary prosopagnosia (HPA). *American Journal of Medical Genetics Part A*, *140*(15), 1617–1622.

Kennerknecht, I., Ho, N. Y., & Wong, V. C. (2008). Prevalence of hereditary prosopagnosia (HPA) in Hong Kong Chinese population. *American Journal of Medical Genetics Part A*, *146*(22), 2863–2870.

McKone, E., Stokes, S., Liu, J., Cohan, S., Fiorentini, C., Pidcock, M., Yovel, G., Broughton, M., & Pelleg, M. (2012). A robust method of measuring other-race and other-ethnicity effects: The Cambridge Face Memory Test format. *PLOS ONE*, *7*(10), Article e47956.

Meschke, E., Hacker, C., & Biederman, I. (2018). How many faces can we recognize? *Journal of Vision*, *18*(10), 158–158.

Murray, E., & Bate, S. (2020). Diagnosing developmental prosopagnosia: Repeat assessment using the Cambridge Face Memory Test. *Royal Society Open Science*, *7*(9), Article 200884.

Osborne, J. W. (2002). Effect sizes and the disattenuation of correlation and regression coefficients: Lessons from educational psychology. *Practical Assessment, Research, and Evaluation*, *8*(1), Article 11.

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin & Review*, *16*(2), 252–257.

Susilo, T., & Duchaine, B. (2013). Advances in developmental prosopagnosia research. *Current Opinion in Neurobiology*, *23*(3), 423–429.

Susilo, T., Germine, L., & Duchaine, B. (2013). Face recognition ability matures late: Evidence from individual differences in young adults. *Journal of Experimental Psychology*, *39*(5), 1212–1217.

Thomas, C., Avidan, G., Humphreys, K., Jung, K. J., Gao, F., & Behrmann, M. (2009). Reduced structural connectivity in ventral visual cortex in congenital prosopagnosia. *Nature Neuroscience*, *12*(1), 29–31.

Warrington, E. K. (1984). *Recognition memory test*. NFER-Nelson.

Wechsler, D. (1997). *WAIS-III, WMS-III technical manual*. The Psychological Corporation.

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Gerbasi, M., & Nakayama, K. (2012). Capturing specific abilities as a window into human individuality: The example of face recognition. *Cognitive Neuropsychology*, *29*(5–6), 360–392.

Wilmer, J. B., Germine, L. T., & Nakayama, K. (2014). Face recognition: A model specific ability. *Frontiers in Human Neuroscience*, *8*, Article 769.

Yardley, L., McDermott, L., Pisarski, S., Duchaine, B., & Nakayama, K. (2008). Psychosocial consequences of developmental prosopagnosia: A problem of recognition. *Journal of Psychosomatic Research*, *65*(5), 445–451.