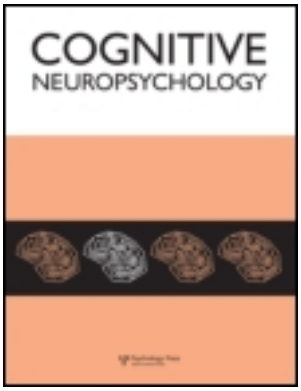

Cognitive Neuropsychology
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/pcgn20

Face ethnicity and measurement reliability affect face recognition performance in developmental prosopagnosia: Evidence from the Cambridge Face Memory Test-Australian

Elinor McKone [a b], Ashleigh Hall [a], Madeleine Pidcock [a], Romina Palermo [a b], Ross B. Wilkinson [a], Davide Rivolta [c], Galit Yovel [d], Joshua M. Davis [a] & Kirsty B. O'Connor [a]

[a] Department of Psychology, Australian National University, Canberra, ACT, Australia

[b] Australian Research Council Centre of Excellence in Cognition and Its Disorders (CCD)

[c] Macquarie Centre for Cognitive Science (MACCS), Macquarie University, Sydney, Australia

[d] Department of Psychology, Tel Aviv University, Tel Aviv, Israel

Available online: 28 Nov 2011

To cite this article: Elinor McKone, Ashleigh Hall, Madeleine Pidcock, Romina Palermo, Ross B. Wilkinson, Davide Rivolta, Galit Yovel, Joshua M. Davis & Kirsty B. O'Connor (2011): Face ethnicity and measurement reliability affect face recognition performance in developmental prosopagnosia: Evidence from the Cambridge Face Memory Test-Australian, Cognitive Neuropsychology, 28:2, 109-146

To link to this article: http://dx.doi.org/10.1080/02643294.2011.616880

COGNITIVE NEUROPSYCHOLOGY, 2011, 28 (2), 109–146

Psychology Press
Taylor & Francis Group

Face ethnicity and measurement reliability affect face recognition performance in developmental prosopagnosia: Evidence from the Cambridge Face Memory Test–Australian

Elinor McKone[1,2], Ashleigh Hall[1], Madeleine Pidcock[1], Romina Palermo[1,2], Ross B. Wilkinson[1], Davide Rivolta[3], Galit Yovel[4], Joshua M. Davis[1], and Kirsty B. O'Connor[1]

[1]Department of Psychology, Australian National University, Canberra, ACT, Australia
[2]Australian Research Council Centre of Excellence in Cognition and Its Disorders (CCD)
[3]Macquarie Centre for Cognitive Science (MACCS), Macquarie University, Sydney, Australia
[4]Department of Psychology, Tel Aviv University, Tel Aviv, Israel

The Cambridge Face Memory Test (CFMT, Duchaine & Nakayama, 2006) provides a validated format for testing novel face learning and has been a crucial instrument in the diagnosis of developmental prosopagnosia. Yet, some individuals who report everyday face recognition symptoms consistent with prosopagnosia, and are impaired on famous face tasks, perform normally on the CFMT. Possible reasons include measurement error, CFMT assessment of memory only at short delays, and a face set whose ethnicity is matched to only some Caucasian groups. We develop the "CFMT-Australian" (CFMT-Aus), which complements the CFMT-original by using ethnicity better matched to a different European subpopulation. Results confirm reliability (.88) and validity (convergent, divergent using cars, inversion effects). We show that face ethnicity within a race has subtle but clear effects on face processing even in normal participants (includes cross-over interaction for face ethnicity by perceiver country of origin in distinctiveness ratings). We show that CFMT-Aus clarifies diagnosis of prosopagnosia in 6 previously ambiguous cases. In 3 cases, this appears due to the better ethnic match to prosopagnosics. We also show that face memory at short (<3-min), 20-min, and 24-hr delays taps overlapping processes in normal participants. There is some suggestion that a

Correspondence should be addressed to Elinor McKone, Department of Psychology, Australian National University, Canberra, ACT 0200, Australia (E-mail: elinor.mckone@anu.edu.au).

Supported by Australian Research Council Grant DP0984558 to E.M. We thank Mary Broughton and Michel Pelleg for participant testing and task scoring in Experiments 2 and 3; Amy Dawel for making the average faces in Experiment 3; Hugh Dennett for scoring the Cambridge Car Memory Task (CCMT); Tirta Susilo for providing Cambridge Face Memory Test (CFMT)-original and Cambridge Face Perception Test (CFPT) scores for Case 1 in Experiment 4; C. Ellie Wilson for designing the Macquarie Centre for Cognitive Science (MACCS) Famous Face Test 2008 (MFFT-08) and testing some controls; Brad Duchaine for providing information on the exact item structure of the original CFMT; and Mike Burton for providing the Glasgow face images. McKone oversaw design of the overall project, conducted the final data analysis, and wrote the paper (with contributions from Hall, Pidcock, Palermo, Yovel, Wilkinson, & Rivolta). Experiment 1 was designed and conducted by Hall, Pidcock, McKone, and Palermo, Experiment 2 by McKone, Experiment 3 by Pidcock, Hall, McKone, and Yovel, Experiment 4 by Palermo, Rivolta, and McKone, and new data in Experiment 5 by Davis, O'Connor, and Palermo. Wilkinson contributed expertise in psychometrics.

http://www.psypress.com/cogneuropsychology http://dx.doi.org/10.1080/02643294.2011.616880

form of prosopagnosia may exist that is long delay only and/or reflects failure to benefit from face repetition.

Keywords: Face recognition; Developmental prosopagnosia; Ethnicity; Memory delay; Measurement error.

Prosopagnosia refers to the inability to recognize faces at the level of individual identity. It occurs in an acquired form after brain injury and also in the developmental or congenital form investigated here, in individuals with no known history of brain insult (e.g., Behrmann, Avidan, Marotta, & Kimchi, 2005; Duchaine, Germine, & Nakayama, 2007; Rivolta, Palermo, Schmalzl, & Coltheart, in press; Schmalzl, Palermo, & Coltheart, 2008). Developmental prosopagnosia has been of theoretical interest in the field of face perception (e.g., Behrmann et al., 2005; Carbon, Grüter, Grüter, Weber, & Lueschow, 2010; Duchaine, Jenkins, Germine, & Calder, 2009, Le Grand et al., 2006; Palermo, Willis, et al., 2011; Rivolta, Schmalzl, Coltheart, & Palermo, 2010; Yovel & Duchaine, 2006). It also has practical significance for affected individuals, being associated with significant functional deficits in everyday life (e.g., need to wear nametags to family reunions in an extended family of prosopagnosics; Duchaine, Germine, et al., 2007) and with problems such as anxiety in social settings (Yardley, McDermott, Pisarski, Duchaine, & Nakayama, 2008). Thus, both researchers and clinicians require access to valid tasks to assist in diagnosis of the disorder.

Diagnosis typically includes a test of the ability to recognize famous faces. A theoretical strength of famous face tests is that they directly measure an ability similar to everyday face recognition—namely, the ability to identify a person, presumably seen many times rather than in a single image, from many hundreds or thousands of possibilities. However, famous face tests also have drawbacks. To ensure that stimuli are in fact famous to the population being tested, faces need to be matched to local conditions and need to be updated over time. Further, participants vary in the extent to which they engage with sources of famous people, such as their interest in popular culture, sports, films, or politics. This problem can be partially dealt with by calculating the number of faces correctly identified as a percentage of the number of target people's *names* with which the subject is familiar (e.g., Lee, Duchaine, Wilson, & Nakayama, 2010; Rivolta et al., in press), but even when this is done it remains possible that apparently "prosopagnosic" performance could arise falsely due to lack of interest in visual media (TV, films, Internet, magazines) where the person's face would be encountered.

Diagnosis also commonly includes novel face learning. The Cambridge Face Memory Test (CFMT; Duchaine & Nakayama, 2006) is illustrated in Figure 1. Participants learn six faces, each in three viewpoints to encourage face rather than image learning. Faces are tested in three stages: recognition of the studied images (*Learn*, for which controls are at ceiling); recognition of the same faces in new images (*Novel*, involving different viewpoint and/or lighting); and recognition of the same faces in new images covered with heavy visual noise (*Noise*). Since its release, the CFMT has quickly become a standard test, used by multiple different laboratories studying developmental prosopagnosia (e.g., Bate, Haslam, Tree, & Hodgson, 2008; Bowles et al., 2009; DeGutis, Bentin, Robertson, & D'Esposito, 2007; Herzmann, Danthiir, Schacht, Sommer, & Wilhelm, 2008; Iaria, Bogod, Fox, & Barton, 2009; Palermo, Willis, et al., 2011; Steede, Tree, & Hole, 2007) and other disorders demonstrating face recognition difficulties (e.g., autism spectrum disorder; O'Hearn, Schroer, Minshew, & Luna, 2010). This popularity is due to a combination of strong demonstrated validity and high measurement reliability.

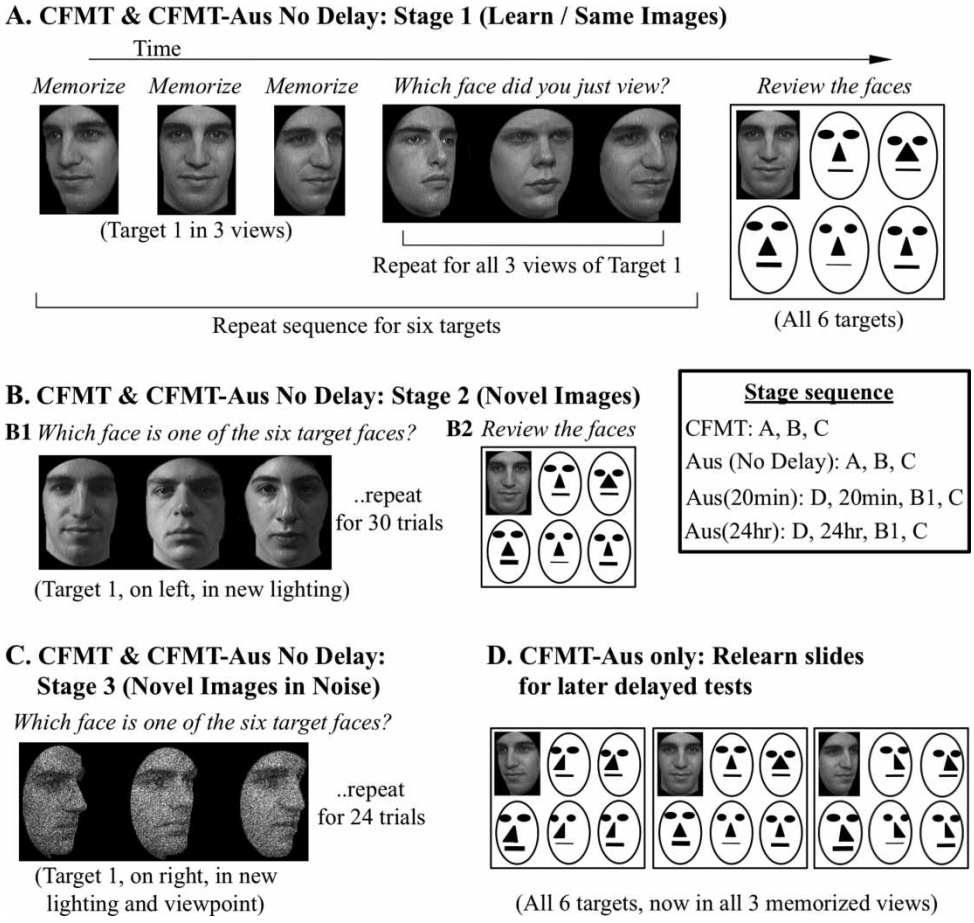

Figure 1. *Illustration of task structure and face format in the Cambridge Face Memory Test (CFMT)-original and the CFMT-Australian (CFMT-Aus). A, B, and C are common to both tests. D is included only in the CFMT-Aus. The 20-min delay condition of the CFMT-Aus begins with relearning (D, presented immediately after C in no delay), then 20-min delay, then repeats B1 then C. The 24-hr condition of the CFMT-Aus begins with relearning (D, presented immediately after C in 20-min test), then 24-hr delay, then repeats B1, then C. The photographic images illustrate the appearance (view, lighting, noise) of the test items in each stage; note that these faces are not actual items from either test (they are the example stimuli provided in Duchaine & Nakayama, 2006. Reproduced with permission). The schematic faces are intended to illustrate the other five target individuals; these appear as photographic images in the tests. Text in italics indicates the participant instructions as given in the tests. Correct answers are illustrated for the examples in B and C.*

Regarding validity, the CFMT requires face recognition rather than hair or clothing recognition (cf. the Warrington Recognition Memory for Faces Test; Warrington, 1984). It also mimics the naturalistic requirement of recognition of faces in different images and viewpoints from previously seen learning images (unlike the Benton Facial Recognition Test, Benton, Sivan, Hamsher, Varney, & Spreen, 1983, which uses simultaneous presentation and allows back-and-forth matching based purely on eyebrows, Duchaine & Weidenfeld, 2003). Validity of the CFMT has also been demonstrated empirically. First, the test substantially improves hit rates for diagnosing developmental prosopagnosia in comparison to the Benton (e.g., Bowles et al., 2009; Duchaine & Nakayama, 2004, 2006). Second, it shows a very substantial inversion effect (Duchaine, Germine, et al., 2007; Duchaine &

Nakayama, 2006; Duchaine, Yovel, & Nakayama, 2007), consistent with expectations from the experimental psychology literature that inversion typically reduces face memory by approximately 15–25 percentage points but reduces within-class discrimination memory for nonface objects by only 0–8 percentage points (e.g., Robbins & McKone, 2007; Yin, 1969). Third, correlation studies demonstrate independence of CFMT performance from general cognitive abilities (very low correlations with verbal memory; Bowles et al., 2009; Wilmer, Germine, Chabris, et al., 2010) and also to a large degree from the rest of visual memory (only a modest correlation with memory for abstract art, Wilmer, Germine, Chabris, et al., 2010, and with cars, Dennett et al., 2011).

Regarding reliability, values for Cronbach's alpha, which assesses internal consistency of a test via split-half reliability taking into account all possible splits, are .88 ($N = 126$ Australians; Bowles et al., 2009), .86–.90 (various groups; Wilmer, Germine, Chabris, et al., 2010), and .83 ($N = 153$ Germans; Herzmann et al., 2008). This reliability is high by the standard of cognitive tasks and is much higher than that of many other face tasks (Herzmann et al., 2008; Susilo et al., 2011; Zhu et al., 2010). It is also high enough to be useful in screening an individual for a possible diagnosis of disorder (Nunnally & Bernstein, 1994).

Rationale for the CFMT-Australian (CFMT-Aus)

The CFMT format offers many advantages, but there are some limitations with the current test. Our laboratories have uncovered several cases of possible prosopagnosia in which the original CFMT test has, in Australian participants, implied a different diagnosis from that arising from a famous faces test. We suggest there are at least three possible sources of this discrepancy: measurement error in the tasks; lack of exact match in ethnicity between the participant (or the participants' typical exposure environment) and the CFMT face stimuli; and testing of only short-delay memory in the CFMT.

The aim of our present study was thus to: (a) develop an additional CFMT-structure test, which (b) uses faces differing in ethnicity (within Caucasians) from the faces in the original CFMT, and (c) add a 20-min and 24-hr delay condition. We now discuss the rationale for each of these aspects of the design.

Replicability and the statistical value of an additional version: Measurement error in an individual's score

Reliability of the CFMT at the individual level is high, but not perfect. With Cronbach's alpha = .88, mean = 55.4 items correct (out of 72), and standard deviation = 8.5 items (the values for $N = 126$ young adults in Bowles et al., 2009), the 95% confidence interval (CI) on a single individual's raw score is ± 6 items (Ley, 1972). Although neuropsychological studies of prosopagnosia have traditionally ignored measurement error in an individual's score,[1] it is an important issue in diagnosis. Specifically, upper and lower bounds of a 95% CI will, for some individuals, translate to different decisions regarding their status as "prosopagnosic" or "normal". For example, an individual with a CFMT score of 41 items correct will have a point z-value of −1.69 (using norms from Bowles et al., 2009). This z suggests they fall in the poorest 5% of the population but not the poorest 2% as typically required for diagnosis of a clinical disorder. However, the *lower bound* of the 95% CI would place this individual as clearly prosopagnosic (score = 35 items, $z = -2.4$, rank in population = poorest 0.8%), while the *upper bound* of the 95% CI would place this individual as completely normal with no suggestion of any deficit (score = 47 items, $z = -1.0$, rank = 15.6%).

This example illustrates that diagnosis status determined using the CFMT can be ambiguous,

[1] Where 95% CIs on a individual's rank in the population are considered, this is usually based on uncertainty in knowledge of the *population* norms (i.e., deriving from small or modest control sample size, particularly $N < 50$; Crawford & Howell, 1998). The 95% CI determined from reliability (even with infinitely large control sample) is a different value.

for simple statistical reasons. Thus, in some cases, there will be statistical value in having available a second test—a test using the same structure and format as those of the CFMT but using different faces—to assess replicability and thus clarify diagnosis status. This issue will be particularly important for potentially prosopagnosic individuals for whom evidence from existing tasks is contradictory (e.g., normal on CFMT-original but impaired on famous faces, or vice versa).

Currently, the only other version available using Caucasian faces is an online version of the CFMT (Germine, Duchaine, & Nakayama, 2010), which has two major disadvantages. First, because the test is easily accessible online, potential prosopagnosics may have already taken it before they come to the laboratory, meaning it cannot be reused, either for diagnosis or for other purposes such as providing a test of generalization to a new item set after a prosopagnosia training programme. Second, the faces (bald male heads) are computer generated using FaceGen (Singular Inversions, Inc.), and most have unusual eye artefacts that give an impression that the face is staring in a manner impossible in natural faces. This impression of bizarre staring could disrupt a participant's attention from the primary task of recognizing facial identity; it could also be problematic if a participant has atypical perception of face expression (as occurs in some prosopagnosics, Duchaine, Murray, Turner, White, & Garrido, 2010) or eye gaze. Thus, there is a need for other versions of the CFMT that are (a) not publically available outside the laboratory or clinic, and (b) use natural face photographs.

Face ethnicity: Can some prosopagnosics be better diagnosed by own-ethnicity faces?

Another potential issue in the original CFMT is the within-race match in face ethnicity to ethnicity of the local test population[2]. The CFMT uses Caucasian faces. However, like any face test derived from a single local database of faces, the CFMT uses faces that reflect the range of specific Caucasian ethnicities available in that location. The source of the CFMT-original stimuli is the Harvard Face Database (Tong & Nakayama), containing images primarily of Harvard University students. This population has a particular demographic profile, which differs noticeably from that in some other European or European-heritage countries. For example, there are demographic differences between Harvard students and Australia. These are best characterized by saying that Harvard/Boston faces are more likely to be of Southern European or Middle Eastern appearance, while Australian faces are more likely to be of British Isles or Northern European appearance. (Note that this north/south difference corresponds to perhaps the broadest distinction that can be drawn within Caucasians.) In Bowles et al. (2009), we estimated the proportion of various demographic groups to be: Jewish 35% (Harvard) versus only $<0.5\%$ (Canberra, Australia); Italian 11.8% (Boston) versus 2.6% (Canberra); and in contrast British 71% (Canberra) versus 33% (Boston). Thus, the CFMT-original face set is well matched ethnically to Harvard participants, and to participants in other locations such as Israel, but less well matched to the ethnicity typical in Australia, Northern Europe, UK, and so on.

In Bowles et al. (2009), we argued that, at a minimum, it is necessary to deal with the ethnic match issue by diagnosing prosopagnosia against CFMT norm data obtained from a similar-ethnicity country/region as the potential prosopagnosic, rather than automatically against the Harvard-based Duchaine and Nakayama (2006) norms. However, this does not necessarily completely solve the problem of ethnic mismatch, because Gilchrist and McKone (2003) reported a

[2] We use "ethnicity" to mean differences between ancestral origin and typical facial appearance *within Caucasians* (e.g., Norwegians typically appear different from southern Italians) and not as a synonym for the larger differences in appearance associated with "race" (e.g., Caucasian versus Asian). We are interested only in Caucasian participants here, because the original CFMT uses Caucasian faces and thus is suitable only for diagnosis of prosopagnosia in Caucasian participants (due to the well-established other-race effect). For readers who wish to test Asian participants, Jia Liu of Beijing Normal University has developed a Chinese face version of the CFMT.

qualitative difference between Australians' response to the Harvard faces and that to local Australian faces: Australians' face memory was improved by increases of featural distinctiveness (e.g., making the eyebrows bushier, or the lips narrower) for both Australian and Harvard faces, but memory was improved by increases in relational distinctiveness (e.g., increasing spacing between the eyes) only for the local Australian faces. This suggests that, where face stimuli are not well matched to local populations, participants can pay undue attention to local features. Importantly, attention to single local features (e.g., eyebrow shape) may prove a useful strategy on laboratory tests, but it is unlikely to lead to successful face recognition in everyday life where there are thousands more possible face identities, and the shape of the local feature alters with changes in view, lighting, make-up, speech, and expression.

This idea raises the possibility that some prosopagnosic participants, while impaired at recognizing own-ethnicity faces, could potentially use specific atypical strategies that allow them to appear normal on tests using other-ethnicity faces. In Australian participants, such strategies might allow some genuine prosopagnosics to appear normal on the CFMT-original. At the same time, they would suffer difficulties in recognizing faces in everyday life (an own-ethnicity environment) and also be revealed as impaired on our local famous faces test (which uses 75% Australian or UK faces).

The present study created a new form of the CFMT using primarily British ethnicity faces, obtained from either Australia (Australian National University Face Database; all 6 targets and some of the distractors) or Scotland (Glasgow Unfamiliar Face Database; remaining distractors). Based on demographics, these faces should be well matched in ethnicity to populations of Australia, New Zealand, UK, and Ireland; they are also likely to be better matched than the Harvard face set to populations in northern European countries and in South Africa. Our question was whether ambiguous cases of prosopagnosia, who were all Australian, might be better diagnosed with the own-ethnicity face set (CMFT-Aus) than with the other-ethnicity face set (the original CFMT). We also tested, in normal-range participants, (a) whether perception of typicality in the two face sets interacted with participants' country of origin (Australia versus Israel), and (b) whether correlational analyses in Australian participants supported the idea that the CFMT-Aus and CFMT-original tap partially different processes.

Relearn-and-delay conditions: Is one form of prosopagnosia an inability to retain faces over long delays?

Many neuropsychological tests involving memory include not only an immediate test, but also a delay phase (e.g., 20-min delay: Warrington Face Recognition Test, Warrington, 1984; California Verbal Learning Test, CVLT, Delis, Kramer, Kaplan, & Ober, 1987). This is because it is possible for individuals to display normal immediate memory in conjunction with impaired longer term memory, for example as found in classic amnesia. Currently, the original CFMT includes only short-delay testing. In the Learn phase, the first test item immediately follows the learning item (very short delay of 500 ms), and in the Novel and Noise stages, the average delay between interstage target review slides (see Figure 1), and a test item is at most 3 min.

The fact that the original CFMT assesses face memory only at short delays may be another reason why it produces conflicting results to famous faces in some individuals. It is already established that some individuals can display intact face perception together with impaired face memory (e.g., Bowles et al., 2009; Lee et al., 2010; Palermo, Willis, et al., 2011; Williams, Berberovic, & Mattingley, 2007), a version of prosopagnosia sometimes referred to a prosopamnesia. Here, however, we are considering the possibility of a distinction within "prosopamnesia", based on the duration of the memory delay (also see Stollhoff, Jost, Elze, & Kennerknecht, 2011). Specifically, there might exist a form of prosopagnosia in which the individual is able to perceive faces normally and to retain faces normally for a short period on immediate memory tasks, but is

unable to retain faces in memory over longer delays. Importantly, both famous face tasks and everyday face recognition require faces to be held in memory over long delays—it might be days, weeks, or months since the participant last saw the target person's face.

If a long-delay subtype of prosopagnosia does exist, then affected individuals would show the following pattern: (a) normal on face perception (e.g., as assessed by the Cambridge Face Perception Test, which uses simultaneous presentation of target and comparison faces); (b) normal on short-delay face memory tests (e.g., CFMT and CFMT-Aus no delay); (c) a deficit emerging on longer delay novel-face learning tests; and (d) a deficit at naturalistic long-delay face recognition as occurs in famous face recognition and everyday life. To assess Stage c, we added two relearn-and-delay conditions to our CFMT-Aus test. Delays were 20 min and 24 hr. We used 20 min because this is similar to delays in established neuropsychological tests and is feasible to include in standard laboratory or clinic testing. The 24-hr delay is less feasible in practical settings, but we tested it here given the absence of any previous information as to how long might be long enough to show up a deficit in any potential "long-delay" prosopagnosics.

We took the simplest possible approach to selecting the items for the delay conditions. At each delay, we simply repeated exactly the same 54 test items as those used in the novel (30 items) and novel-plus-noise (24 items) stages of the basic no delay test. Retesting the same items is a common procedure in other neuropsychological tests (e.g., CVLT, Delis et al., 1987). Also, this method is easily transferable to the original CFMT test: There are not enough photographs of the CFMT-original targets (or distractors) available in the Harvard Face Database for a researcher to be able to set up delay versions of that test using new images in delay conditions, whereas it is straightforward to repeat the last 54 trials as they stand.

We also included review slides showing all six target faces at the end of each delay phase. Opportunities for relearning the targets are already provided within the basic CFMT structure (review slides after the Learn stages, and again after the Novel stage; Figure 1). We added further review slides (Figure 1D), showing all three viewpoints, after the no delay phase (preliminary to the 20-min condition) and again after the 20-min condition (preliminary to the 24-hr condition). This mimics naturalistic face recognition requirements. That is, in everyday settings, when people meet they do not usually have only a single brief learning opportunity, but instead participate in an event of some duration where attention will return to a new to-be-learned person's face on several distinct occasions over some period of time. Similarly, famous faces are usually learned over multiple, spaced, exposures.

Given the item repetition and review slides, the delay conditions are accurately termed *relearn-and-delay* conditions. We expected control participants, who have normal face learning ability, to be able to take advantage of both the relearning slides and the practice with repeated items, and so to partially (or potentially even entirely) overcome effects of increasing time delay reducing memory. Our question was whether prosopagnosics can do the same. Failure to benefit from the additional exposures to faces, and/or poor retention ability over time delay per se, would lead to impaired performance on the relearn-and-delay conditions. Such a result—in combination with normal performance at short-delay and impaired performance for Famous Faces—would support the proposal that some prosopagnosics might show deficits specifically in long-delay retention of faces under naturalistic-style (spaced) learning conditions.

PRESENT STUDY

The structure of the experimental sections is as follows. Experiment 1 presents the new CFMT-Aus and describes its basic psychometric properties including reliability, norms, and convergent and divergent (discriminant) validity; the experiment then uses this new task to addresses the extent to which, in normal subjects, overlapping processes are tapped by the immediate and relearn-and-delay conditions, and by own- and other-ethnicity faces. Experiment 2 further addresses validity,

testing the size of the inversion effect on the CFMT-Aus. Experiment 3 further addresses ethnicity, reporting distinctiveness ratings for the CFMT-Aus and CFMT-original face stimuli, from own- and other-ethnicity participants (Australians vs. Israelis). In Experiment 4, we present CFMT-Aus results for 6 potential prosopagnosics and discuss the extent to which improved diagnosis capability is attributable to better ethnic match, simple measurement error, and addition of relearn-and-delay conditions. Finally, in Experiment 5, we confirm the validity of our *z* scores for the potential prosopagnosics on the original CFMT in Experiment 4, by addressing an issue relevant to many laboratories—namely, whether scores from an individual collected under one testing regime can be fairly compared to control norms collected under another (i.e., using a session structure containing different tasks and/or in a different order).

EXPERIMENT 1: THE CFMT-AUSTRALIAN (CFMT-AUS)

In developing the CFMT-Aus, our aims were to achieve a test with good psychometric properties suitable for both diagnosis purposes and normal-range individual-differences studies. We matched the task format to that of the original CFMT. Our specific aims were to obtain internal[3] reliability $>.85$ and to set task difficulty such that, in the main no delay task, there was both sufficient room below the normal distribution to see prosopagnosic-level performance, yet no ceiling effect so that the distribution was normal. We also present norms for young adults (age 18–32 years).

In addition to the CFMT-Aus, each participant completed the CFMT-original (Duchaine & Nakayama, 2006) and a within-class object discrimination task matched exactly to the CFMT structure (Cambridge Car Memory Task, CCMT; Dennett et al., 2011). These tasks allowed us to assess convergent and divergent validity.

Method

Participants

Participants were Caucasians raised in Australia or New Zealand and so were matched to the face stimuli in terms of ethnic exposure. (New Zealand has similar demographics of Caucasians to Australia; in all experiments, proportion of New Zealanders was $<5\%$). Of our participants, 72% reported solely British Isles ancestry (i.e., England, Ireland, Wales, Scotland). Participants were unselected for face recognition ability beyond the fact that previously diagnosed prosopagnosics were excluded. None reported major head injury or psychiatric disorder likely to affect face recognition and had no personal familiarity with any face used in the CFMT-Aus (relevant because of the locally photographed faces).

The participant sample used to construct control norms comprised 75 individuals (41 female, 34 male; age 18–32 years, mean = 21.67 years, *SD* = 2.96). Two additional participants were tested, and they were excluded from the control norms due to very poor performance on the CFMT-Aus (but were included in the calculations of reliability, because in that situation coverage of the full range is desired). Participants were a convenience sample—namely, members of the Australian National University community, mostly undergraduate students. Average education level would be higher than that in the general population; this is unlikely to limit applicability of norms because face memory does not correlate with general cognitive abilities (IQ, verbal memory; e.g., Bowles et al., 2009; Wilmer, Germine, Chabris, et al., 2010; Zhu et al., 2010). Participants were tested individually and received $24 or 2 hours course credit.

Design

Each participant completed two 1-hour sessions conducted 24 hours apart. The order of tasks was as follows. On Day 1, participants completed: CFMT-Aus no delay; CCMT; filler task to

[3] Note that test–retest is a less appropriate measure of reliability because practice effects from repeating the same items may be different for different participants (Wilmer, Germine, Loken, et al., 2010).

complete 20-min delay period (choice of find-a-word or sudoku); CFMT-Aus 20 min. On Day 2, participants completed: CFMT-Aus 24 hr; CFMT-original; other tasks not relevant to this article.

Apparatus and software

Apparatus for all tasks was CRT-screen eMac with 16-inch monitor running Mac OS X, screen resolution 1,152 × 864, refresh rate 80 Hz, brightness and contrast maximized. PsyScope X software (http://psy.ck.sissa.it; Cohen, MacWhinney, Flatt, & Provost, 1993) presented the CFMT-Aus and CCMT tasks.

CFMT-Aus

Figure 1 illustrates the CFMT-Aus. Availability and conditions of use of the CFMT-Aus are set out in Appendix A. As in the original CFMT, the general structure of the task is that participants learn six target faces each seen in three views. Each test trial then presents a three-alternative forced-choice (3AFC) format display, containing one of the six learned faces plus two nonlearned distractors (all three in a common view and with common external template; hair always excluded), and the participant's task is to select the learned face.

Stimuli

Stimuli were greyscale images of 52 Caucasian males aged in their 20s and 30s (6 targets and 46 distractors, matching the numbers in the original CFMT) with no facial hair or markings. To match the original CFMT, none were noticeably overweight. All 6 targets plus 8 distractors were from the Australian National University (ANU) Face Database (faces photographed in Canberra, Australia, of individuals raised in Australia or New Zealand). The remaining 38 distractors were from the Glasgow Unfamiliar Face Database (Burton, White, & McNeill, 2010), showing individuals photographed in Glasgow, UK. As in the original CFMT, some distractors were repeated across more than one test trial: The number repeated, and how often, was exactly matched to the trial-by-trial structure used in the original CFMT.

Specific viewpoints of images matched those used in the corresponding phase of the original CFMT. Images had neutral facial expression, resolution 72 pixels/inch, and were cropped to include only face and part of neck. Pimples, moles, and any hair over the face were edited out using Adobe Photoshop CS4 Extended Version 11.0. A small amount of noise (greyscale, uniform, strength 2% in the Photoshop "Add Noise" function) was added to ANU database faces to ensure that the ANU targets could not be discriminated easily from Glasgow distractor faces on the basis of sharper image quality.

Average size across all face images appearing in the 3AFC test items (including Learn, Novel, and Noise phases) was 5.2 cm on screen for vertical height of the visible face region (includes some neck and most forehead; see Figure 1). At the approximate viewing distance of 55 cm, this is vertical visual angle 5.4°.

Procedure

No delay. All participants saw the items in constant order. In all stages, on each 3AFC test item the participant pressed a key to indicate which position ("1", "2", or "3") contained the target.

Learn stage (Stage 1; Figure 1A). All faces had top/front lighting. For Target Face 1, three learning images were presented sequentially (one-third-profile-left, then front facing, then one-third-profile-right), each for 3,000 ms with 500 ms interstimulus interval (ISI). A one-third-left 3AFC test item followed, containing the identical Target 1 image to that learned plus two distractor individuals also in one-third-left view, until response. Similar 3AFC test items followed for the front-facing image, then the one-third-right image. This procedure for Target Face 1 was repeated for Target Faces 2–6, giving 18 test trials for the Learn phase (i.e., 3 views × 6 target faces). Finally, participants saw a review slide containing the six target faces presented simultaneously for 20 s, using the front-view learn images.

Novel stage (Stage 2; Figure 1B). Images of each target differed from the learn phase photographs in viewpoint, lighting, and/or shape of external template applied around the face (see Appendix

B for details). There were 30 of the 3AFC trials (6 target faces × 5 viewpoint/lighting/template combinations), presented until response ("1", "2", or "3" to indicate which position contained one of the targets), in fixed pseudorandom order. The same target face never appeared on more than two consecutive trials. At the end, the review slide of the six target faces in front view appeared for 20 seconds.

Novel-in-Noise stage (Stage 3; Figure 1C). Another new set of images of the targets was used, varying viewpoint, lighting direction, brightness and contrast, and/or template shape (see Appendix B for details). Then, heavy noise was applied (coloured, Gaussian, 30% strength using Photoshop "Add Noise" function); this level matched the noise appearance in the original CFMT. There were 24 of the 3AFC trials (6 target faces × 4 viewpoints), presented in fixed pseudorandom order using the same procedure as that for the Novel phase.

Relearn Slides 1 (Figure 1D). Immediately following the last test trial of the no delay phase, relearn slides for the later 20-min delay test were presented. Participants saw three review slides, each for 20 s, with 500 ms between. The first showed all six target faces in the one-third-profile-left images from the Learn phase, the second the front view target images from the Learn phase, and the third the one-third-profile-right images from the Learn phase. If researchers wish to use only the no delay phase of the CFMT-Aus, these review slides can be omitted.

20-min phase. Following a 20-min filled delay (which included completing the Cambridge Car Memory Test), the Novel and Novel-in-Noise sections of the no delay test were repeated (now called Stages 4 and 5). The exact same slides were used as those in Stages 2 and 3, in the same order. The only difference was that the review slide previously between the Novel and Novel-in-Noise stages was removed.

Directly following the last trial of the 20-min-delay test, the same sequence of three relearn slides as that used at the end of the no delay phase was repeated.

24-hr phase. Following a 24-hr delay, the Novel and Novel-in-Noise stages were again repeated (now referred to as Stages 6 and 7). This phase was an exact repeat of the 20-min phase; the only exception was that the relearn slides at the end did not appear.

CFMT-Aus scoring

As in the original CFMT, the measure of interest was accuracy, not reaction time. There were no instructions to participants to respond quickly.

For the *no delay* phase, participants' Total score is the number of trials correct out of 72 (with chance being 24). For the component sections, scores for Learn (Stage 1) are out of 18; for Novel (Stage 2) out of 30; for Novel-in-Noise (Stage 3) out of 24.

For the *20-min relearn-and-delay* phase, Total is number correct out of 54 (chance = 18), comprising Novel out of 30 and Novel-in-Noise out of 24.

For the *24-hr relearn-and-delay* phase, scoring is the same as that for the 20-min phase.

Cambridge Face Memory Test

The CFMT-original was administered in accordance with standard procedure (see Duchaine & Nakayama, 2006). Stimulus size and viewing distance were the same as those for the CFMT-Aus. Scoring was identical to that for the CFMT-Aus no delay phase.

Cambridge Car Memory Test (CCMT)

The Cambridge Car Memory Test was developed and kindly provided by Bradley Duchaine and Raka Tavashmi. The test method and psychometric properties are described in full in Dennett et al. (2011). Briefly, the CCMT follows the same structure and procedure as the CFMT, except stimuli are cars rather than faces. Stimuli are computer-generated images of real car models, all in the same colour, with identifying badges, logos, and insignias removed. Participants learn six target cars, each in three views. On each test trial, they discriminate the target vehicle from two distractors (12 trials per vehicle, 46 distractors in total) matched to the

target for angle and lighting conditions. In the Learn phase, the 18 test trials use the same target image as that learned. In the Novel phase (30 trials), the target car can be any one of the six learned cars and appears in different viewpoint and/or lighting. In the Novel-in-Noise phase (24 trials), the target car can be any one of the six learned cars, appears in different viewpoint and/or lighting, and has Gaussian noise applied.

Scoring was identical to CFMT-Aus no delay. Note that we had valid Cambridge Car Memory Test scores from only $N = 67$ of the control sample; the others were mistakenly given an incorrect version of the experiment script, which presented items in random rather than fixed order.

Results

CFMT-Aus (no delay)

We first discuss results for the no delay phase of the CFMT-Aus. This is the phase equivalent in format to the original CFMT.

Distribution shape and range. Figure 2A shows CFMT-Aus no delay frequency distribution for Total scores. Individuals included in the norm sample are shown in the tan colour. Special-case individuals not included in the norm sample are in blue/red/green (see Appendix B for rationale for excluding MB13 and AM80 from norm sample). Statistics concerning distribution shape are based on the 75 individuals in the norm sample.

The frequency distribution demonstrates that the new test has appropriate properties. Scores in the control sample were normally distributed (Kolmogorov–Smirnov statistic = .098, $df = 75$, $p = .069$) and, most importantly, had no evidence of skew (skew = −0.334, $SE = 0.277$, $z = 1.242$, $p > .1$) including 0 participants scoring the test maximum of 72 items correct. There was also a broad range of scores, making the test valuable for individual differences analysis in the typical population. Yet, the range of scores was also suitable for use in prosopagnosia: The substantial room between the bottom of the normal range and the chance score of 24 items correct allows room to diagnose prosopagnosics and, moreover, to distinguish different severities of prosopagnosia (noting that the chance score of 24 corresponds to $z = -4.6$).

Figure 2B shows that the subtotal of the *Novel + Novel-with-Noise stages* (Stages 2 + 3) was also normally distributed (Kolmogorov–Smirnov statistic = .099, $df = 75$, $p = .065$; skew = −0.355, $SE = 0.277$, $z = 1.282$, $p > .1$). We report this subtotal separately because (a) some researchers may wish to use only these two stages in individual differences studies (because Learn is at ceiling in controls), and (b) it is the *Novel + Novel-with-Noise stages* only that are repeated in the 20-min and 24-hr conditions, meaning that comparison across delays requires this subtotal from no delay.

Norm scores, sex effects, and appropriate method of comparison to norms. For the CFMT-Aus no delay, where scores were normally distributed, comparison to norms in potential prosopagnosics can be achieved in the standard way via *z* scores. (Or, where relevant, via *T*-scores that take into account effects of modest control sample size; see Crawford & Garthwaite, 2002; Crawford & Howell, 1998). That is, in the case of the normal distribution, the number of standard deviations below the mean can be easily converted to the rank order of the case in the population using standard *z* tables.

Table 1 gives means and standard deviations for the norm sample for the no delay phase. There was no sex difference on the means: advantage to females only 0.1 items correct on Total, $t(73) = 0.061$, $p = .952$; for individual stages, smallest $p = .382$. There was also no sex difference on the standard deviations: Levene's test for equality of variances on Total, $F = 1.187$, $p = .279$; for individual stages, smallest $p = .098$. Given the lack of sex differences, we recommend comparing cases to the full control sample regardless of sex. This is because the standard deviation, in particular, is most reliable with the largest sample size (see Bowles et al., 2009).

Comparison to accuracy of original CFMT. Our intention was to set the accuracy of the CFMT-

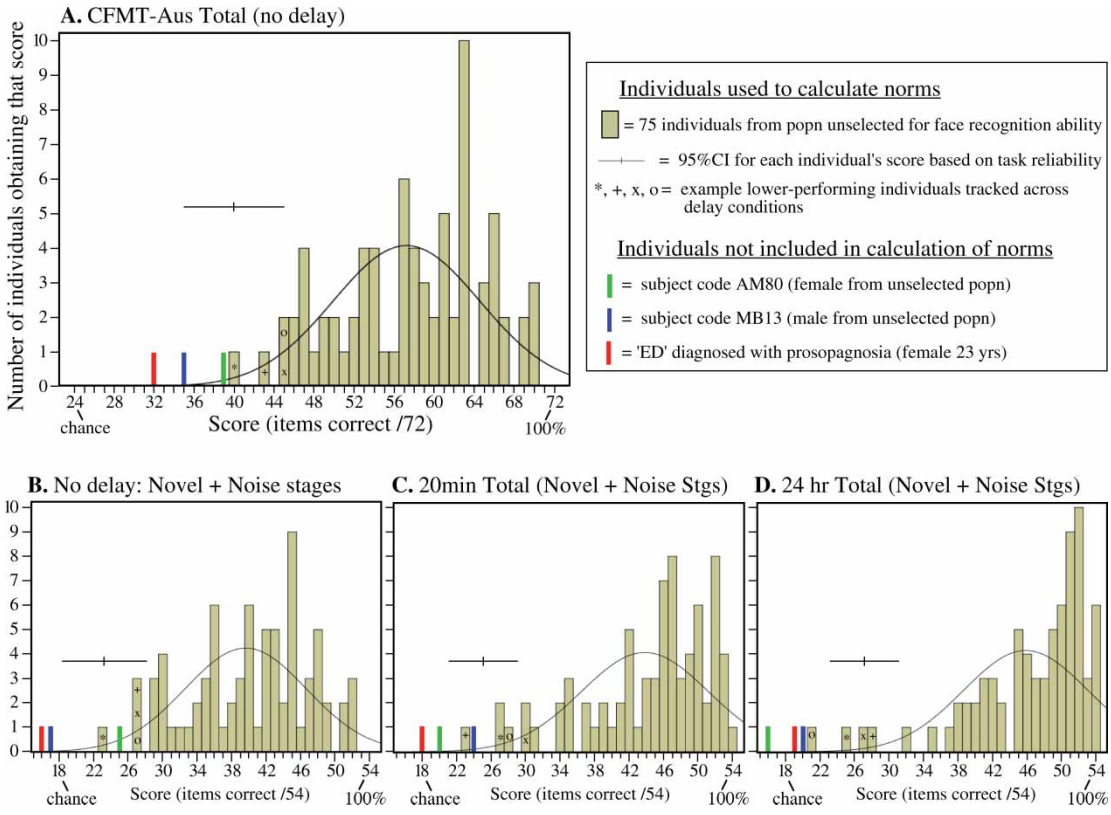

Figure 2. *Experiment 1. Frequency distributions for the Cambridge Face Memory Test–Australian (CFMT-Aus) test, plus best Gaussian fits to the data of the 75 individuals included in the norm sample. Note that distribution is normal at no delay (both for Total scores, and for summed scores of Novel plus Novel-in-Noise stages, see A and B). There is improvement in the relearn-and-delay conditions, with distributions becoming non-normal and bunched against ceiling in the 20-min condition (C) and particularly in the 24-hr condition (D). To view a colour version of this figure, please see the online issue of the Journal.*

Table 1. *Norms for young adults on the CFMT-Aus no delay condition*

	CFMT-Aus Total (Stages 1 + 2 + 3) [Scale range 24–72]			*CFMT-Aus Learn (Stage 1) [Scale range 6–18]*			*CFMT-Aus Novel (Stage 2) [Scale range 10–30]*			*CFMT-Aus Noise (Stage 3) [Scale range 8–24]*			*CFMT-Aus Novel + Noise (Stages 2 + 3) [Scale range 18–54]*		
	M	SD	N	M	SD	N	M	SD	N	M	SD	N	M	SD	N
All upright	57.73	7.34	75	17.63	0.73	75	25.27	3.96	75	14.84	3.98	75	40.11	7.05	75
Females upright	57.78	6.75	41	17.59	0.84	41	25.63	3.43	41	14.56	3.78	41	40.20	6.54	41
Males upright	57.68	8.10	34	17.68	0.59	34	24.82	4.54	34	15.18	4.24	34	40.00	7.72	34
Females inverted	40.00	6.49	16	15.94	1.61	16	15.13	4.11	16	8.94	2.64	16	24.06	5.44	16

Note: CFMT-Aus = Cambridge Face Memory Test-Australian. The standard test has faces upright (Experiment 1); table also includes the inverted version (Experiment 2). Scale range listed is from chance (each trial is three-alternative forced-choice, 3AFC) to maximum value. Minimum value on all scales is 0.

Aus no delay, in our Australian sample of participants, to be similar to the accuracy of the original CFMT, in that test's ethnically matched Boston sample of participants. This was achieved. The mean for CFMT-Aus no delay (80.3% correct, $N = 75$, males and females combined) was almost identical to the mean of the original CFMT in Duchaine and Nakayama (2006; 80.4% correct, $N = 50$, males and females combined). The standard deviations were also similar ($SD = 7.6$ items here; $SD = 7.9$ items in Duchaine & Nakayama 2006).

Analysis by face. We also analysed scores by face. We computed percentage correct (for no delay Total) for each of the six target faces. Sorted from highest to lowest, our six faces produced mean accuracy of: 88, 87, 82, 77, 74, and 74% (see Appendix B for face code numbers). For comparison, the original CFMT in Duchaine and Nakayama (2006) means by face were: 88, 88, 81, 80, 77, and 69%. Note that, in both tests, the origin of differences amongst the faces would be expected to include a mixture of facial distinctiveness and trial order (e.g., an advantage for first-learned face due to lack of proactive interference).

Internal reliability. For CFMT-Aus no delay phase, reliability for the Total score (all 72 trials), including both sexes of participant, was Cronbach's alpha = .88. There were no sex differences, alpha = .88 for females ($N = 42$), alpha = .89 for males ($N = 35$). The full sample reliability value of .88 means that an individual's score can be considered to have a 95% CI error of ± 5 items correct (Ley, 1972). For example, an individual who scored 55 items correct has a 95% confidence interval range on their score of 50–60 items correct out of 72 (i.e., 69–83% in percentage correct).

There was no noticeable difference in reliability if the scores considered are the subtotal of the *Novel + Novel-in-Noise* stages. For these 54 trials, alpha = .88.

We also examined the reliability of the original CFMT (Duchaine & Nakayama, 2006) in our sample. Cronbach's alpha was .84 (all 72 trials, both sexes), which is similar to our previous study reporting .88 for another young-adult sample from the same Australian population ($N = 126$, Bowles et al., 2009).

Convergent and divergent validity (faces versus cars). One way of demonstrating validity of the CFMT-Aus—that is, as tapping face memory rather than general object processing or broad cognitive abilities—is to show both convergent and divergent validity. Convergent validity means that the CFMT-Aus should correlate strongly with other face memory tests, including CFMT-original. Divergent validity means it should correlate weakly with tests of nonface object memory, even where the test uses exactly the same structure and format as the CFMT-Aus (i.e., the Cambridge Car Memory Test). Thus, a finding that CFMT-Aus correlates with the CFMT far higher than it correlates with the car task (CCMT) can only be attributed to shared processes specific to face perception and/or face memory, and not to the generic cognitive components also shared by the two face tasks (i.e., general memory requirements, task concentration, distribution of visuospatial attention).

Correlations were analysed using all participants in the main norm sample ($N = 75$, although note that one had missing data on the CFMT-original, and only 67 had valid scores on the CCMT). We did not include the "special cases" in Figure 2 in this analysis (AM80, MB13, or E.D.) because a research question (which we address in Experiment 4) was whether cases of potential prosopagnosia are diagnosed differently by the different face stimulus sets.

Results indicated convergent and divergent validity. Correlation of the CFMT-Aus with the CFMT-original face task was $r = .61$, $p < .001$ ($N = 74$), which was strong in the context of an upper bound correlation of .86. (The upper bound correlation is given by the product of the square roots of the reliabilities of the two tasks. The observed correlation became .71 when corrected for attenuation due to the less than perfect upper bound correlation; correction for attenuation is given by dividing the correlation between

the two measures by the square root of the product of the reliabilities of the two measures.) The face correlation was not affected by participant sex: female $r = .56$ ($N = 41$), male $r = .68$ ($N = 33$), with no significant difference, $z = .79$, $p = .430$ (two-tailed).

In contrast, the correlation of the CFMT-Aus with the CCMT car task was weak: $r = .21$, $p = .091$ ($N = 67$), and much less than upper bound correlation .86 (calculated using CCMT Cronbach's alpha .84 from $N = 153$ in Dennett et al., 2011). Note that the low correlation cannot be attributed to any problematic psychometric properties of the car task: The CCMT produces a normal distribution, scores do not approach either ceiling or floor accuracy ($M = 53.2$ out of 72 items), and there is a large spread of scores ($SD = 8.3$ items) similar to the CFMT-Aus ($SD = 7.3$ items; Dennett et al., 2011). The car correlation was again not significantly affected by sex: female $r = .32$ ($N = 41$), male $r = .11$ ($N = 28$), $z = .86$, $p = .390$.

Finally, the correlation between the CFMT-Aus and the CFMT-original face task (.62 with $N = 67$) was much larger and significantly different from its correlation with the car task (.21), $t(64) = 3.146$, $p = .0026$ (two-tailed; test for nonindependent correlations).

Overall, these results support the interpretation that the CFMT-Aus taps processes that are substantially, although not entirely, shared by the original CFMT and are very different from the processes tapped by the CCMT car task.

Relearn-and-delay conditions (20 min and 24 hr)

Distribution shape and range. In the 20-min (Figure 2C) and 24-hr (Figure 2D) relearn-and-delay conditions, scores show a bunching up against ceiling. As a result, distributions were non-normal and showed significant skew at 20 min (Kolmogorov–Smirnov = .164, $df = 75$, $p < .001$; skew = −1.012, $SE = 0.277$, $z = 3.653$, $p < .001$) and 24 hr (Kolmogorov–Smirnov = .164, $df = 75$, $p < .001$; skew = −1.516, $SE = 0.277$, $z = 5.473$, $p < .0001$). The improvement in mean performance across the three successive phases was significant. This was conducted comparing the subtotal of *Novel + Novel-in-Noise stages*. Using parametric analyses, the 20-min condition was significantly better than no delay (means of 44.3 and 40.1 items correct, respectively), $t(74) = 7.842$, $p < .0001$, and 24 hr better than 20 min (46.3 and 44.3), $t(74) = 5.235$, $p < .0001$. Nonparametric tests were also conducted given the skew and replicated the findings (Wilcoxon Signed Rank Test, 20 min versus no delay, $z = 6.08$, $p < .001$; 24 hr versus 20 min, $z = 4.8$, $p < .001$). The theoretical origin of this improvement relative to no delay is presumably a combination of relearning opportunities with the targets and practice with repeated test items. Our results argue that, in the normal population, the positive effects of these variables were more than sufficient to offset the general decline in memory that one would expect due to increasing time delay.

Norm scores, sex effects, and appropriate method of comparison to norms. Table 2 provides means and standard deviations for the relearn-and-delay phases of the CFMT-Aus. Both conditions showed a very small female advantage, but this was far from significant (of the six conditions in Table 2, largest $t = 1.774$, smallest $p = .080$). Thus, we again recommend using the combined-sex full sample for norms.

In the 20-min and 24-hr delay conditions, it is not valid to compare an individual case to the norms via z scores. This is due to the non-normal distributions. Instead, to compute the rank order of an individual case within the population—for example, to determine whether they meet a clinical-level impairment criterion such as poorest 2%—it is necessary to have access to the complete frequency distribution of control sample scores. This is provided in Appendix B (or can be extracted from Figure 2). For a simple formula for calculating the point rank, and software able to calculate the 95% CI on this value if desired, see Crawford, Garthwaite and Slick (2009).

Reliability. For relearn-and-delay phases, reliability was even higher than for no delay. For

Table 2. *Norms for young adults on the CFMT-Aus relearn-and-delay conditions*

	CFMT-Aus 20 min									*CFMT-Aus 24 hr*								
	Total (Stages 4 + 5) [Scale range 18–54]			*Novel (Stage 4) [Scale range 10–30]*			*Noise (Stage 5) [Scale range 8–24]*			*Total (Stages 6 + 7) [Scale range 18–54]*			*Novel (Stage 6) [Scale range 10–30]*			*Noise (Stage 7) [Scale range 8–24]*		
	M	SD	N	M	SD	N	M	SD	N	M	SD	N	M	SD	N	M	SD	N
All	44.31	7.36	75	26.96	3.67	75	17.35	4.39	75	46.33	7.22	75	27.39	3.73	75	18.95	4.12	75
Females	44.85	6.36	41	27.49	2.80	41	17.37	4.11	41	47.15	5.97	41	28.07	2.76	41	19.07	3.73	41
Males	43.65	8.47	34	26.32	4.46	34	17.32	4.77	34	45.35	8.48	34	26.56	4.55	34	18.79	4.60	34

Note: CFMT-Aus = Cambridge Face Memory Test-Australian. All faces upright (Experiment 1).

20 min, alpha = .91 for all 54 trials (i.e., *Novel* + *Novel-in-Noise* stages) combining males and females. Split by sex, alpha was .91 (females) and .92 (males). For 24 hr, alpha = .93 for all 54 trials (i.e., *Novel* + *Novel-in-Noise* stages) combining males and females. Split by sex, alpha was .93 (females) and .94 (males).

Correlation between no delay and relearn-and-delay conditions: Same or different processes?

We next computed the correlations across the three phases (no delay, 20 min, 24 hr). These address the issue of the extent to which the relearn-and-delay conditions tap the same psychological processes as do the no delay conditions. Correlations were based on the 78 participants in Figure 2; data were number correct from the 54 items of the *Novel* + *Novel-in-Noise* stages, because these stages were included in all three delay conditions.

The correlation[4] between no delay (Figure 2B) and 20 min (Figure 2C) was $r = .84$. This is very high both in absolute terms and also relative to the upper bound correlation of $r = .90$. The same finding emerged for no delay and 24 hr, again with $r = .84$ compared to upper bound .90. The results argue that, in control participants at least, the processes being tapped by both the 20-min and 24-hr conditions were mostly, although not entirely, the same as those driving recognition with no delay.

We also assessed the value of testing at 24 hr as opposed to the more practical (i.e., within-session) delay of 20 min. The correlation between 20 min and 24 hr was $r = .92$, which was not smaller at all than the upper bound of $r = .92$. This implies that, in controls, no new theoretical construct is being tapped by including the 24-hr condition over and above the 20-min condition.

Discussion

The CFMT-Aus demonstrated good internal reliability, high enough for use in screening individuals in clinical settings. Validity as a test of face memory was supported by convergent and divergent validity (i.e., high correlation with the original CFMT and low correlation with the equivalent-format car task, CCMT). There were no sex differences, meaning potential prosopagnosics can be compared to norms from mixed-sex control samples. The standard no delay condition was normally distributed, and so comparison to controls can be done using *z* scores. Improvement in the 20-min and 24-hr relearn-and-delay conditions led to non-normal distributions, and so comparison to controls must use the full frequency distribution to determine population rank (Crawford et al., 2009).

Regarding our question of ethnic match, results were as predicted by the view that face recognition processes are not identical for own- and other-ethnicity faces. That is, the correlation between the CFMT-Aus no delay and the original CFMT, while strong (.61), was not at the upper bound level (.86). This was despite the fact that we used the same item-by-item structure as that in the original CFMT, the same viewpoints, lighting conditions that were as similar as possible, and so on. Thus, the results support some degree of difference in psychological processes[5] for the two tests. Importantly, the present experiment does not directly address the origin of this difference and particularly whether it is face ethnicity.

[4] All *r*-values in this section are parametric (i.e., Pearson's *r*). Nonparametric correlations were also high (Spearman's rho = .82 for no delay with 20 min; .81 for no delay with 24 hr; .85 for 20 min with 24 hr). We have presented the parametric versions in the text because the upper bounds are based on parametric analyses.

[5] An alternative, theoretically uninteresting, interpretation could be that because the CFMT-Aus and CFMT-original were given on successive days, our upper bound value is falsely biased upwards: The upper bound was determined from internal (i.e., within-session) reliability, and differences in variables such as motivation, alertness, or mood across days could reduce the actual upper bound correlation compared to our calculation. However, arguing against this interpretation, Wilmer, Germine, Loken, et al. (2010) tested the split-half reliability of the original CFMT and found this was only 0.02 lower when the halves were tested across sessions rather than within sessions; and, alternate forms reliability (correlation of CFMT with the online version using FACEGEN stimuli) was 0.03 higher across sessions than within sessions.

Results do, however, support the source of difference being a face-level process. The correlation with the car task was almost identical for the CFMT-Aus (.208) and for the original CFMT (.207). This argues that the lack of perfect (i.e., upper bound value) correlation between the two face tests does not reflect different amounts of reliance on object processing contributions to performance; that is, it was not the case that one was relying more on face-specific mechanisms while the other was more heavily tapping general object-recognition mechanisms.

Regarding the value of adding the relearn-and-delay conditions, the results of Experiment 1 argue that, in control participants at least, these gave only a very small amount of additional information regarding normal-range individuals' abilities. That is, the correlations between no delay and the two later phases were very high and only a small amount below upper bound. Importantly, this does not necessarily imply that the delay conditions would not be valuable in diagnosing prosopagnosia, where no delay and delay could still potentially dissociate (see Experiment 4).

EXPERIMENT 2: INVERSION EFFECT ON THE CFMT-AUS

In Experiment 1, all stimuli in all tasks were upright. In Experiment 2, we further tested validity of the CFMT-Aus by determining the size of its inversion effect—that is, the reduction in accuracy when faces are presented upside down at learning and test compared to upright at learning and test. For the original CFMT, Duchaine and Nakayama (2006) supported the task as tapping face-specific processes by reporting a very large inversion effect of 22 percentage points (upright = 80.4% correct, inverted = 58.4%; note chance = 33%).

Method

Experiment 2 tested only the no delay phase of CFMT-Aus. Everything was exactly the same as in the upright version except that the face stimuli were flipped vertically. There were 16 new participants (mean age = 20.6 years, $SD = 1.9$). All were female; females were easier to recruit than males, and we made a between-subjects comparison to the 41 females from Experiment 1. The literature gives no reason to think that females' face inversion effect is any different in size than males'. Performance for the inverted version could be fairly compared to that for the earlier upright version because the no delay condition was tested first up in Experiment 1.

Payment was $5 for the 10-min session if participants completed this task only. Participants completing other tasks in a longer session received $15 for 1 hour or undergraduate course credit. Participants were from the same pool as that in Experiment 1 and were tested individually.

Results

The CFMT-Aus demonstrated a very large inversion effect (Table 1). On Total score (all 72 trials), the inversion decrement on mean accuracy was 24.7 percentage points (upright = 80.3%, inverted = 55.7%), $t(55) = 9.025$, $p < .001$. There was no difference in variance between upright and inverted (Levene's test, $F << 1$, $p = .990$). The inversion effect on means was also significant for each stage separately (Learn, $p = .001$; Novel, $p < .001$, Novel-in-Noise, $p < .001$).

We also examined the internal reliability of the CFMT-Aus Inverted. Cronbach's alpha ($N = 16$) was .68. The 95% CI derived from this reliability and the standard deviation in Table 1 are ± 7 items correct (e.g., for a measured score of 45, the 95% CI range is 38–52; Ley, 1972). This compares to ± 5 items for females upright, and the inverted reliability of .68 was noticeably lower than that for upright (.88 for females).

Discussion

Overall, results of the inversion experiment argue strongly for the validity of the CFMT-Aus as specifically tapping face processing. Its inversion effect (25%) is as large or larger than that in

previous face studies (typical range 15–25%, e.g., Diamond & Carey, 1986; Robbins & McKone, 2007; Yin, 1969) and much larger than inversion effects for objects: Multiple studies report only small inversion effects ranging from 0–8% in any one recognition memory experiment for cars, dogs, clothing, and so on (e.g., de Gelder, Bachoud-Levi, & Degos, 1998; Reed, Stone, Bozova, & Tanaka, 2003; Robbins & McKone, 2007; Scapinello & Yarmey, 1970; Yin, 1969). Theoretically, large face inversion effects are associated with holistic coding of upright faces but not inverted faces (e.g., for review, see Rossion, 2009; also our particular stimuli, McKone, 2008).

Turning to the lower reliability for CFMT-inverted, this phenomenon has also been previously reported for another face task (Cambridge Face Perception Test; Bowles et al., 2009) although not previously given theoretical interpretation. For upright, the high reliability argues that all 72 items tap, to a large extent, a common process or processes used for recognizing all upright faces. The lower reliability inverted argues for more internal variability across items in the processes used.

EXPERIMENT 3: DISTINCTIVENESS OF CFMT-AUS AND CFMT-ORIGINAL FACES IN AUSTRALIAN AND ISRAELI PARTICIPANTS

In Experiment 1, the correlation between CFMT-Aus and CFMT-original was below upper bound, indicating the use of partially different processes for the two face sets. Given that the two face sets are taken from different subpopulations of Caucasians, with different demographics (i.e., more British/Northern European for the CFMT-Aus, and more Southern European/Middle Eastern for the CFMT-original), the origin of the memory differences could potentially lie in own- versus other-ethnicity faces. However, this could only remain plausible if (a) there are physical differences between the two face sets, and (b) participants can perceive these in a manner that is clearly associated with ethnic match. Experiment 3 tests these proposals.

To test for *physical* differences, we examined morphed average faces of the people who appear as stimuli in the CFMT-Aus and CFMT-original. To test *perception*, we obtained distinctiveness ratings (i.e., how much a face would "stand out in a crowd") for the individual faces in a full cross-over design in which the CFMT-Aus and CFMT-original faces were rated by Australian and Israeli participants. Given demographic information, ethnic match to the everyday life exposure of the participants is higher for CFMT-Aus faces than for CFMT-original faces in Australian participants, but the reverse in Israeli participants. Thus, if participants perceive the match in ethnicity, we predict a cross-over interaction in which Australians rate the CFMT-Aus faces as less distinctive (i.e., more typical) than the CFMT-original faces, and vice versa in the Israeli participants.

Method

Participants

Australians were 14 new participants (10 female; mean age = 23.7; $SD = 2.2$) from the same pool as that in Experiment 1, with 71% ($N = 10$) reporting solely British Isles ancestry. They received course credit or $5 (30 min testing).

Israelis were 13 Tel Aviv University undergraduates (9 female; aged 20 to 30 years, mean age = 22.6 years, $SD = 2.7$) who received course credit. All were Caucasian and raised in Israel; 10 had Jewish names and 3 Arab.

Stimuli

Both for creating average-face morphs and for the distinctiveness rating task, the face stimuli were all front view, standard lighting (i.e., from the top and/or front), and without noise. This was because (a) it is impossible to average across 2D images of different face views, and (b) distinctiveness ratings could be affected by the use of unusual views, unusual lighting, or noise. Our aim was to include as many of the people appearing in the CFMT-Aus and CFMT-original as possible, even where they appeared in those tests only in a

nonfront view or with poor lighting. Where necessary, we extracted front-view photographs from the original databases. The final stimuli included all 52 of the CFMT-Aus faces and 40 of the 52 CFMT-original faces (all 6 targets and 34 of the distractors; missing ones were where the Harvard Face Database did not contain a usable image of the person in front view/lighting). Faces were formatted as shown in Figure 3. A common cropping template was applied to images from all databases. All faces were presented on a black background. Brightness and contrast were matched across databases. A small amount of monochromatic, uniform noise (3 or 4% in the Photoshop "Add Noise" function, depending on the photo) was added to the ANU and Glasgow database faces to match image quality to the Harvard database photos (which had been taken with a lower resolution camera).

Procedure

Morphed average faces (average of 52 faces for CFMT-Aus, and average of 40 faces for CFMT) were created using software from FaceResearch.org (Lisa DeBruine and Ben Jones, of University of Aberdeen; available at http://www.faceresearch.org/demos/average).

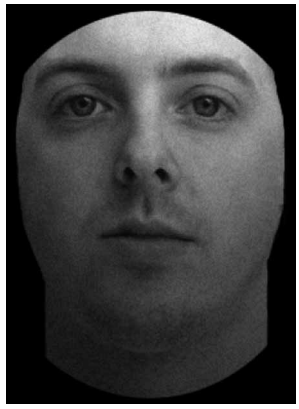

Figure 3. *Illustration of face format for the distinctiveness rating task, where all images were front-view, top/front lighting versions of the people whose photographs appear in the Cambridge Face Memory Test (CFMT)-original and CFMT-Australian (CFMT-Aus). The face shown is not an actual test item.*

For the rating task, participants rated all 92 individual faces one at a time, with faces from the different sources randomly intermixed and, crucially, with no mention of different sources, of ethnicity, or of social groups; that is, participants were given face information only to make their judgements. Participants were told to imagine they were at their local shopping centre and to rate each face for how difficult it would be to "spot in the crowd". Each face appeared alone in the centre of the screen until response, together with the rating scale: 1 "not very distinctive" (difficult to spot in a crowd), to 9 "very distinctive" (easy to spot in a crowd).

The rating task was presented using PsyScope X software (http://psy.ck.sissa.it; Cohen et al., 1993), on CRT screens with 17-inch monitors, specifically an eMac running Mac OS X (Australia) and Mac Power PC G4 Dual 450 MHZ running OS X 10.4.11 (Israel).

Regarding stimulus size, for Australian participants screen resolution was 1,152 × 864, average vertical height of the stimuli (including some neck, some forehead, no hair; see Figure 3) on the screen was 6.2 cm, and viewing distance was approximately 55 cm, giving vertical visual angle of 6.5°. For Israeli participants, screen resolution was 1,024 × 768, average vertical height of stimuli was 7.0 cm, and viewing distance was approximately 65 cm, giving vertical visual angle of 6.2°.

Results

The morphed CFMT-Aus average and CFMT-original average confirmed the existence of physical differences between the stimulus sets. The nature of the differences in the average faces is illustrated and described in Figure 4.

Results of the rating study (see Figure 5) then illustrate a cross-over interaction: Observers rated as more distinctive the individual faces from the set with average ethnicity less common in their home countries. A two-way analysis of variance (ANOVA) found a significant interaction between subject group (Australian, Israeli) and face set (CFMT-Australian, CFMT-Original), $F(1, 25) = 18.401$, $MSE = 0.281$, $p < .001$.

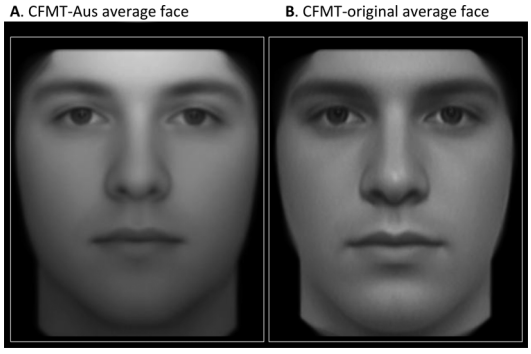

Figure 4. *Morphed average faces of front-view versions of stimulus faces from (A) Cambridge Face Memory Test–Australian (CFMT-Aus; Australian and Glasgow origin faces) and (B) CFMT-original (Harvard origin faces). Images are matched for height of eyes and width between eyes. White boxes are identical for the two faces, allowing comparison of face width). Relative to the CFMT-original average, the CFMT-Aus average is broader, shorter in vertical distance from eyes to mouth, has a shorter nose that is also smaller at the end, has a different jaw shape, has thinner lips, has less bushy eyebrows, and also has eyes that are slightly less round on the top.*

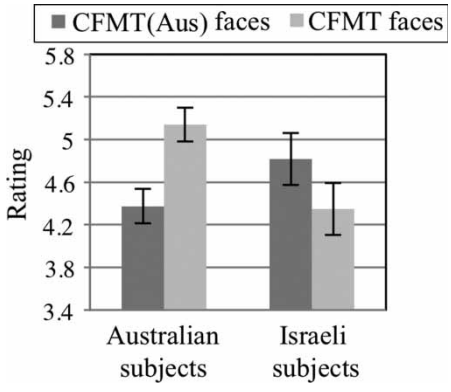

Figure 5. *Experiment 3. Distinctiveness ratings showing cross-over interaction based on within-race match in ethnicity. CFMT = Cambridge Face Memory Test. Rating scale, for "how much would this face stand out in the crowd at a busy shopping centre", is 1 = "not at all" to 9 = "very much". Error bars show* ±1 SEM *of the within-subject difference score (i.e., as appropriate for the comparison of the CFMT-Aus and CFMT-original means within a participant group).*

Follow-up t tests showed that Australians rated the CFMT-Australian faces as significantly less distinctive (i.e., more typical) than the CFMT-original faces, $t(13) = 4.765$, $p < .001$, and the opposite-direction trend in the Israeli participants approached significance, $t(12) = 1.926$, $p = .078$.

Also of interest, there was no main effect of face set, $F(1, 25) = 1.055$, $MSE = 0.281$, $p = .314$. That is, averaged across both groups of observers, distinctiveness ratings for the CFMT-Australian faces were no different from those for the CFMT-original faces. This argues that the CFMT-Australian stimuli are well matched to the stimuli from the original CFMT in terms of intrinsic physical distinctiveness (i.e., independent of the particular ethnic group of the observer).

Discussion

Experiment 3 confirms both that there are physical differences present between the CFMT-Aus and CFMT-original face sets, and that these are differently perceived by observers in a manner consistent with their status as own versus other ethnicity. These results demonstrate the plausibility of the hypothesis that the "different processes" tapped in memory by the two tasks in Experiment 1 (and Experiment 4) are differences in processes applied to same- and other-ethnicity faces.

EXPERIMENT 4: ABILITY OF THE CFMT-AUS TO ENHANCE DIAGNOSIS OF DEVELOPMENTAL PROSOPAGNOSIA

Our results from our analyses so far imply that (a) the CFMT-Aus is a valid test of face recognition, (b) it has high reliability in the no delay and in both relearn-and-delay conditions, and that (c) the exact face processes it taps are largely but not entirely in common with those tapped by the original CFMT, with the differences plausibly attributed to the differences in face ethnicity.

Our question now concerns the ability of the CFMT-Aus to assist in the diagnosis of developmental prosopagnosia. Although the hit rate of the CFMT-original is far better than that of previous face recognition tests (the Benton or the Warrington; see Bowles et al., 2009; Duchaine & Nakayama, 2006), we have still observed a number

of cases of people who report everyday symptoms consistent with prosopagnosia, but who show internally contradictory performance on behavioural tests.

The 6 potential prosopagnosics we considered here all self-reported as having everyday face recognition difficulties (e.g., trouble following the identities of the characters in films, failing to recognize colleagues out of their usual setting, etc.). They were selected because all had a conflict between the diagnosis status suggested by a famous face test and by the CFMT-original. Five were impaired (defined as point z score in poorest 2% of population) on identifying famous faces yet were within the normal range on the CFMT-original. One showed the reverse pattern. The intention here was to see whether the CFMT-Aus could resolve diagnosis and, where possible, comment on whether this was due to increased statistical reliability from adding another test, the better match in ethnicity (all participants were Australian), or the addition of relearn-and-delay conditions.

Method

Participants

The 6 participants (labelled Case 1 through Case 6) all self-reported to our laboratories as having everyday face recognition difficulties. For example, Case 2 said: "I regularly sit in meetings with people, sometimes for hours but can't recognize them afterwards, i.e., in a different setting", and "I tend to look at hair or clothes, but when clothes change I'm stuck", and "I think people think I'm ignoring them and rude when I walk past them in the corridor. They may say hello and I have no idea who they are." Participants varied in age from 19 to 50 years when tested on the CFMT and CFMT-Aus. All were Caucasian. Five were raised in Australia; the remaining individual (Case 5) was raised in England and had been living in Australia for approximately 8 years at time of testing. Regarding ethnic background, 5 reported 100% British Isles ancestry; the remaining individual (Case 3) had a mix of Belgian, French, and English ancestry. None reported known brain injury.

A criterion for selection was that poor performance on face recognition tasks could not be attributed to general factors such as low cognitive abilities or lack of motivation. Table 3 shows that all 6 performed completely normally on the CCMT car task (z scores ranging from −1.04 to +1.27), indicating good general learning and memory skills. Also note that the car test was conducted in the same session as two phases of the CFMT-Aus (no delay and 20 min); thus, impaired performance on CFMT-Aus cannot be attributed to participants having a "bad day". Regarding intelligence, Raven's matrices scores were available for 3 participants and were average or above average (Raven Coloured Progressive Matrices; Raven, Raven, & Court, 1998); it was not felt necessary to test intelligence formally in the other 3 participants because all held at least bachelor-level university degrees (2 were current doctoral-level students), implying above-average intelligence (Table 3).

Interestingly, all 6 participants were completely normal on the Cambridge Face Perception Test (CFPT; Duchaine, Germine, et al., 2007), with no evidence of even mild impairment, z score range −1.14 to −0.04; Table 3); this task involves ordering morphed faces by similarity to a simultaneously presented different-view target face. Other articles have noted previously that, where participants have been tested on the CFPT in addition to the standard face memory tests used for diagnosis (e.g., Famous Faces), it appears quite common to find developmental prosopagnosics who fail to recognize faces only on the tests involving memory (Lee et al., 2010; Palermo, Willis et al., 2011).

For present purposes, the CFPT results confirm that our participants had no general deficits in cognitive abilities and also allow us to rule out any suggestion that the participants were specifically unmotivated to try hard on tasks involving faces. On this latter point, we also had available autism spectrum quotient scores (AQ; Baron-Cohen, Wheelright, Skinner, Martin, & Clubley, 2001) for 5 participants, and all were in the normal range (highest observed score = 26, where score above 32 is required to be suggestive of autism

Table 3. *Potential prosopagnosic participants in Experiment 4*

Subject code	*Sex*	*Age (years)*	*CCMT (cars)*[a]	*Intelligence*	*CFPT*[b]	*AQ*[c]
Case 1	F	19–21	43 ($z = -1.04$)	Raven = 12/12 (popn rank >95%)	34 ($z = -0.04$)	n/a
Case 2[d]	F	50–51	53 ($z = 0.36$)	Raven = 9/12 (popn rank = 50)	50 ($z = -1.14$)	22
Case 3	F	47	56 ($z = 0.78$)	Bachelor degree + graduate diploma	40 ($z = -0.39$)	5
Case 4	F	25	54 ($z = 0.50$)	PhD student	36 ($z = -0.28$)	26
Case 5	M	50	68 ($z = 1.27$)	Raven 12/12 (popn rank >95)	46 ($z = -0.40$)	20
Case 6	F	25	55 ($z = 0.64$)	Doctoral level student	34 ($z = -0.11$)	6

Note: Sex: F = female, M = male. Age: in years, at time of testing. Nonface object memory: Cambridge Car Memory Test (CCMT). Information relevant to intelligence: Raven Coloured Progressive Matrices (popn = population); or education level. Face perception: Cambridge Face Perception Test (CFPT). AQ = autism spectrum quotient. Note that there was no evidence of impairment on any task.

[a]CCMT z scores were calculated using sex-specific norms for young adults from Dennett et al. (2011). [b]CFPT coding is such that scores higher than the mean are worse than average and are shown with negative z scores. CFPT z scores are calculated compared to norms from Bowles et al. (2009), which are age adjusted (because the CFPT is a speeded task and age-related worsening of performance begins by 40 years). [c]Autism spectrum quotient scores are considered indicative of potential autism spectrum disorder (ASD) if ≥32 (Baron-Cohen et al., 2001). We were unable to test Case 1 on AQ, but there was nothing in her everyday behaviour to suggest ASD. [d]Case 2 was previously referred to as SP_50 in Bowles et al. (2009).

spectrum disorder, ASD; Table 3). Thus, there was no reason to suspect that our participants displayed lack of social interest.

Procedure

The CFMT-Aus and CCMT were tested using exactly the same procedure and session structure as those in Experiment 1 (with the exception of change to a laptop computer for some participants; note that stimulus size measured in visual angle remained as for Experiment 1). Scores on other tasks, including CFMT-original, had generally been collected in other sessions, in most cases prior to the testing for the present study. In Experiment 5, we present evidence that session structure does not affect performance on the CFMT.

Our famous face task was the MACCS Famous Face Test 2008 (MFFT-08; Palermo, Rivolta, Wilson, & Jeffery, 2011b; Palermo, Willis, et al., 2011; Rivolta et al., in press). This uses faces famous in Australia and measures accuracy for naming (or providing other identifying information about) 20 famous faces as a percentage of the number of the famous people familiar to participants from their names. Ethnicity is heavily biased towards Australian/British: Of the 20 target faces, 11 are Australian, 4 are from the UK, and 5 are from the US.

Results

As noted earlier, our 6 participants were chosen specifically because their diagnosis status was ambiguous, with apparent conflict between their results for Famous Faces and CFMT-original. Table 4 illustrates that, using the traditional criteria of "impaired" as point z score in the poorest 2% of the population (i.e., $z < -2.05$ where measure is normally distributed; indicated in bold in Table 4) and "not impaired" as point z score not in the poorest 2%, each participant was either impaired on our famous faces test and simultaneously not impaired on the CFMT-original, or vice versa.

Turning to their CFMT-Aus performance (Table 4), our participants grouped into three categories, as follows.

Face ethnicity

First, in three cases, the CFMT-Aus resolved the previous discrepancy in favour of demonstrating prosopagnosia. In Table 4, Cases 1–3 all showed clearly impaired performance on the CFMT-Aus. This is inconsistent with their CFMT-original results (even including the lower bound of their 95% CIs on this task; see Table 4), but is consistent with their self-reports of real-world difficulties, and their famous face performance. This

Table 4. *Face memory scores for the potential prosopagnosic participants in Experiment 4*

Subject code		*Famous faces (MFFT-08)*	*CFMT original (Harvard faces)*	*CFMT-Aus*			*Our categorization*
				No delay	*20 min*	*24 hr*	
Case 1	Raw score	22 (4/18)[a]	58	32	18	19	prosopagnosic
	z score	**−3.83**[b]	0.33[c]	**−3.51**[d]	n/a[e]	n/a[e]	
	Point rank	**0.01**[b]	62[c]	**0.02**[d]	**< 0.1**[f]	**< 0.1**[f]	
	95% CI rank	n/a[b]	[35−85][c]	[**< 0.01**−0.24][d]	[**0.0**−*3.3*][f]	[**0.0**−*3.3*][f]	
Case 2	Raw score	30 (6/20)[a]	42	37	27	27	prosopagnosic
	z score	**−2.40**[b]	−1.55 (−1.39)[g]	**−2.82**	n/a	n/a	
	Point rank	**0.82**[b]	6.0	**0.24**	*2.7*	*3.3*	
	95% CI rank	n/a[b]	[*2.2*−20]	[**0.02**−**1.6**]	[**0.3**−8.9]	[**0.7**−9.5]	
Case 3	Raw score	33 (6/18)[a]	46	41	26	27	prosopagnosic
	z score	**−2.29**[b]	−1.08 (−0.99)[g]	**−2.28**	n/a	n/a	
	Point rank	**1.1**[b]	13.9	**1.1**	**1.3**	*3.3*	
	95% CI rank	n/a[b]	[*3.7*−35]	[**0.15**−5.5]	[**0.1**−6.1]	[**0.7**−9.5]	
Case 4	Raw score	53 (10/19)[a]	36	49	37	38	poorer end of
	z-score	***−1.75***[b]	**−2.30**	−1.19	n/a	n/a	normal range
	Point rank	*4.0*[b]	**1.2**	11.7	17.3	10.7	
	95% CI rank	n/a[b]	[**0.15**−6.0]	[*3.1*−31]	[9.9−27]	[*4.9*−20]	
Case 5	Raw score	27 (4/15)[a]	40	55	41	39	poorer end of
	z-score	**−2.61**[b]	−1.79 (−1.62)[g]	−0.37	n/a	n/a	normal range
	Point rank	**0.45**[b]	*3.7*	36	25.3	13.3	
	95% CI rank	n/a[b]	[0.**63**−13.9]	[14.6−62]	[16−36]	[6.8−23]	
Case 6	Raw score	47.4 (9/19)[a]	44	49	30	33	? mildly
	z-score	**−2.21**[b]	−1.34	−1.19	n/a	n/a	prosopagnosic
	Point rank	**1.4**[b]	9.4	11.7	6.7	6.7	with relearn
	95% CI rank	n/a[b]	[*2.2*−27]	[*3.1*−31]	[*2.3*−14.4]	[*2.4*−14.0]	and delay

Note: Face memory scores for 6 people reporting face recognition difficulties in everyday life whose scores on Famous Faces and CFMT-original are apparently contradictory. CFMT = Cambridge Face Memory Test. CFMT-Aus = CFMT-Australian. MFFT-08 = Macquarie Centre for Cognitive Science (MACCS) Famous Face Test 2008. CI = confidence interval. Values in bold fall in poorest 2% of population (i.e., rank < 2, $z < -2.05$); values in bold italic fall in poorest 5% (i.e., rank < 5, $z < -1.65$).

[a]Famous Face scores are percentage correctly identified (and, in parentheses, the number of faces correctly recognized out of the number of people that participant indicated familiarity with from the name). [b]For MFFT-08, norms for *z* scores (and corresponding population rank) were derived from Australian Caucasians aged 19−72 years recruited from MACCS and the general Sydney community ($N = 28$) using the fit-and-residual procedure described in Bowles et al. (2009). Note that 95% CI on population rank is not reported for MFFT-08 (Famous Faces) because it is not clear how to calculate internal reliability for this task (the number of trials used varies across participants). [c]For CFMT-original, norm values used to obtain *z*, population rank, and 95% CI rank, were $M = 55.23$, $SD = 8.51$ (see Table 5 in Experiment 5), reliability = .88 (chosen as an approximate average of values in present study, Bowles et al., 2009, and Wilmer, Germine, Chabris, et al., 2010; note that we ignored the lower .83 value from Herzmann et al., 2008, as this could have been reduced by their very long testing session). Because the control distribution was normal, 95% CIs on ranks were obtained by: calculating lower bound raw score and upper bound raw score (based on reliability and *SD*; see Ley, 1972); converting this to *z*; converting *z* to population rank. [d]For CFMT-Aus no delay, norm values used to obtain *z*, population rank, and 95% CI rank were $M = 57.73$, $SD = 7.34$, reliability = .88 (control data from Experiment 1). Because distribution is normal, 95% CI calculated as for CFMT-original. [e]For 20-min and 24-hr conditions, *z* scores not reported because they are invalid due to non-normal control distributions. [f]For 20-min and 24-hr conditions, point ranks and 95% CIs were calculated based on full control distribution (from Experiment 1), using software described in Crawford et al. (2009). [g]For the 3 older participants, alternative CFMT-original *z* scores in parentheses are shown calculated relative to Bowles et al. (2009) age-adjusted norms. Because age-related decline is minimal on this task prior to 50 years, age adjusting makes little difference to the results.

Table 5. *Norms used for CFMT-original, based on combined data*

	CFMT Total (Stages 1 + 2 + 3) [Scale range 24–72]			*CFMT Learn (Stage 1) [Scale range 6–18]*			*CFMT Novel (Stage 2) [Scale range 10–30]*			*CFMT Noise (Stage 3) [Scale range 8–24]*		
	M	SD	N	M	SD	N	M	SD	N	M	SD	N
All	55.23	8.51	248	17.71	0.69	248	22.38	4.76	248	15.14	4.16	248
Females	55.85	8.77	144	17.72	0.73	144	22.78	4.87	144	15.35	4.37	144
Males	54.38	8.11	104	17.71	0.63	104	21.83	4.56	104	14.84	3.85	104

Note: CFMT = Cambridge Face Memory Test. The norms are those that we used here for the CFMT-original (Harvard stimuli), based on combining several studies (Bowles et al., 2009, young adults; present Experiment 1; "fatigue" condition from Experiment 5) to obtain a very large sample size.

is evidence that the CFMT-Aus can produce a different conclusion from the CFMT-original and can be useful in diagnosing prosopagnosia. That is, three out of four measures agree for these participants and imply prosopagnosia, and the odd one out is the original CFMT.

In terms of the theoretical origin of the extra diagnosis capability, in these three cases it was not due to the addition of the delay conditions. For Cases 1–3, impairments at no delay were at least as strong as their impairments in the relearn-and-delay conditions. It also seems unlikely that the difference between the CFMT-Aus and CFMT-original was due to measurement error on the tasks. Particularly for Cases 1 and 2, there is a very striking difference in z scores for the CFMT-Aus no delay and CFMT-original. For both these individuals, there is no overlap in the 95% CIs for the two tasks. That is, for Case 1, the upper bound of the CFMT-Aus is a population rank of 0.24, while the lower bound of the CFMT-original is a rank of 35; for Case 2, the corresponding values are 1.6 and 2.2.

This leaves differences in ethnicity of the faces as the most plausible source of the difference in diagnosis, with more accurate diagnosis arising from the use of stimuli ethnically matched to the participants. Self-report from 2 participants further supports this hypothesis. After testing, Case 2 commented to the researcher that the Australian version "was harder than the original CFMT because on that task she could use more features (eyebrows) than in this Australian version". Case 3 also reported attempting to use eyebrows to recognize the faces and said that this was possible for two of the CFMT-Aus faces that were distinctive to her, but that the other four were very hard to recognize. Note that these self-reports are consistent with previous evidence that prosopagnosics sometimes try to use local part strategies to perform laboratory face tasks (Behrmann & Avidan, 2005; Bentin, De Gutis, D'Esposito, & Robertson, 2007; Duchaine & Weidenfeld, 2003; Lee et al., 2010) and also with the evidence that, in control participants, Australians find that attention to distinctive local features is relatively more useful in remembering the Harvard face set than the local Australian face set (Gilchrist & McKone, 2003). Thus, our explanation of the most likely reason why prosopagnosia has been revealed only on the ethnically matched CFMT-Aus, and not on the mismatched CFMT-original, is (a) the prosopagnosics attempted to discriminate the faces via local features, (b) they attempted this strategy for both own- and other-ethnicity faces, but (c) the local feature strategy is more effective (i.e., making the participant's score appear more normal) where local features are perceived as more distinctive, which occurs more often for other-ethnicity faces (CFMT-original) than for own-ethnicity faces (CFMT-Aus).

Measurement error

Second, we have two cases where the CFMT-Aus results resolved the discrepancy in favour of concluding that the individuals were at the poorer end of the normal range but not poor enough to be labelled prosopagnosic. Cases 4 and 5 were not impaired (using point rank) on any phase of the CFMT-Aus. Moreover, in general the differences between their population ranks on different tests are able to be explained by measurement error on the tasks (i.e., the 95% CIs overlap). This is most obvious in the case of Case 4. Her point rank even on the famous faces task in fact overlaps with her 95% CI on CFMT-original. That is, the apparent discrepancy between famous faces and the CFMT—based on point population ranks in both tasks—is potentially attributable simply to measurement error on the tasks. Her CFMT-Aus findings are consistent with this interpretation: The 95% CIs on all stages of the CFMT-Aus overlap with her 95% CI on the CFMT-original.

Relearn and delay

Finally, we consider Case 6. She demonstrated normal performance on the perception task (CFPT), the CFMT-original, and also the CFMT-Aus no delay, yet clinically impaired level of performance on famous faces. In this case, the results of the relearn-and-delay conditions of the CFMT-Aus are illustrative. Case 6′s point rank in the population is perfectly normal where no memory requirement is involved (percentile rank of 46 on CFPT), drops somewhat when a short-delay memory requirement is introduced (rank in poorest 9.0 and 11.7% on CFMT-original and CFMT-Aus no delay), drops again with relearning and delay (rank 6.7 for both 20-min and 24-hr delay), and drops even further with the types of spaced exposures and long delays present in everyday life (rank 1.7 on famous faces). These suggest that Case 6 may be genuinely prosopagnosic (albeit at a mild level, given her 95% CIs), but specifically her difficulty is in retaining learned faces over long delays and/or in failing to take the usual benefit from relearning when faces are repeated.

Discussion

These results demonstrate that the CFMT-Aus can provide a useful tool for diagnosis, particularly where there is conflict between famous faces and the CFMT-original. In participants who have self-reported as possible prosopagnosics, we have illustrated three cases where the CFMT-Aus no delay results confirm the diagnosis of prosopagnosia, apparently due to the better match in ethnicity. We have also illustrated two cases in which the pattern of results across all tests suggests performance in the lower end of the normal range rather than prosopagnosia, and, importantly, the apparent differences between tasks can be largely or entirely explained simply by taking measurement error into account. Finally, we have described one case in which the participant appears to show a coherent pattern of increasing difficulty relative to controls in conditions with longer delays and more learning repetitions.

EXPERIMENT 5: DOES SESSION STRUCTURE AFFECT CFMT-ORIGINAL NORMS?

Our Experiment 4 conclusions regarding the ability of the CFMT-Aus to enhance prosopagnosia diagnosis rely on the assumption that our *z* scores were valid for the CFMT-original. However, as is common in neuropsychological testing, our potential prosopagnosic participants were tested on the CFMT-original in different session structures from the control participants. Thus, the *z* scores will only be valid if there are no effects of session structure on norms.[6]

Here we tested this proposal. We were specifically interested in the possible effects of *practice* with the general task format—due to earlier

[6] Note that this issue does not affect the *z* scores for the CFMT-Aus or CCMT, because controls and potential prosopagnosics were tested on identical session structures for those tasks.

exposure to a similar face or object learning task—and *fatigue*, arising where the CFMT-original was completed late in an experimental session. We assessed the influence of these variables on the CFMT-original by comparing mean performance in typical individuals across three studies using different session structures.

Method

All three studies used participants from the same population—namely, young adult Caucasians aged 18–32 years raised in Australia/New Zealand, unselected for face recognition ability. All three contained mixed-sex samples, and the proportion of women was similar in each.

Our "baseline" group was from Bowles et al. (2009; $N = 114$, 59% female). For these participants, the CFMT was the first test conducted, in a single session, meaning there were no fatigue effects from previous tasks and also no prior practice with any CFMT-like face learning (or object learning) tasks.

Our "practice" group was the control participants from present Experiment 1 ($N = 75$, 55% female). For these, CFMT-original was conducted on Day 2 after extensive practice at CFMT-style tests (two phases of the CFMT-Aus using faces and the CCMT using cars, on Day 1, plus the final phase of the CFMT-Aus on Day 2). However, there was little opportunity for offsetting effects of fatigue, because CFMT-original started only 10 min into the Day 2 session.

Our "fatigue" group was a new sample for whom there had been no previous practice with CFMT-style tasks, but for whom the CFMT-original was the sixth element in a continuous session, beginning after 1 hr of completing other tasks (demographic questionnaire, facial emotion matching task, two anxiety questionnaires, facial emotion labelling task, and verbal emotion labelling). There were 58 participants (35 females, i.e., 60% female; ages 18–32 years, mean age 22.6 years, $SD = 3.2$), some tested individually ($N = 21$) and others in groups (third-year undergraduate laboratory class, $N = 37$), combined here because there was no effect of individual versus group testing (mean CFMT accuracy for individually tested = 77.9%; for group tested = 77.1%). They were members of the Australian National University community, mostly undergraduate students.

Results

Compared to the baseline group, any effect of practice predicts *improvement* in performance, and any effect of fatigue predicts *worsening*. Results showed neither. For the "baseline" group, mean CFMT-original score was 55.1 items correct, or 76.5% ($N = 114$). For the "practice" group, performance was not better than baseline, with mean CFMT-original score 54.8 items correct, or 76.1% ($N = 75$). And, for the "fatigue" group, performance was not worse than baseline, with mean CFMT-original score 55.5, or 77.1% ($N = 58$).

Given these findings, the CFMT-original z scores in Experiment 4 were based on combining the data from the three samples. This gave the most reliable young-adult norms for Australian/New Zealand participants (Table 5; $N = 248$; we used norms collapsed over sex because there was no significant sex difference on either mean, $p = .182$, or SD, $p = .589$).

Discussion

We conclude that performance on the original Duchaine and Nakayama (2006) CFMT was unaffected by practice (with other similar-style tasks) or fatigue (within the range of a 1.25-hr testing session). This validates our CFMT-original z scores for our potential prosopagnosics. Our results also support the validity of the common practice of diagnosing prosopagnosia against control norms obtained from other studies where session structure for the norm participants was probably different from that for the potential prosopagnosics.

GENERAL DISCUSSION

The key conclusions from the present study are as follows. First, face ethnicity within a race has

subtle but definite effects on face processing even in normal participants. Second, face ethnicity can significantly impact prosopagnosia diagnosis, with hit rate apparently increased by using own-ethnicity faces. Third, the CFMT-Aus, which complements the CFMT-original by using face ethnicity better matched to the other half of the European population, is a valid and reliable test. Fourth, in some cases of prosopagnosia, apparent conflict between tasks can be attributed simply to measurement error. Fifth, face memory at short (<3-min) delay taps overlapping processes as 20-min and 24-hr relearn-and-delay in normal participants, although there is some suggestion that a form of prosopagnosia may exist that is revealed with long delays and/or spaced face learning repetitions.

In addition, we have shown that scores from prosopagnosics can be validly compared to those of controls tested with other session structures.

Finally, our results also imply that the creation of truly parallel forms of the CFMT (e.g., for prosopagnosia training studies) requires equating ethnicity of the face stimuli in the two versions.

We now discuss these findings, plus some more peripheral results, in more depth.

Theoretical implications for understanding face recognition in typically developing individuals

Does within-race ethnicity affect face recognition processes?

The existence of other-race effects on face memory is well established. However, the question of whether exact ethnicity of faces within a race influences face recognition has received less attention. Results of four previous studies have implied that within-Caucasian ethnicity might matter, using two measures—mean face memory performance (Bowles et al., 2009; Chiroro et al., 2008; Sporer, Trinkl, & Guberova, 2007) and relative sensitivity of memory to distinctiveness of featural versus relational information (Gilchrist & McKone, 2003). However, note that of these studies, only Sporer et al. (2007) tested a full cross-over design—that is, two ethnicities of Caucasian participants tested on two ethnicities of Caucasian faces—as is ideally necessary to demonstrate that ethnic *match* of participants to stimuli is the relevant variable. Further, the Sporer et al. (2007) two-way interaction involving memory in Turkish and German faces/participants appears to have been at best very weakly replicated in Sporer and Horry (2010; see their Figure 1).

In this context, our results provide important new evidence that participants' face processing is affected by within-Caucasian ethnic match. A cross-over design—Face Stimuli (CFMT-original vs. CFMT-Aus) × Participants (Israeli vs. Australian)—testing perception of distinctiveness showed a two-way interaction of the required form, with other-ethnicity faces rated as less typical than own-ethnicity faces (Experiment 3). This effect must be driven by bottom-up perceptions of physical differences between the faces, rather than any top-down influence of social groups, given the lack of any mention to participants of groups or ethnicity. We also observed that, in Australian participants, the correlation between performance on the CFMT-Aus and that on the CFMT-original (.61) was below upper bound (.86), indicating that memory also relied on partially different processes for the two face sets (Experiment 1). Finally, we found that some participants were diagnosed as prosopagnosic only when tested with own-ethnicity faces (Experiment 4).

Overall, the results argue that other-ethnicity faces are perceived (Experiment 3) and remembered (Experiments 1 and 4) using partially different processes from own-ethnicity faces. Although our results do not specify exactly which processes differ between own- and other-ethnicity faces, our results did argue that the difference is not greater reliance on nonface object recognition mechanisms (i.e., CFMT-original correlation with car task was no higher than CFMT-Aus correlation with car task); instead, it appears to be a difference in face-level processing. One possibility is differential reliance on certain subcomponents of the face recognition system, such as unusual reliance on local feature strategies (associated with occipital face area; Pitcher, Walsh, Yovel, &

Duchaine, 2007) for other-ethnicity faces. This is suggested by Australians' sensitivity to featural but not relational distinctiveness in the Harvard faces (Gilchrist & McKone, 2003) and by the self-report from 2 of our prosopagnosic participants that the CFMT-original was easier because they could use the eyebrows more effectively in the CFMT-original (Experiment 4).

An open question is how small ethnicity differences can be and still affect face processes. In the present study, the ethnicity difference between the face stimuli was perhaps the broadest distinction that can be drawn within Caucasians—that is, deriving from southern European/Middle Eastern ancestry versus British Isles/Northern European ancestry (this also corresponds approximately to speakers of Romance versus Germanic languages). The fact that we have found other-ethnicity effects for these face types does not necessarily mean there would be similar effects for finer variations in within-race appearance. For example, other-ethnicity effects would seem unlikely for Scottish versus Danish faces.

Does upright and inverted face recognition tap different processes?

In addition to confirming the expected large inversion effect for faces (24.7%), we found that internal reliability for the CFMT-Aus was lower for inverted than for upright. Together with a similar finding we reported previously on the Cambridge Face Perception Task (Bowles et al., 2009), this argues that the specific mechanisms used to process inverted faces are less consistent from trial to trial than those used to process upright faces. This reliability evidence provides a novel approach to confirming differences in processing between upright and inverted faces, adding to substantial other evidence indicating such a difference (for reviews, see McKone, 2010; Rossion, 2009). Theoretically, a possible origin of the reliability differences is that normal participants use a holistic processing strategy consistently for all upright faces, while the part-based processing used for inverted faces is more variable because participants can choose to focus on different specific parts across trials or faces.

Sex differences in face recognition ability

Previous studies have found that women show, on average, better memory than men for female faces (Lewin & Herlitz, 2002; McKelvie et al., 1993). For male face stimuli, previous results have been conflicting. Three studies have found no female advantage (Duchaine & Nakayama, 2006; Lewin & Herlitz, 2002; McKelvie et al., 1993). In Bowles et al. (2009), we reported a numerically small female advantage on the CFMT-original, which was not significant with $N = 126$ young adults, but was significant with $N = 236$ adults aged 18–88 years. At that time, we attributed our finding to increased power arising from a larger sample compared to the original studies. Our present results disagree with that interpretation. In Experiment 5, we added more participants to our norms for the CFMT-original test for young adults (i.e., now $N = 248$, cf. $N = 126$ in Bowles et al.), and the female advantage was numerically tiny (1.1%) and not significant ($p = .182$). The CFMT-Aus, which also uses male faces, also showed no sex difference in means (Experiment 1). It is possible that the significant female advantage in our earlier paper arose from the older adults who were included in the large-sample analysis (we reported in Bowles et al. that a female advantage on the Cambridge Face Perception Test emerged only in older adults).

Use of CFMT-style tests in diagnosing prosopagnosia

We now discuss issues in prosopagnosia diagnosis relevant to all CFMT-style tests.

Replicability and statistical uncertainty in diagnosis

Unless a task has perfect reliability, the raw score produced on that task for a single participant has uncertainty associated with it (Ley, 1972). Our reliability of .88 for the CFMT-Aus, together with the task standard deviation, gives a 95% CI of ± 5 items correct (out of 72). The reliability and standard deviation of the original CFMT are similar, also giving an error of ± 5.

In some individuals, this will not matter for diagnosis, because even the full 95% CI range

would place them as either clearly impaired or clearly normal. However, in some cases, the point *z* score will indicate a "no deficit" (i.e., more than 2 *SD*s below mean) on one test (e.g., CFMT-original) and "deficit" (i.e., less than 2 *SD*s below mean) on another (e.g., CFMT-Aus, or Famous Faces) purely due to statistical chance. In the present article, we have presented 2 individuals for whom basic task replicability issues—that is, measurement error—appear to have been the origin of their previously conflicting results between tasks (Experiment 4).

Ethnicity can matter in prosopagnosia diagnosis

A key result of the present paper is that there can be significant diagnosis value in having available an ethnically matched version of the CFMT. We reported three cases in which an own-ethnicity CFMT-format test (CFMT-Aus) indicated prosopagnosia—consistent with Famous Faces scores and reports of face recognition difficulties in everyday life in the affected individuals—while an other-ethnicity CFMT-format test (CFMT-original) failed to pick up these cases.

In terms of why ethnicity matters, a plausible hypothesis based on self-reports of strategies from the prosopagnosics is that they try to use distinguishing local features to discriminate the targets, and that these features are more distinctive and therefore more discriminating in other-ethnicity face sets than in own-ethnicity face sets. We do not, however, wish to argue that this is the only possible route by which ethnicity could affect prosopagnosics' performance. For example, there is some recent evidence from adaptation aftereffects (Palermo, Rivolta, Wilson, & Jeffery 2011a) that complex face-space dimensions might be less finely tuned in developmental prosopagnosics than in controls; this might make it more difficult to discriminate amongst "typical" own-ethnicity faces because these lie close to the centre of face-space and have many competing neighbours, while still allowing some ability to discriminate amongst "distinctive" other-ethnicity faces if these lie in lower density regions of face-space (cf. Valentine, 1991).

Is there a need to create ethnically matched versions of the CFMT for multiple world locations?

At the same time as emphasizing that ethnicity can have a major effect on diagnosis in some individual cases, it is important to note that our earlier results in Australians, plus results of other researchers testing in Britain, show that the CFMT-original picks up many cases of prosopagnosia despite less than ideal ethnic match of the Harvard stimuli to the local participants (e.g., Bate et al., 2008; Bowles et al., 2009; Palermo, Willis, et al., 2011; Rivolta et al., in press; Steede et al., 2007). This shows that there is no wholesale need for researchers in countries and regions other than Boston, USA, to throw out the CFMT-original and develop their own local versions. Extensive time and expense would be required to develop a version of the CFMT for every country or ethnic group of Caucasians.

We suggest that, in practice, testing an individual on both the CMFT-original and CFMT-Aus may resolve many ambiguous cases of prosopagnosia, given that the two tests between them cover a large ethnic range. Where the two tests disagree (beyond mere measurement error), we suggest the version with the better ethnic match should be preferred.

However, even with both tests available, there may remain some cases where a person reports everyday problems and is significantly impaired on Famous Faces, but appears normal on both the CFMT-original and the CFMT-Aus. We suggest that if neither of CFMT-original or CFMT-Aus are well matched ethnically to the participants' likely history of face exposure, then a diagnosis of normal face recognition should not be made without retest with better ethnic match stimuli.

Is there a long-delay-only form of prosopagnosia?

Our study has presented some suggestive evidence that a specific long-delay form of prosopagnosia might exist (also see Stollhoff et al., 2011, who tested a 1-year delay). Our present data set included two cases who were candidate individuals to test for long-delay prosopagnosia—namely, individuals for whom (a) deficits were apparent on Famous Faces and were reported in everyday life, but (b) performance was normal on short-delay face recognition tests (here, CFMT-original

and CFMT-Aus no delay). Of these, one showed a pattern potentially consistent with a long-delay form (Case 6 in Table 4). That is, her rank in the population gradually declined as the delay between face learning and test increased.

Given the small number of candidate cases we have been able to identify in our laboratories to date, it will be important for future studies to test more individuals. Should other labs wish to do this, the process can be assisted by use of the present norms for the 20-min and 24-hr relearn-and-delay conditions (if ethnic group of the potential prosopagnosic is suitable). If a participant shows a specific deficit in these conditions, independently manipulating delay and number of face repetitions would allow the researcher to determine whether the deficit lies with inability to retain faces in memory per se and/or with inability to benefit normally from relearning opportunities that are spaced over time.

Value of adding both 20-min and 24-hr delay conditions

The present study has provided preliminary evidence (from 1 participant) that adding relearn-and-delay conditions to a CFMT-style test can assist in diagnosis of prosopagnosia. Thus, there may be value in including a 20-min condition as a standard part of the testing regime. There was no evidence in the present study that adding the 24-hr condition added any additional information over and above the 20-min condition (i.e., the correlation between 20 min and 24 hr was at upper bound in control participants; plus our one relevant prosopagnosic participant showed equal population rank in these two phases).

Note that the 20-min and 24-hr conditions both produce non-normal distributions. This causes statistical inconvenience in calculating population ranks (see Experiment 1). However, it has a potential trade-off—namely, that the individuals at the low end of the distribution become separated more clearly from the rest of the group and from each other. This potentially allows more accurate rank ordering of deficit severity within prosopagnosics. This could, for example, give more power for correlating degree of prosopagnosia with behavioural or neural measures of interest (e.g., strength of holistic face processing; level of activation in the fusiform face area, Williams et al., 2007; integrity of white matter tracts in occipitotemporal cortex, Thomas et al., 2009, etc.).

Comparing prosopagnosics to controls tested with other session structures

It is common in prosopagnosia research to take a potential prosopagnosic's score from one session structure (e.g., Task A, Task B, target task, Task C) and to compare it to control norms for that task obtained from quite different session structures (e.g., target task, Task D, Task A). The present Experiment 5 supports the validity of this procedure, for CFMT-style tasks. We found no effects on control norms of earlier practice with similar-style tasks; that is, prior exposure to CFMT-Aus and CCMT did not boost CFMT-original performance. There was also no effect of fatigue due to late placement in the session, up to a session length of 1.25 hr.

Note that this does not rule out potential effects of fatigue with extremely long sessions. For example, Herzmann et al. (2008) tested the CFMT-original as Task Number 11 presented approximately 3 hr into a continuous 4-hr session (with 10-min breaks every hour). Although we had previously attributed the German participants' particularly poor performance on the CFMT-original to the lack of ethnic match (Bowles et al., 2009), fatigue may have explained some or all of the decrement.

Practical issues with using the CFMT-Australian

Validity and reliability

Our results demonstrated that the CFMT-Aus is both a valid and a reliable test of face recognition. Regarding validity, it shows a very large inversion effect (Experiment 2), a low correlation with an equivalent format test for objects (Cambridge Car Memory Test; $r = .21$, Experiment 1), and a much higher correlation with the original CFMT ($r = .62$). Regarding reliability, Cronbach's alpha of all three phases (no delay = .88; 20 min = .91, 24 hr = .93) is high enough that the test can be

used for prosopagnosia screening in individuals. High reliability is important because it reduces both false negatives and false positives.

Use of the inverted form

One question in studies of prosopagnosia has been whether prosopagnosics' poor recognition of upright faces occurs in conjunction with a lack of inversion effect for faces (e.g., Behrmann et al., 2005), which is taken as suggestive of a lack of holistic processing. Reliability of the inverted version of the CFMT-Aus (Experiment 2), at .68, is suitable for group comparisons addressing this question (i.e., a group of prosopagnosics compared to a group of controls). However, it is lower than would traditionally be desired for use with individual cases, where the aim would be to ask whether a single prosopagnosic has an atypically small inversion effect. However, note that in certain situations an individual's inversion score could still be interpretable. This is where conclusions remain clear once confidence intervals are taken into account: For example, an individual prosopagnosic could reliably be concluded to have an impaired inversion effect if the upper bound of the 95% CI on their inversion score did not fall in the normal range. To calculate the 95% CI, the reliability of the upright-minus-inverted difference score is required. This depends on the correlation between upright and inverted (for formula see, e.g., Kaplan & Saccuzzo, 2005), which we have not assessed here.

Applicability of our norms: Ethnicity; ageing; Glasgow Face Matching Test

We have presented control data for the CFMT-Aus in the present paper (Experiment 1). These control scores provide appropriate norms only for certain populations.

Regarding ethnicity, the control data in the present study were collected on an Australian population (including a small proportion of New Zealanders). We would expect control norms (Experiment 1) to apply to other countries to the extent that typical facial appearance of that country is similar to that of Australia (which is largely British heritage amongst Caucasians). Our values should be applicable in Australia, New Zealand, UK, and Ireland. We would also expect that they would apply reasonably well in Northern Europe and South Africa, although ideally this would need explicit testing. We would expect our norms to be less applicable to Southern Europe and Israel, and researchers wishing to use the test in those countries would be advised to collect their own control data.

Regarding age, the present norms were collected from young adults (age inclusion criterion 18–32 years). Age-related decline compared to 20-year-old performance begins at approximately 50 years of age on the CFMT-original (Bowles et al., 2009; Germine et al., 2010; also see Hildebrandt, Sommer, Herzmann, & Wilhelm, 2010), so the present norms are not valid beyond this age. Future studies will be needed to collect norms for older participants.

Finally, the Glasgow Face Matching Test (GFMT; Burton et al., 2010) uses some of the same faces as those that appear in the CFMT-Aus. The number of CFMT-Aus faces also appearing in the Glasgow Face Matching Test–Short Form is 4 (i.e., 4 of the 38 distractor faces that were taken from the Glasgow Unfamiliar Face Database), comprising 8% of the total 52 faces used in the CFMT-Aus. The number of CFMT-Aus faces also appearing in the Glasgow Face Matching Test–Long Form is 12, comprising 24% of the total 52 faces used in the CFMT-Aus (specific face code numbers are listed in Appendix B). Thus, a potential prosopagnosic's CFMT-Aus score may not be able to be compared to our norms if they have previously been tested on the GFMT.[7]

[7] This issue does not affect conclusions in the present paper. Of the 6 potential prosopagnosics we tested, only 1 (Case 2) had previously done the Glasgow Face Matching Test. She did the short form (only 4 overlapping faces out of 52), and this was 5 months before the CFMT-Aus. Also, she was impaired on the CFMT-Aus, while previous exposure to the items would be expected to, if anything, improve performance; thus, the interpretation of her results would have been problematic only if she were not impaired on the CFMT-Aus.

Use of CFMT-Aus in other settings

Beyond diagnosing prosopagnosia, the CFMT-Aus is also valuable in several other settings.

Experimental psychology

CFMT-style tests can offer potential statistical benefits in experimental psychology studies. Increasing internal reliability for individual subject's scores also decreases random error in the condition means, increasing power. The reliability of CFMT-style tests (typically around .88) is likely to be noticeably higher than that for most face recognition tasks traditionally used in experimental psychology. Of three face memory tests (delays of 4 min to 2.5 hours), Herzmann et al. (2008) reported that two had Cronbach's alpha of only .52 and .58. Even the best task (learn 16 faces, then old–new decision on 32 faces) achieved only .75 for reaction time and .69 for accuracy.

As an example illustrating the benefits of increased power arising from using CFMT-style tasks, we have recently used the CFMT and a Chinese-face version of the task (developed by Jia Liu) in an other-race effect study. The mean difference in accuracy for own-and other-race faces was similar to that in a previous study in our population of participants (approximately 10%), yet with similar numbers of participants the significance level was strikingly higher with the CFMT-format task ($N = 20$, $p < .001$, Stokes, McKone, Darke, & Aimola Davies, 2010) than in our previous study that used a more traditional-format recognition memory task ($N = 25$, $p = .021$, McKone, Brewer, MacPherson, Rhodes, & Hayward, 2007).

Individual-differences studies

The high internal reliability also makes CFMT-style tests valuable in studies utilizing individual differences in the normal range. Such studies have become increasingly popular as theoretical tools. For example, they have shown that face recognition ability is heritable, independent of general visual and cognitive abilities (twin studies, Wilmer, Germine, Chabris, et al., 2010; Zhu et al., 2010; see McKone & Palermo, 2010, for discussion). Like the original CFMT, the CFMT-Aus is ideal for individual differences studies because it displays, in addition to the desired high internal reliability, (a) a wide range of scores across individuals, (b) no ceiling or floor effects, and (c) a normal distribution.

Creating truly parallel forms: Use in prosopagnosia training studies

In arguing that there are important cognitive effects of the ethnicity difference between CFMT-original and CFMT-Aus faces, our results imply that the two tests are not truly parallel forms—that is, two forms where all aspects of the test are psychologically equivalent. We have argued that, for prosopagnosia diagnosis, this is a benefit rather than a negative, because researchers have the option of choosing the test that is best ethnically matched to the potential prosopagnosic.

However, in some circumstances, researchers may require access to two truly parallel versions of CFMT-structure tests. For example, there has been recent interest in whether face training can improve face recognition ability in prosopagnosia (De Gutis et al., 2007; Schmalzl, Palermo, Green, Brunsdon, & Coltheart, 2008) and autism spectrum disorder (Tanaka et al., 2010). Training studies can benefit from the availability of two closely matched tests of face recognition ability, one to assess ability pre training and one to assess ability post training. It is problematic to use the same test twice (e.g., CFMT-original both pre and post) in that there are practice effects arising from repeating the exact items: Wilmer, Germine, Chabris, et al. (2010) found improvement on readministration of the CFMT-original even with a mean delay of 6 months (mean score first occasion = 76.9%; mean score second occasion = 83.2%, $N = 389$). It is then difficult to disentangle the effects of item practice from the effects of the training programme. Further, adding a typically developing control group does not necessarily resolve this problem because it is theoretically plausible that the size of the practice effect could differ between the disorder group and the control group.

We suggest that, where required, two parallel CFMT-structure tests can be made by intermixing items across the CFMT-original and CFMT-Aus tests to equate average ethnicity. Appendix B provides readers with accuracy information by face to assist in this process. The aim would be to take, say, three target faces from the CFMT-Aus (with their corresponding distractors, i.e., the entire set of test triplets for that face) and three from the CFMT-original (with their corresponding distractors), such that ethnicity is equated across the two new tests, and mean accuracy is also equated.

Manuscript received 27 May 2011
Revised manuscript received 14 August 2011
Revised manuscript accepted 17 August 2011
First published online 25 November 2011

REFERENCES

Baron-Cohen, S. B., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, *31*(1), 5–17. doi:10.1023/A:1005653411471

Bate, S., Haslam, C., Tree, J., & Hodgson, T. L. (2008). Evidence of an eye movement-based memory effect in congenital prosopagnosia. *Cortex*, *44*, 806–819. doi:10.1016/j.cortex.2007.02.004

Behrmann, M., & Avidan, G. (2005). Congenital prosopagnosia: Face-blind from birth. *Trends in Cognitive Sciences*, *9*(4), 180–187. doi:10.1016/j.tics.2005.02.011

Behrmann, M., Avidan, G., Marotta, J. J., & Kimchi, R. (2005). Detailed exploration of face-related processing in congenital prosopagnosia: 1. Behavioral findings. *Journal of Cognitive Neuroscience*, *17*(7), 1130–1149. doi:10.1162/0898929054475154

Bentin, S., De Gutis, J. M., D'Esposito, M., & Robertson, L. C. (2007). Too many trees to see the forest: Performance, event-related potential, and functional magnetic resonance imaging manifestations of integrative congenital prosopagnosia. *Journal of Cognitive Neuroscience*, *19*, 132–146. doi:10.1162/jocn.2007.19.1.132

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

Bowles, D. C., McKone, E., Dawel, A., Duchaine, B., Palermo, R., Schmalzl, L., et al. (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.

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

Carbon, C.-C., Grüter, T., Grüter, M., Weber, J. E., & Lueschow, A. (2010). Dissociation of facial attractiveness and distinctiveness processing in congenital prosopagnosia. *Visual Cognition*, *18*(5), 641–654. doi:10.1080/13506280903462471

Chiroro, P., Tredoux, C., Radaelli, S., & Meissner, C. (2008). Recognizing faces across continents: The effect of within-race variations on the own-race bias in face recognition. *Psychonomic Bulletin & Review*, *15*(6), 1089–1092. doi:10.3758/pbr.15.6.1089

Cohen, J., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods*, *25*(2), 257–271. doi:10.3758/bf03204507

Crawford, J. R., & Garthwaite, P. H. (2002). Investigation of the single case in neuropsychology: Confidence limits on the abnormality of test scores and test score differences. *Neuropsychologia*, *40*(8), 1196–1208. doi:10.1016/S0028-3932(01)00224-X

Crawford, J. R., Garthwaite, P. H., & Slick, D. J. (2009). On percentile norms in neuropsychology: Proposed reporting standards and methods for quantifying the uncertainty over the percentile ranks of test scores. *The Clinical Neuropsychologist*, *23*(7), 1173–1195. doi:10.1080/13854040902795018

Crawford, J. R., & Howell, D. C. (1998). Comparing an individual's test score against norms derived from small samples. *The Clinical Neuropsychologist*, *12*(4), 482–486. doi:10.1076/clin.12.4.482.7241

de Gelder, B., Bachoud-Levi, A., & Degos, J. (1998). Inversion superiority in visual agnosia may be common to a variety of orientation polarised objects besides faces. *Vision Research*, *38*, 2855–2861.

De Gutis, J. M., Bentin, S., Robertson, L. C., & D'Esposito, M. (2007). Functional plasticity in ventral temporal cortex following cognitive

rehabilitation of a congenital prosopagnosic. *Journal of Cognitive Neuroscience*, *19,* 1790–1802.

Delis, D., Kramer, J., Kaplan, E., & Ober, B. (1987). *California Verbal Learning Test*. San Antonio, TX: The Psychological Corporation.

Dennett, H. W., McKone, E., Tavashmi, R., Hall, A., Pidcock, M., Edwards, M., et al. (2011). The Cambridge Car Memory Test: A task matched in format to the Cambridge Face Memory Test, with norms, reliability, sex differences, dissociations from face memory, and expertise effects. *Behavior Research Methods*.

Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General*, *115*(2), 107–117. doi:10.1037/0096-3445.115.2.107

Duchaine, B., Germine, L., & Nakayama, K. (2007). Family resemblance: Ten family members with prosopagnosia and within-class object agnosia. *Cognitive Neuropsychology*, *24*(4), 419–430. doi:10.1080/02643290701380491

Duchaine, B., Jenkins, R., Germine, L., & Calder, A. J. (2009). Normal gaze discrimination and adaptation in seven prosopagnosics. *Neuropsychologia*, *47*(10), 2029–2036. doi:10.1016/j.neuropsychologia.2009.03.011

Duchaine, B., Murray, H., Turner, M., White, S., & Garrido, L. (2010). Normal social cognition in developmental prosopagnosia. *Cognitive Neuropsychology*, *25,* 1–15.

Duchaine, B. C., & Nakayama, K. (2004). Developmental prosopagnosia and the Benton Facial Recognition Test. *Neurology*, *62*(7), 1219–1220. doi:10.1212/01.wnl.0000118297.03161.b3

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. doi:10.1016/j.neuropsychologia.2005.07.001

Duchaine, B. C., & Weidenfeld, A. (2003). An evaluation of two commonly used tests of unfamiliar face recognition. *Neuropsychologia*, *41*(6), 713–720. doi:10.1016/S0028-3932(02)00222-1

Duchaine, B., Yovel, G., & Nakayama, K. (2007). No global processing deficit in the Navon task in 14 developmental prosopagnosics. *Social Cognitive and Affective Neuroscience*, *2*(2), 104–113. doi:10.1093/scan/nsm003

Germine, L. T., Duchaine, B., & Nakayama, K. (2010). Where cognitive development and aging meet: Face learning ability peaks after age 30. *Cognition*, *118*(2), 201–210. doi:10.1016/j.cognition.2010.11.002

Gilchrist, A., & McKone, E. (2003). Early maturity of face processing in children: Local and relational distinctiveness effects in 7-year-olds. *Visual Cognition*, *10*(7), 769–793.

Herzmann, G., Danthiir, V., Schacht, A., Sommer, W., & Wilhelm, O. (2008). Toward a comprehensive test battery for face cognition: Assessment of the tasks. *Behavior Research Methods*, *40*(3), 840–857. doi:10.3758/brm.40.3.840

Hildebrandt, A., Sommer, W., Herzmann, G., & Wilhelm, O. (2010). Structural invariance and age-related performance differences in face cognition. *Psychology and Aging*, *25*(4), 794–810. doi:10.1037/a0019774

Iaria, G., Bogod, N., Fox, C. J., & Barton, J. J. S. (2009). Developmental topographical disorientation: Case one. *Neuropsychologia*, *47*(1), 30–40. doi:10.1016/j.neuropsychologia.2008.08.021

Kaplan, R. M., & Saccuzzo, D. P. (2005). *Psychological testing principles, applications and issues* (6th ed.). Belmont, CA: Thomson Wadsworth.

Kennerknecht, I., Grueter, T., Welling, B., Wentzek, S., Horst, J., Edwards, S., et al. (2006). First report of prevalence of non-syndromic hereditary prosopagnosia (HPA). *American Journal of Medical Genetics Part A*, *140A*(15), 1617–1622. doi:10.1002/ajmg.a.31343

Lee, Y., Duchaine, B., Wilson, H. R., & Nakayama, K. (2010). Three cases of developmental prosopagnosia from one family: Detailed neuropsychological and psychophysical investigation of face processing. *Cortex*, *46*(8), 949–964.

Le Grand, R., Cooper, P. A., Mondloch, C. J., Lewis, T. L., Sagiv, N., de Gelder, B., et al. (2006). What aspects of face processing are impaired in developmental prosopagnosia? *Brain and Cognition*, *61*(2), 139–158. doi:10.1016/j.bandc.2005.11.005

Lewin, C., & Herlitz, A. (2002). Sex differences in face recognition—Women's faces make the difference. *Brain and Cognition*, *50*(1), 121–128. doi:10.1016/S0278-2626(02)00016-7

Ley, P. (1972). *Quantitative aspects of psychological assessment*. London, UK: Duckworth.

McKelvie, S. J., Standing, L., St Jean, D., & Law, J. (1993). Gender differences in recognition memory for faces and cars: Evidence for the interest hypothesis. *Bulletin of the Psychonomic Society*, *31*(5), 447–448.

McKone, E. (2008). Configural processing and face viewpoint. *Journal of Experimental Psychology:*

Human Perception & Performance, *34*(2), 310–327. doi:10.1037/0096-1523.34.2.310

McKone, E. (2010). Face and object recognition: How do they differ? In V. Coltheart (Ed.), *Tutorials in visual cognition* (pp. 261–303). New York, NY: Psychology Press.

McKone, E., Brewer, J. L., MacPherson, S., Rhodes, G., & Hayward, W. G. (2007). Familiar other-race faces show normal holistic processing and are robust to perceptual stress. *Perception*, *36,* 224–248. doi:10.1068/p5499

McKone, E., & Palermo, R. (2010). A strong role for nature in face recognition. *Proceedings of the National Academy of Sciences of the USA*, *107*(11), 4795–4796. doi:10.1073/pnas.1000567107

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

O'Hearn, K., Schroer, E., Minshew, N., & Luna, B. (2010). Lack of developmental improvement on a face memory task during adolescence in autism. *Neuropsychologia*, *48*(13), 3955–3960. doi:10.1016/j.neuropsychologia.2010.08.024

Palermo, R., Rivolta, D., Wilson, C. E., & Jeffery, L. (2011a, May). *Abnormal adaptive coding of identity in congenital prosopagnosia*, Paper presented at Vision Sciences Society meeting, Naples, FL, USA.

Palermo, R., Rivolta, D, Wilson, C. E., & Jeffery, L. (2011b). *Adaptive face space coding in congenital prosopagnosia: Typical figural aftereffects but abnormal identity aftereffects*, Manuscript submitted for publication.

Palermo, R., Willis, M. L., Rivolta, D., McKone, E., Wilson, C. E., & Calder, A. J. (2011). Impaired holistic coding of facial expression and facial identity in congenital prosopagnosia. *Neuropsychologia*, *49*(5), 1226–1235. doi:10.1016/j.neuropsychologia.2011.02.021

Pitcher, D., Walsh, V., Yovel, G., & Duchaine, B. (2007). TMS evidence for the involvement of the right occipital face area in early face processing. *Current Biology*, *17,* 1568–1573.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. Oxford, UK: Oxford Psychologists Press.

Reed, C., Stone, V. E., Bozova, S., & Tanaka, J. W. (2003). The body-inversion effect. *Psychological Science*, *14*(4), 302–308.

Rivolta, D., Palermo, R., Schmalzl, L., & Coltheart, M. (in press). Covert face recognition in congenital prosopagnosia: A group study. *Cortex*, Advance online publication. doi:10.1016/j.cortex.2011.01.005

Rivolta, D., Schmalzl, L., Coltheart, M., & Palermo, R. (2010). Semantic information can facilitate covert face recognition in congenital prosopagnosia. *Neuropsychology, Development, and Cognition. Section A, Journal of Clinical and Experimental Neuropsychology*, *32*(9), 1002–1016.

Robbins, R., & McKone, E. (2007). No face-like processing for objects-of-expertise in three behavioural tasks. *Cognition*, *103*(1), 34–79. doi:10.1016/j.cognition.2006.02.008

Rossion, B. (2009). Distinguishing the cause and consequence of face inversion: The perceptual field hypothesis. *Acta Psychologica*, *132*(3), 300–312. doi:10.1016/j.actpsy.2009.08.002

Scapinello, K. F., & Yarmey, A. D. (1970). The role of familiarity and orientation in immediate and delayed recognition of pictorial stimuli. *Psychonomic Science*, *21,* 329–331.

Schmalzl, L., Palermo, R., & Coltheart, M. (2008). Cognitive heterogeneity in genetically based prosopagnosia: A family study. *Journal of Neuropsychology*, *2*(1), 99–117. doi:10.1348/174866407x256554

Schmalzl, L., Palermo, R., Green, M., Brunsdon, R., & Coltheart, M. (2008). Training of familiar face recognition and visual scan paths for faces in a child with congenital prosopagnosia. *Cognitive Neuropsychology*, *25,* 704–729.

Sporer, S. L., & Horry, R. (2010). Recognizing faces from ethnic in-groups and out-groups: Importance of outer face features and effects of retention interval. *Applied Cognitive Psychology*, *25*(3), 424–431.

Sporer, S. L., Trinkl, B., & Guberova, E. (2007). Matching faces. *Journal of Cross-Cultural Psychology*, *38*(4), 398–412. doi:10.1177/0022022107302310

Steede, L. L., Tree, J. J., & Hole, G. J. (2007). I can't recognize your face but I can recognize its movement. *Cognitive Neuropsychology*, *24*(4), 451–466. doi:10.1080/02643290701381879

Stokes, S., McKone, E., Darke, H., & Aimola Davies, A. (2010). Poor memory for other race faces is not associated with deficiencies in holistic processing [Abstract]. *Journal of Vision*, *10*(7), 696, doi:10.1167/10.7.696

Stollhoff, R., Jost, J., Elze, T., & Kennerknecht, I. (2011). Deficits in long-term recognition memory reveal dissociated subtypes in congenital prosopagnosia. *PLoS ONE*, *6*(1), e15702, doi:10.1371/journal.pone.0015702

Susilo, T., McKone, E., Dennett, H., Darke, H., Palermo, R., Hall, A., et al. (2011). Face recognition impairments despite normal holistic processing and

face space coding: Evidence from a case of developmental prosopagnosia. *Cognitive Neuropsychology*, *27*, 636–664.

Tanaka, J. W., Wolf, J. M., Klaiman, C., Koenig, K., Cockburn, J., Herlihy, L., et al. (2010). Using computerized games to teach face recognition skills to children with autism spectrum disorder: The Let's Face It! program. *Journal of Child Psychology and Psychiatry*, *51*(8), 944–952. doi:10.1111/j.1469-7610.2010.02258.x

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*, 29–31.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology A*, *43*(2), 161–204.

Warrington, E. K. (1984). *Recognition Memory Test*. Windsor, UK: NFER-Nelson.

Williams, M. A., Berberovic, N., & Mattingley, J. B. (2007). Abnormal fMRI adaptation to unfamiliar faces in a case of developmental prosopamnesia. *Current Biology*, *17*(14), 1259–1264. doi:10.1016/j.cub.2007.06.042

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., et al. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences*, *107*(11), 5238–5241. doi:10.1073/pnas.0913053107

Wilmer, J. B., Germine, L., Loken, E., Guo, X. M., Chatterjee, G., Nakayam, K., et al. (2010). Response to Thomas: Is human face recognition ability entirely genetic? *Proceedings of the National Academy of Sciences*, *107*(24), E101.

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. doi:10.1016/j.jpsychores.2008.03.013

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*(1), 141–145. doi:10.1037/h0027474

Yovel, G., & Duchaine, B. (2006). Specialized face perception mechanisms extract both part and spacing information: Evidence from developmental prosopagnosia. *Journal of Cognitive Neuroscience*, *18*(4), 580–593. doi:10.1162/jocn.2006.18.4.580

Zhu, Q., Song, Y., Hu, S., Li, X., Tian, M., Zhen, Z., et al. (2010). Heritability of the specific cognitive ability of face perception. *Current Biology*, *20*(2), 137–142. doi:10.1016/j.cub.2009.11.067

APPENDIX A

CFMT-Aus availability and conditions of use

Copies of the CFMT-Aus are available from the first author, including stimuli (in either .pct or .jpg), PsyScope X script files, a list of item order to allow users of other experiment-running software to set up their own scripts, and scoring sheets in Excel format. We hope to produce a .jar version in future.

Users with other screen sizes and/or resolutions from ours will need to adjust viewing distance if they wish to maintain the same visual angle of stimuli as that used here (this is because the faces appear as different physical sizes on different screens).

It is a condition of use of the test that researchers do not make it, or any of its items, publically available on the web or in publications. The purpose of this limitation is to ensure that the test remains a test of *novel* face learning to potential participants.

APPENDIX B

Further details on the use of the CFMT-Aus

1. Further details on construction of the CFMT-Aus test trials

Learn phase stimuli. Within each viewpoint, a common shape of cropping template was applied to all targets and distractors. For the 3AFC test trials, image trios were determined such that distractor individuals were reasonably similar in appearance to the target; viewpoints of the photographs were also closely matched. A given distractor was used for a maximum of two target faces in any one viewpoint. Once these trios were determined, the brightness, contrast, and lighting of the distractor images within each trio were manipulated to match the

Table B1. *Frequency distribution for CFMT-Aus 20-min and 24-hr delay*

	Condition			
	20-min		*24-hr*	
Score	k	m	k	m
21	0	0	1	0
22	0	0	0	1
23	1	0	0	1
24	0	1	0	1
25	0	1	1	1
26	0	1	0	2
27	2	1	1	2
28	1	3	1	3
29	0	4	0	4
30	2	4	0	4
31	1	6	0	4
32	0	7	1	4
33	0	7	0	5
34	2	7	0	5
35	3	9	1	5
36	0	12	0	6
37	2	12	1	6
38	1	14	2	7
39	2	15	2	9
40	1	17	2	11
41	2	18	3	13
42	5	20	3	16
43	1	25	2	19
44	3	26	0	21
45	3	29	5	21
46	7	32	4	26
47	8	39	3	30
48	3	47	3	33
49	4	50	5	36
50	6	54	6	41
51	2	60	9	47
52	8	62	10	56
53	4	70	3	66
54	1	74	6	69

Note: CFMT = Cambridge Face Memory Test. CFMT-Aus = CFMT-Australian. $N = 75$. k = number of controls obtaining that score. m = number of controls obtaining less than that score.

target, to ensure the target image could not be identified on the basis of lighting differences from the distractors.

Novel phase stimuli. Viewpoints were one-third-profile right, two-thirds-profile left, two-thirds-profile right, and 2 images facing front. For front views, the starting stimuli of the 6 targets were the original photographs used in the Learn phase, and, from these, two or three changes were made: (a) A different external template shape was always used (common to all items in the trio); (b) lighting conditions were always altered to mimic the CFMT;[8] and (c) for one set of the front-view faces (those given bottom lighting), the brightness and contrast were also changed. For one-third-profile right views (i.e., approximately 30-degree rotation from front view), the starting stimuli were again the original photographs used in the Learn phase: Template shape and lighting changes were applied, with new lighting directed from the right for the one-third-profile right images, from the left for the first set of front-facing images, and from the bottom for the second set of front-facing images, causing the top part of the face to become darker. For the two-thirds-profile images (i.e., approximately 60-degree rotation from front view), new photographs of the targets not used in the Learn phase were available: Therefore, the images were only given a template, and lighting changes were not made because they were not made for these viewpoints in the original CFMT. Pairing of distractor faces to particular targets was based partly on availability of photographs meeting the above criteria and partly on pilot testing to set task difficulty; choice of exact template shape and lighting changes was also determined by pilot testing.

Novel-in-Noise phase stimuli. Viewpoints in this stage were one-third-profile left, two-thirds-profile right, and the front-facing image. Manipulations were made to lighting so the images matched, as best as possible, this stage of the CFMT. Lighting was directed from the right for one set of the front-facing images and the one-third-profile-left images, and from the left for the two-thirds-profile-right images. Brightness and contrast were manipulated for the second set of front-facing images, making them appear lightly shadowed. New templates were given to both sets of front-facing images and the one-third-profile-left images.

2. Data analysis: Rationale for inclusion or exclusion of individuals in CFMT-Aus control means

As noted in Participants, of the 77 subjects tested from a population unselected for face recognition ability, we excluded

[8] The original CFMT used photographs taken from a database in which faces were physically lit from different angles (e.g., left, right, and bottom, in addition to the more normal top/front lit); the databases we used did not contain such images, but a reasonable approximation to the CFMT lighting conditions was possible and was performed using the FILTER > RENDER > LIGHTING EFFECTS function in Photoshop.

2 from our control means as outliers. These cases had very poor performance and are marked in blue and green in the frequency distributions in Figure 2[9]. Our rationale for excluding case MB13 was that, across all CFMT-Aus conditions, he performed very nearly as poorly as a diagnosed prosopagnosic (E.D., see red bar in Figure 2; note that E.D. is Case 1 in Experiment 4). Indeed, although 100% accurate at same-image items, MB13 showed no ability to discriminate new-image items at all: He was slightly below chance for no delay subtotal for Novel + Novel-in-Noise (see Figure 2B). MB13's z score on the no delay total was −3.10. We also excluded case AM80. After MB13, she was the next poorest performer on the no delay total, with a z score of −2.55. Although she was not much poorer than the next poorest subject at this delay, AM80 was the only participant in the entire sample to get consistently worse across the three delay conditions. In the 20-min and 24-hr conditions, her performance is clearly not part of the control distribution (Figures 2C and 2D), and she performed as poorly as diagnosed prosopagnosic E.D.

Finding 2 individuals out of 77 as having scores poor enough to imply possible prosopagnosia is not unexpected. It is consistent with previous results arguing that approximately 2.0−2.9% of the population have developmental prosopagnosia (Bowles et al., 2009; Kennerknecht et al., 2006).

We also considered 4 other low performers for possible deletion (see individuals marked *, +, x, o in Figure 2), but decided there was no convincing reason to exclude these. They do not fall clearly outside the control distribution, and their ordering was not consistent across the various delay conditions: * was the poorest at no delay, yet similar to or better than +, o, and/or x on the 20-min and 24-hr conditions. (We also checked there was no evidence that * was prosopagnosic on other tests: His score on the original CFMT was 47, i.e., a z score of −0.98, well within the normal range.)

3. Frequency distribution tables for non-normally distributed phases of CFMT-Aus

The CFMT-Aus 20-min and 24-hr delay conditions have non-normal (skewed) control data. An individual case's percentile rank can be calculated using the information in Table A1, using the formula (Crawford, Garthwaite & Slick, 2009):

$$\text{Percentile rank} = [(m + 0.5k)/N] \times 100$$

where m = number of members of the normative sample scoring below a given score

k = number of members of the normative sample obtaining the given score

N = total control sample size.

For example, the percentile rank of a case who scores 27 in the 24-hr condition is 3.33 (i.e., bottom 3.33% of population), calculated as: $\{[2 + (0.5 \times 1)]/75\} \times 100$. To calculate 95% CI on this percentile rank, use the software provided by Crawford, Garthwaite, and Slick (2009), entering the frequency distribution table (i.e., score and k) for the desired relearn-and-delay condition.

4. Creating true parallel forms of CFMT: Filename codes for accuracy by face

If researchers wish to create two ethnically equated versions of the CFMT by intermixing items from the CFMT-original and the CFMT-Aus, the aim would usually be to equate average accuracy (i.e., not put the 6 hardest target faces in one version and 6 easiest in the other). To assist with this, the face code numbers in square brackets corresponding to each target in the by-face analysis (Experiment 1) are: 88% correct [004], 87% [049], 82% [011], 77% [022], 74% [040], 74% [076]. These codes are as they appear in the CFMT-Aus stimulus filenames and in the Australian National University face database. Regarding the CFMT-original, the by-face accuracy scores in Duchaine and Nakayama (2006, Section 2.1.3; listed there as 77, 69, 80, 81, 88, and 88) are presented in the same order as the target faces appear in the Learn phase when the test is run (B. Duchaine, personal communication, March 31st, 2011); hence, 77% was for the first face, 69% was the second, and so on.

5. Filename codes for faces that overlap with Glasgow Face Matching Test

The Glasgow face stimuli that appear in the CFMT-Aus and also appear in the GFMT-short form are: 049, 076, 091, 092 (using code numbers from the Glasgow Unfamiliar Face Database). The Glasgow face stimuli that appear in the CFMT-Aus and also appear in the GFMT-long form are: 012, 049, 051, 076, 091, 092, 114, 125, 126, 239, 246, 270.

[9] All references to Figure 2 in the Appendix refer to Figure 2 in the text.