Behavior Research Methods
https://doi.org/10.3758/s13428-022-01805-8

Introducing the female Cambridge face memory test – long form (F-CFMT+)

Myles Arrington[1] · Daniel Elbich[1] · Junqiang Dai[1] · Bradley Duchaine[2] · K. Suzanne Scherf[1]

Accepted: 24 January 2022

Abstract

The Cambridge Face Memory Test (CFMT) is one of the most used assessments of face recognition abilities in the science of face processing. The original task, using White male faces, has been empirically evaluated for psychometric properties (Duchaine & Nakayama, 2006), while the longer and more difficult version (CFMT+; Russell et al., 2009) has not. Critically, no version exists using female faces. Here, we present the Female Cambridge Face Memory Test – Long Form (F-CFMT+) and evaluate the psychometric properties of this task in comparison to the Male Cambridge Face Memory Test – Long Form (M-CFMT+). We tested typically developing emerging adults (18 to 25 years old) in both Cambridge face recognition tasks, an old-new face recognition task, and a car recognition task. Results indicate that the F-CFMT+ is a valid, internally consistent measure of unfamiliar face recognition that can be used alone or in tandem with the M-CFMT+ to assess recognition abilities for young adult White faces. When used together, performance on the F-CFMT+ and M-CFMT+ can be directly compared, adding to the ability to understand face recognition abilities for different kinds of faces. The two tasks have high convergent validity and relatively good divergent validity with car recognition in the same task paradigm. The F-CFMT+ will be useful to researchers interested in evaluating a broad range of questions about face recognition abilities in both typically developing individuals and those with atypical social information processing abilities.

Keywords Face recognition · Recognition memory · Visual processing · Face processing · Face memory · Neuropsychological test

Identifying people as unique individuals is critical for social interaction. While there are multiple sensory cues that facilitate person identification, visual information from the *face* is disproportionately useful for recognizing individual people. The Cambridge Face Memory Test (CFMT) is a commonly used assessment to measure face recognition abilities, particularly for unfamiliar faces (Duchaine & Nakayama, 2006; Russell et al., 2009). Duchaine and Nakayama's original work has been cited over 800 times since its publication in 2006 and has been used to assess multiple aspects of face recognition, including poor face identity recognition skills among those with developmental prosopagnosia (Bowles et al., 2009) and autism (O'Hearn et al., 2010; Whyte et al., 2016), the heritability of face recognition (Shakeshaft & Plomin, 2015; Wilmer et al., 2010), individual differences in face recognition abilities (Elbich & Scherf, 2017; Russell et al., 2009; Susilo et al., 2013), and the development of face recognition skills (Germine et al., 2011).

The CFMT measures unfamiliar face identity memory for White American male faces[1]. Therefore, performance is dependent on both visuoperceptual mechanisms and recognition memory. It was designed to mimic aspects of everyday face recognition experiences. The authors reported that they chose male faces exclusively to avoid potential gender differences in responses to female faces (i.e., own-gender bias; see Herlitz & Lovén, 2013). However, without a similarly standardized task for evaluating recognition of female faces, questions about gender differences in face recognition

✉ K. Suzanne Scherf
suzyscherf@psu.edu

[1] Department of Psychology, Pennsylvania State University, 113 Moore Building, University Park, PA 16802, USA

[2] Department of Brain and Psychological Sciences, Dartmouth University, Hanover, NH, USA

[1] Subsequently, researchers have developed version of the CFMT to test recognition of male faces from other ethnicities and nationalities, including the CFMT- Chinese (McKone et al., 2017) and the CFMT-Australian (McKone et al., 2011).

Published online: 22 February 2022

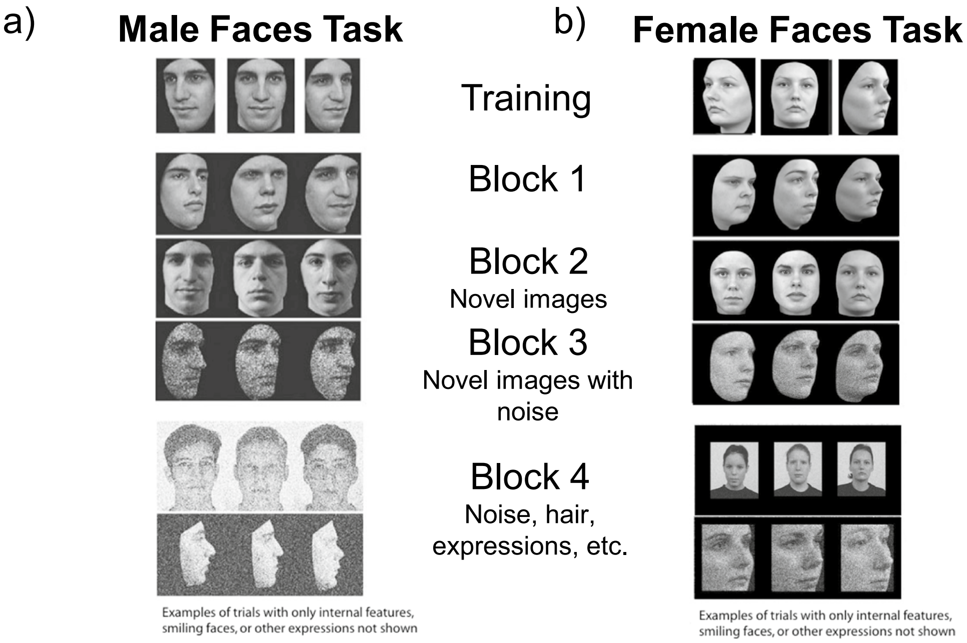

a)
Male Faces Task
b)
Female Faces Task
Training
Block 1
Block 2
Novel images
Block 3
Novel images with noise
Block 4
Noise, hair, expressions, etc.
Examples of trials with only internal features, smiling faces, or other expressions not shown
Examples of trials with only internal features, smiling faces, or other expressions not shown

Fig. 1 Male and female face recognition tasks. Task outlines of the (**a**) male (figure adapted from Russell et al., 2009) and (**b**) female (created for Scherf et al., 2017) versions of the CFMT+ (images of female faces are published with permission from the Radboud (Langner et al., 2010) and Karolinska (Lundqvist et al., 1998) faces databases and include images AF16NES, AF19NES, and AF29NES). In these tasks, participants view target identities from multiple viewpoints and then must recognize the target faces among distractors with increasing levels of difficulty across blocks, which add noise via changes in lighting and viewpoint (Block 2), visual noise (Block 3), hair, affect, and repeating distractors (Block 4)

and the own-gender bias cannot be evaluated using the current version of the CFMT alone. Here, we formally present the Female Cambridge Face Memory Test and evaluate the psychometric properties of this task in comparison to the CFMT with male faces in the same participants.

The Cambridge face memory test structure and properties

The original version of the CFMT (Duchaine & Nakayama, 2006) is a three-alternative forced choice paradigm in which participants study six unfamiliar target faces without hair or clothing making neutral facial expressions in each of three viewpoints (see Fig. 1a). During recognition trials, participants identify the target faces under conditions of increasing difficulty, including lighting changes and visual noise. The authors reported the psychometric properties of this task in both college-aged control participants and in individuals with face recognition impairments. They found internal consistency of performance across the recognition blocks, consistency in item level sensitivity, and external validity with other measures of face recognition ability. Subsequently, a long form of the task was created (CFMT+) to evaluate individual differences in superior face recognition abilities by adding a fourth block of trials in which hair and changing expressions are introduced on the target faces and in which a smaller set of distractor identities repeat more frequently (Russell et al., 2009). The psychometric properties of this long form of the task have not been empirically reported, which is critical for evaluating how well the task assesses individual differences in face recognition abilities.

Developing the female Cambridge face memory test

Here, we present a new version of the CFMT+ using only White female faces (F-CFMT+). This task parallels all parameters of the male task (M-CFMT+), including the editing and presentation of the face images, stimulus timing, and response protocol. As in the M-CFMT+, the photographs were high-resolution images of individual women in the age range of 20–30 years. In contrast to the development of the M-CFMT+, we did not photograph actors ourselves, opting instead to use face images from databases of women from European and North American

Table 1 Demographic characteristics of sample

	Total sample	Men	Women
N	126	59	67
Age (years)	19.56	19.56	19.57
Proportion White	80.95%	86.44%	76.12%

countries. Initial findings with a previous version of the F-CFMT+ were presented elsewhere (Scherf et al., 2017).

Current study

Our primary goal was to evaluate the psychometric properties of the new F-CFMT+ in comparison to those of the M-CFMT+ in the same participants. In doing so, we determined the psychometric properties of the M-CFMT+, given that they have not been reported to date. To evaluate the external validity of both Cambridge face tests, we assessed correlations in performance with another test of unfamiliar face recognition using an old-new recognition memory paradigm (see Picci & Scherf, 2016). We were also interested in evaluating the extent to which these measures assess unfamiliar face recognition and not just visual recognition memory more generally. Therefore, we included the Car Cambridge Memory Test (CCMT; Dennett et al., 2012) for comparison. We compared the internal consistency of each Cambridge face test, both on the block level and the item level, and the external validity across all three tests, to evaluate the specificity of these properties.

Methods

Participants

The final sample included 126 typically developing emerging adults (range 18–25 years). The demographic characteristics of the participants are given in Table 1 as a function of participant gender. The ethnic distribution of the participants reflected the ethnic distribution of the university from which we recruited them. Our prior work with a pilot version of the F-CFMT+ indicated that overall performance on the F-CFMT+ and M-CFMT+ would be moderately correlated ($r = 0.61$), but that a large effect size ($d = 3.34$) may characterize a difference in overall task performance (Scherf et al., 2017). As a result, an a priori power analysis indicated that a total same size of 75 would give us 95% power to detect a more moderately sized within-subjects fixed effect of task at a significance level of .01. Therefore, we sought to double this estimate so that we could safely evaluate the effect separately in male and female participants. However, the COVID-19 pandemic led to an early termination of data collection in March 2020. Importantly, our final sample size of 126 still exceeds (and nearly doubles) the minimum sample size needed to evaluate the question of task-related differences using a within-subjects design. A subset of these participants ($N = 112$ participants, 57 women; $M_{\mathrm{Age}=}$ 19.59, $SD = 1.64$) also competed an old-new face recognition memory task.

Participants were healthy and had no history of neurological (e.g., Parkinson's disease, Huntington's disease) or psychiatric disorders (e.g., mood, body dysmorphic, anxiety, and developmental disorders) in themselves or their first-degree relatives. They were also screened for autism-like symptoms using the Autism Quotient (AQ: Baron-Cohen et al., 2001) and behavioral symptoms indicative of undiagnosed psychopathology. Each of these conditions are reportedly associated with atypical face processing behavior.

Written informed consent was obtained using procedures approved by the Internal Review Board of Pennsylvania State University. Participants were recruited through the Psychology Department undergraduate subject pool and via fliers on campus.

Behavioral measures

Male Cambridge face memory test – long form (M-CFMT+)

The M-CFMT+ is a test of unfamiliar face recognition of White male faces (Duchaine & Nakayama, 2006; Russell et al., 2009). We used the long form, which has been shown to capture a broad range of unfamiliar face recognition abilities (Elbich & Scherf, 2017; Russell et al., 2009). The task is divided into four blocks that increase in difficulty (Fig. 1a). Across the 18 items in Block 1, participants study six target faces with no hair and neutral expressions in each of three viewpoints. Each of the target faces is first studied and then immediately tested to be sure the participant can identify that specific target. During recognition trials, participants identify the target faces by button-press in a three-alternative forced choice paradigm. The 46 distractor faces are novel unfamiliar faces. At the conclusion of Block 1, participants are presented with frontal images of the six target faces simultaneously and study them for 20 seconds. For the 30 items in Block 2, images of the target faces are novel (i.e., altered lighting, pose, or both) and must be identified in contrast to the distractor faces. Participants again review the frontal images of the target faces at the end of Block 2. For the 24 items in Block 3, Gaussian noise is added to the novel images of the target faces and the distractor faces. Finally, across 30 items in Block 4, additional novel images of the target and distractor faces are introduced that include hair and facial expressions and a smaller set of distractor

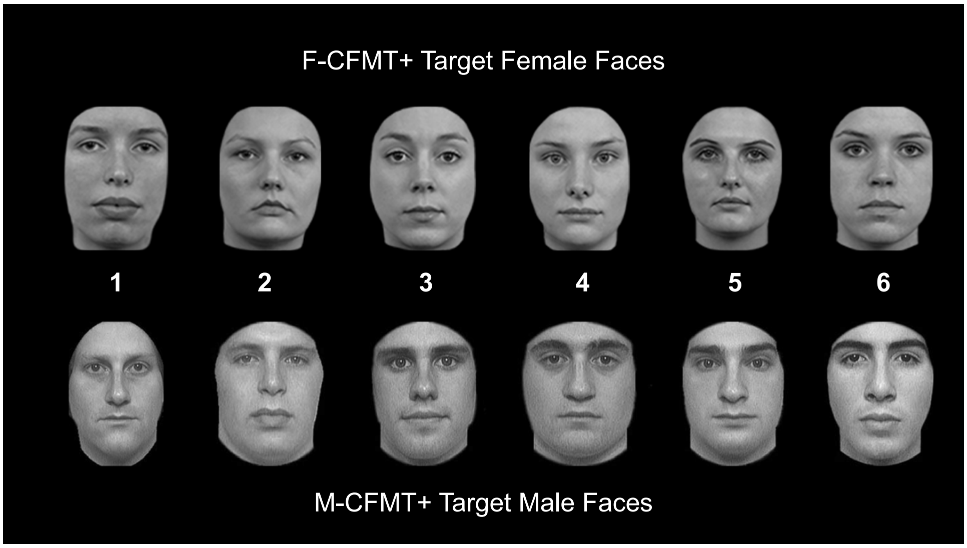


Fig. 2 Target face stimuli in each of the face recognition tasks. The target face stimuli selected for the F-CFMT+ included high-resolution images of individual White women in the age range of 20–30 years. The images were grey-scaled and cropped to exclude the background. The hair, skin blemishes, and skin discoloration (e.g., freckles) were removed from the faces. The range of perceptual similarity among the six target female faces was selected to be comparable to the range of perceptual similarity among the six target male faces in the M-CFMT+

faces occur more frequently than in previous blocks, making identification of the target faces very difficult. The 102 trials across the four blocks are presented with a fixed order for all participants.

Female Cambridge face memory test – long form (F-CFMT+)

Like other versions of the CFMT (McKone et al., 2011, 2012), the F-CFMT+ is designed to match all the parameters of the M-CFMT+ at both the block (see Fig. 1b) and trial levels. The stimuli include high-resolution images of individual White women in the age range of 20–30 years. They were selected from the Radboud (Langner et al., 2010) and Karolinska (Lundqvist et al., 1998) faces databases. We selected images of multiple face identities from different viewpoints (frontal, 45°, 135°) to match the training trials in the M-CFMT+, and with different expressions to match the manipulations in the fourth block of the M-CFMT+. The images were converted to grayscale and cropped to exclude the background. The hair and all skin blemishes and discoloration (e.g., freckles) were removed from the faces. The six target White female faces were selected to exhibit a similar range of perceptual variability as exhibited among the six target White male faces in the M-CFMT+ (see Fig. 2). As in the original CFMT and the M-CFMT+, the target identities appear on 1/6 trials throughout all four blocks of the task. Each target appears approximately 17 times throughout the task (range 16–18); the variability in rate of presentation is a function of the relative perceptual similarity between the target and the distractor identities.

To create the task, each female target face (e.g., target face 1, target face 3) was substituted for the matching target male face (e.g., target face 1, target face 3) at the trial level throughout each block of the task. To match the M-CFMT+ task, 46 distractor identities were required with frontal and profile images. However, because of limitations with the stimuli databases that were available at the time of task development, only 38 distractor identities met these criteria. On average, each face appeared 5.37 times throughout the task. Therefore, the frequency of the distractor faces in the F-CFMT+ was a bit higher (1/19 trials) than in the M-CFMT+ (1/23 trials) in the first three blocks.

To make the noise-distorted images for Blocks 3 and 4, we applied a 30% level of Gaussian noise to the images, as described in the creation of the M-CFMT (Duchaine & Nakayama, 2006). However, given the high resolution of the female face images, we had to apply multiple iterations of this Gaussian noise to degrade the images to a similar level as in Blocks 3 and 4 of the M-CFMT+. In the final block 4, as in the M-CFMT+, we selected 14 of the most frequently occurring distractor faces to repeat more frequently and increase the difficulty of this block by reducing "the difference in familiarity between target and distractor faces" (Russell et al., 2009).

Pilot testing on emerging adults at Penn State University was used to adjust the difficulty level of the task via selection of target and distractor items. Preliminary data with an earlier version of the task were published previously (Scherf et al., 2017). Results from this study were used to generate the power analysis and to refine the task for this round of testing and release.

In this final version of the F-CFMT+, the instructions, task structure, and task timing are exactly the same as the M-CFMT+. The task is available for experimental purposes on Testable.org (tstbl.co/174-523) and via download on Databrary (Scherf, 2021; https://doi.org/10.17910/b7.1396**).**

Car Cambridge memory test (CCMT)

The CCMT tests recognition memory for cars in the same task structure as the original CFMT (Dennett et al., 2012). Participants study six target cars and subsequently identify them in a three-alternative forced-choice paradigm under conditions of increasing difficulty (i.e., changes in viewpoint, lighting, and noise levels). In so doing, the CCMT functions as a comparable task of visual object recognition that is matched for general cognitive requirements (memory, processing speed). There are 72 trials across three blocks in the task.

Old-new face identity recognition memory task

To determine the convergent validity of the two long-form Cambridge memory face tasks to measure unfamiliar face recognition abilities, we evaluated the degree of association between performance in each Cambridge task to recognition performance in an old-new face recognition memory task (see Picci & Scherf, 2016). The stimuli consisted of grayscale photographs of faces with neutral and happy expressions obtained from multiple face databases: NimStim (Tottenham et al., 2009), Karolinska (Lundqvist et al., 1998), and the Radboud Face Database (RaFD; Langner et al., 2010). In the encoding phase, participants were shown 10 target faces with neutral expressions for three seconds each. The encoding phase was immediately followed by a 90-second delay period. In the test phase, participants were shown 20 individual images of the target (10) and distractor (10) faces, which were all smiling. Participants were instructed to indicate by button-press whether they recognized each face. The original task includes five recognition blocks in which the stimuli varied as a function of face age (child, early adolescent, late adolescent, emerging adult, parent). For the purposes of these analyses, we only included assessed performance from the block in which participants recognized emerging adult faces, which is overlapping with the age of the faces in the Cambridge face tests. Performance from the other blocks in the task will be reported in a subsequent manuscript. We computed hit rate, false alarm rate, and d-prime scores as dependent variables from this task.

Procedure

Participants first completed an online screening form to assess initial eligibility and the likelihood that they would meet the inclusion and exclusion criteria for the study. Eligible participants were invited to the lab to conduct a full assessment of eligibility. The experimental design was within-subjects. Eligible participants completed all three of the Cambridge recognition tasks (F-CFMT+, MCFMT+, CCMT), the order of which was counterbalanced across male and female participants separately. In addition, all participants also completed the old-new face recognition task; the order of which was also counterbalanced with the set of Cambridge across participants.

Data analysis

All analyses were conducted using R software using the R Studio interface (R Core Team, 2019; RStudio Team, 2020). R packages included tidyverse (Wickham et al., 2019), lme4 (Bates et al., 2015), lmerTest (Kuznetsova et al., 2017), psych (Revelle, 2020), rstatix (Kassambara, 2021), lavaan (Rosseel, 2012), and semTools (Jorgensen et al., 2021). Accuracy (i.e., percent correct items) was the primary dependent measure from the Cambridge tasks. The distributions of accuracy scores on all three recognition memory tests exhibited normal skew [−1, 1] and kurtosis [−3, 3]; therefore, none of the data were transformed.

In evaluating the psychometric properties of the M-CFMT+ and the F-CFMT+, we report many of the same metrics described in the original publication of the M-CFMT (Duchaine & Nakayama, 2006). First, we examined average total accuracy across all items in each task and separately within each block of each task. Second, we evaluated accuracy for each target face. To compare the tasks in these accuracy metrics we employed separate two-tailed paired-samples t-tests with a Bonferroni correction of 5 (0.05/5, $p=0.01$).

We measured the reliability of each of the tasks using *coefficient omega* (ω_{cat}), which assesses how reliably a total score for a test measures a single construct that is common to all items in the test (see Flora, 2020). We measured internal consistency at the block and item levels separately for each task. For the block level, we computed Pearson product moment correlations between task blocks (i.e., M-CFMT+ Block 1 correlated to M-CFMT+ Block 2). To measure the item-level internal consistency, we computed point biserial correlations comparing each participant's overall accuracy with their performance on each trial-level item (i.e., M-CFMT+ overall accuracy correlated to performance on M-CFMT+ Trial 1). We compared the size of point biserial correlations between the tasks by using a Fischer's z-score transformation on the correlations.

We also measured external and divergent validity for the Cambridge face tasks. External validity was measured by computing Pearson correlation coefficients between the two face tasks, both by comparing total accuracy in each task (i.e., M-CFMT+ total accuracy and F-CFMT+ total accuracy) and accuracy in each block (i.e., M-CFMT+ Block 1 and F-CFMT+ Block 1). We further evaluated external validity by comparing performance in each of the Cambridge face tasks to that in the old-new face recognition paradigm. Because prior work indicates that performance in the CFMT predicts false alarms, but not hits, in tasks of old-new recognition memory (see Devue et al., 2019), we measured how well performance in each block of the two Cambridge tasks predicted total performance (i.e., d-prime), hits, and false alarms separately. Divergent validity was measured using a similar approach with the car task.

Results

Prior to analyzing the data for task-related effects, we estimated potential order effects. Among the three tasks, there were six possible task orders. We created five contrast codes for the six orders. The regression model, with each contrast code as a predictor, revealed no main effect of task order on performance for either of the face recognition tasks, M-CFMT+: $F(5, 120) = 0.67$, $p = 0.646$, $\eta_p^2 = 0.028$; F-CFMT+: $F(5, 120) = 0.62$, $p = 0.682$, $\eta_p^2 = 0.026$. There was a main effect of task order for car recognition, $F(5, 120) = 2.32$, $p = 0.048$, $\eta_p^2 = 0.092$, such that participants who executed the car recognition first had higher scores than those who took it in the other orders. However, this effect did not survive Bonferroni correction.

M-CFMT+

Average total performance on the M-CFMT+ is plotted in Fig. 3a. When all four blocks are included (dark blue distribution), average total performance (dark blue line) is reduced ($M = 67.01\%$, $SD = 10.98$) compared to when only the first three blocks (light blue distribution and line) are included ($M = 77.30\%$, $SD = 12.23$). The mean scores from the first three blocks replicate original findings from Duchaine and Nakayama (2006), as well as findings in other studies with large samples (see Table 2). The more difficult fourth block leads to substantially reduced mean performance scores ($M_{\text{reduction}} = 10.29\%$, $SD = 3.37$). In general, performance decreased across blocks of the M-CFMT+ (see Table 2). For recognition performance as a function of target face, see Table 3.

For the M-CFMT+, the ω_{cat} values are as follows: Block 1 $\omega_{cat} = 0.79$, Block 2 $\omega_{cat} = 0.84$, Block 3 $\omega_{cat} = 0.76$, Block 4 $\omega_{cat} = 0.48$. This indicates that the task is reliable in the first three blocks as participants are introduced to targets and the instructions of the test, but the difficulty of the fourth block appears to make the measure less reliable.

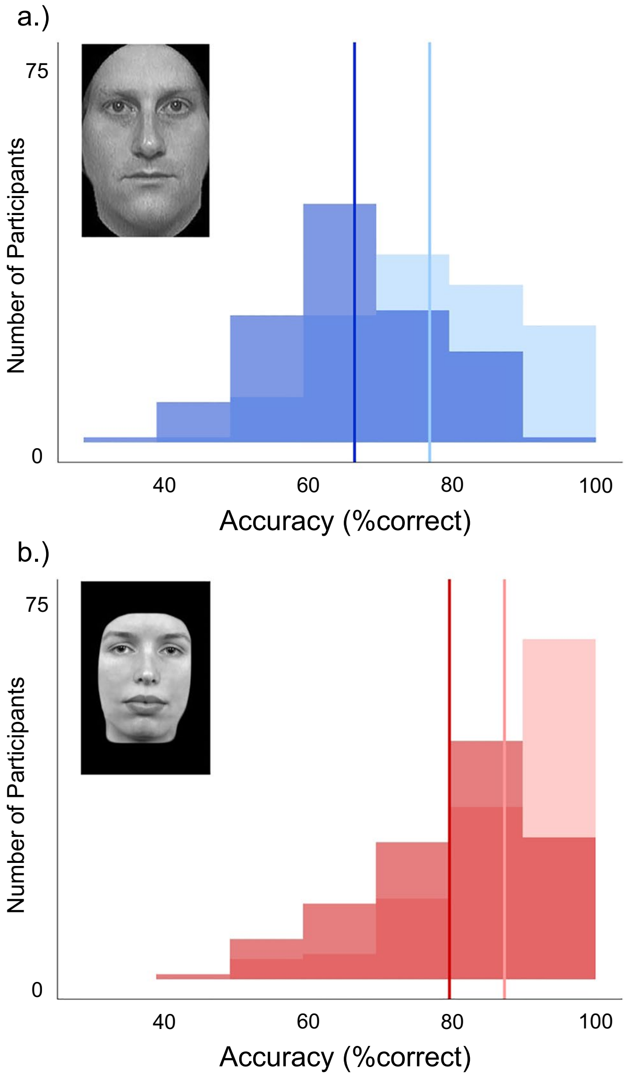


Fig. 3 Distributions of performance scores on the M-CFMT+ and F-CFMT+. Distributions and means (solid lines) for total accuracy (percent correct) plotted as a function of the M-CFMT (**a**) and F-CFMT (**b**) tasks. For each task, the data are plotted as a function of the original first three blocks (i.e., M-CFMT) in light colors, and including the fourth blocks (i.e., M-CFMT+) in dark colors. For both face recognition tasks the distribution of scores and the mean of the distribution are positively shifted with the inclusion of Block 4, indicating the relative difficulty of this block in both tasks

Internal consistency

Block level Performance was largely consistent across blocks for the M-CFMT+. Table 4 shows the average Pearson product correlations between each pair of blocks for the current study as well as for the original study by Duchaine and Nakayama (2006). There is strong internal consistency between all the blocks with one exception; there is no significant association between Blocks 1 and 4. In other words, performance in the easiest recognition

Table 2 Average performance as a function of block in the face recognition tasks

	N	Block 1	Block 2	Block 3	Block 4	Blocks 1–3	Blocks 1–4
M-CFMT+ (current study)	126	98% (5)	76% (16)	64% (17)	42% (12)	77% (12)	67% (11)
F-CFMT+ (current study)	126	98% (4)	90% (12)	77% (17)	62% (16)	88% (10)	80% (11)
Duchaine and Nakayama (2006)	50	n/a	n/a	n/a	n/a	80% (11)	n/a
Bobak et al. (2016)	254	98%	77%	68%	45%	82%	67%
Bate et al. (2018)	40	n/a	n/a	n/a	n/a	n/a	68% (10)
McKone et al. (2011)[a]	75	98%	84%	62%	n/a	80%	n/a
McKone et al. (2012)[a]	20	n/a	n/a	n/a	n/a	66%	n/a

Cells represent mean accuracy (and standard deviation where available) for each sample. [a] McKone et al. (2011) reports mean scores in the CFMT-Aus, and McKone et al. (2012) reports means for European participants in CFMT-Chinese

Table 3 Average performance as a function of target face

	Face 1	Face 2	Face 3	Face 4	Face 5	Face 6
Duchaine and Nakayama (2006)	77%	69%	80%	81%	88%	88%
M-CFMT+ Blocks 1–3	72%	64%	82%	73%	83%	88%
M-CFMT+ Blocks 1–4	63%	56%	73%	64%	70%	77%
F-CFMT+ Blocks 1–3	86%	90%	89%	89%	88%	82%
F-CFMT+ Blocks 1–4	81%	84%	78%	84%	80%	73%

Cells represent mean accuracy

Table 4 Internal consistency of M-CFMT+ across blocks

	Duchaine and Nakayama (2006) (df = 50)	Correlation (df = 126)	Lower limit	Upper limit
B1:B2	0.27	0.44***	0.29	0.57
B1:B3	0.35*	0.36***	0.20	0.51
B1:B4	N/A	0.14	−0.03	0.31
B2:B3	0.74**	0.70***	0.60	0.78
B2:B4	N/A	0.47***	0.32	0.59
B3:B4	N/A	0.59***	0.46	0.69

B1= Block 1; B2 = Block 2; B3 = Block 3; B4 = Block 4; Cells represent Pearson product correlation values; *p < .05; **p < .01; ***p < .001

block was not related to recognition performance in the most difficult block. This may be due to the relatively high, near-ceiling performance in Block 1 across participants. Converging evidence of this interpretation is that correlations between Block 1 and the other blocks (2–3) are relatively lower than are the correlations among the more difficult recognition blocks (2–4), which range from 0.47 to 0.70 on average.

Item level Next, we conducted a series of point-biserial correlations to evaluate how well each individual item was related to participants' overall accuracy in the M-CFMT+. Trials 3 and 10 were excluded because there was no response variability on these items (i.e., 100% correct responses). The average correlation across all trials was relatively low (M = 0.27, SD = 0.13), which is expected given the low variability in performance in Block 1. As a result, we evaluated the correlations separately within each block. As expected, the average correlation was smallest in Block 1 (M = 0.16, SD = 0.10), and increased in Block 2 (M = 0.35, SD = 0.11) and Block 3 (M = 0.34, SD = 0.10). Block 4 (M = 0.19, SD = 0.10) also exhibited a low average correlation, reflecting that performance on items within this block was not consistently associated with overall performance in the task. Only two trials produced negative correlations: Trial 4, $r(126) = -0.02$, $p = 0.785$, *95% CI* [−0.20, 0.15]; Trial 94: $r(126) = -0.07$, $p = 0.454$, *95% CI* [−0.24, 0.19], but the 95% CI for both overlapped with 0. Thus, nearly every trial contributed to the overall sensitivity of the task.

F-CFMT+

Average total performance on the F-CFMT+ is plotted in Fig. 3b. When all four blocks are included (dark red distribution), average total performance (dark red line) is reduced (M = 80.03%, SD = 11.19) compared to when only the first three blocks (light red distribution and line) are included (M = 87.56%, SD = 10.25). The more difficult fourth block leads to substantially reduced performance scores ($M_{\text{reduction}}$ = 7.52%, SD = 3.39). Participants' scores

Table 5 Internal consistency of F-CFMT+ across blocks

	Correlation ($df = 126$)	Lower limit	Upper limit
B1:B2	0.51***	0.37	0.63
B1:B3	0.31***	0.15	0.46
B1:B4	0.28**	0.11	0.43
B2:B3	0.68***	0.57	0.76
B2:B4	0.61***	0.48	0.70
B3:B4	0.72***	0.62	0.79

B1= Block 1; B2 = Block 2; B3 = Block 3; B4 = Block 4; Cells represent Pearson product correlation values; *$p < .05$; **$p < .01$; ***$p < .001$

decreased across each block (see Table 2). For recognition performance as a function of target face, see Table 3.

For the F-CFMT+ the ω_{cat} values are as follows: Block 1 $\omega_{cat} = 0.78$, Block 2 $\omega_{cat} = 0.86$, Block 3 $\omega_{cat} = 0.80$, Block 4 $\omega_{cat} = 0.77$. This indicates consistently high reliability in the F-CFMT+ across the blocks, including the most difficult Block 4, in contrast to the M-CFMT+.

Internal consistency

Block level Performance was also consistent across blocks in the F-CFMT+. Table 5 shows the average Pearson product correlations between each block of the F-CFMT. There is strong internal consistency between all the blocks. As in the M-CFMT+, the correlations between Block 1 and the other blocks are lower ($r = 0.28–0.51$), while the correlations among the more difficult recognition blocks (2–4) are higher, ($r = 0.61–0.72$).

Item level Next, we calculated the point-biserial correlations to evaluate how well each individual item was related to participants' overall accuracy in the F-CFMT+. Trials 10, 15, and 39 were excluded because there was no response variability on these items (i.e., 100% correct responses). The average correlation across all trials was relatively low ($M = 0.31$, $SD = 0.13$). Therefore, we evaluated the correlations separately for each block. As in the M-CFMT+, the average correlation was smallest in Block 1 ($M = 0.14$, $SD = 0.13$), but increased in Block 2 ($M = 0.37$, $SD = 0.11$), Block 3 ($M = 0.37$, $SD = 0.08$), and remained significant in Block 4 ($M = 0.31$, $SD = 11$). Four trials produced negative correlations: Trial 6: $r(126) = -0.05$, $p = 0.545$, *95% CI* [−0.23, 0.12]; Trial 8: $r(126) = -0.003$, $p = 0.975$, *95% CI* [−0.18, 0.17]; Trial 12: $r(126) = -0.02$, $p = 0.836$, *95% CI* [−0.19, 0.16]; Trial 13: $r(126) = -0.06$, $p = 0.519$, *95% CI* [−0.23, 0.12]. Notably, the 95% CI overlapped with 0 for each of these correlations.

M-CFMT+ versus F-CFMT+

Score distribution

Participants performed better overall on the F-CFMT+, $t(125) = 17.30$, $p < 0.001$, $d = 0.53$. When we investigated potential task-related differences in performance at the block level using Bonferroni corrected two-tailed pair-samples t-tests, there were no differences in Block 1, $t(125) = 0.19$, $p = 0.85$, $d < 0.01$. However, scores on the F-CFMT+ were higher in Block 2: $t(125) = 12.20$, $p < 0.001$, $d = 0.41$; Block 3: $t(125) = 9.00$, $p < 0.001$, $d = 0.25$; and Block 4: $t(125) = 14.60$, $p < 0.001$, $d = 0.68$. Importantly, once scores are z-scored, there is no difference in performance between the two tasks, $t(125) < 0.001$, $p = 1.00$, $d < 0.01$.

Following Duchaine and Nakayama's original work (2006), we also plotted the distribution of scores across participants for each item. We visualized this as a function of participants, where the cumulative score for each participant is displayed (Fig. 4a), and as a function of average score at each item (Figs. 4b and 5). These graphs demonstrate two important features of the tasks. First, they demonstrate how scores change across each item of the tasks as they increase in difficulty. For example, in the beginning of the task, participants' scores are largely grouped together, and the standard deviations of mean scores per item are relatively small. However, as participants progress through the task, the difficulty increases, and scores are more widely distributed. Second, the graphs also demonstrate how scores in the F-CFMT+ are differentiated from the M-CFMT+. Figure 5 demonstrates how scores in the F-CFMT+ begin to differentiate from the M-CFMT+ in Block 2, resulting in overall scores that are higher in the F-CFMT+ by the end of Block 4.

Task validity

Divergent validity Divergent validity was assessed by comparing the Cambridge face recognition scores with the Cambridge car recognition scores. As reported in prior work (Dennett et al., 2012; Shakeshaft & Plomin, 2015), total performance on each of the two Cambridge tasks was related to total performance in the CCMT, M-CFMT+: $r(126) = 0.26$, $p = 0.003$, *95% CI* [0.09, 0.42]; F-CFMT+: $r(126) = 0.28$, $p = 0.001$, *95% CI* [0.11, 0.43]. This association likely reflects the similarity in task demands of the recognition paradigm. Importantly, the 95% confidence intervals of the correlations between the face and car tasks do not overlap at all with those of the correlations between the two face recognition tasks (*95% CI* [0.61, 0.79]).

Convergent validity We evaluated the convergent validity of the F-CFMT+ with the M-CFMT+ for both total average performance and separately for block level performance. Average total performance on the two tasks was strongly correlated, $r(126) = 0.71$, $p < 0.001$, *95% CI* [0.61, 0.79]. This correlation is slightly higher than values reported when comparing the CFMT and other versions of the task, such as the original CFMT and the Australian CFMT, $r(74) = 0.61$, $p < 0.01$ (McKone et al., 2011). Across all trials, point biserial correlations were significantly higher in the F-CFMT+ than M-CFMT+, $t(97) = 2.92$, $p = 0.004$, $d = 0.05$, which was driven by differences specifically in Block 4, $t(29) = 4.63$, $p < 0.001$, $d = 0.55$. For a more fine-grained analysis of block- and item-level consistency across tasks, see the Supplementary Materials.

Finally, we compared face recognition scores in the Cambridge tasks to performance in an old-new face recognition task (see Tables 6, 7). We z-transformed recognition scores for the two Cambridge tasks and compared them to d-prime scores from the old-new task[2]. Also, because of prior reports that performance in the CFMT predicts false alarms, but not hits, in old-new face recognition memory tasks (see Devue et al., 2019), we also evaluated how well performance in each of the Cambridge task predicted false alarms and hits in the old-new face recognition memory task as a function of each block.

Overall performance in the M-CFMT+ across all four blocks was not significantly related to total performance in the old/new recognition memory task (as measured by d-prime), $r(112) = 0.16$, $p = 0.091$, *95% CI* [−0.03, 0.34]. However, when the fourth block was excluded and d-prime was based on the first three blocks, there was a significant positive relation with performance in the old-new face recognition task, $r(112) = 0.21$, $p = 0.025$, *95% CI* [0.03, 0.38].

As in prior work, overall performance in the M-CFMT+ did not predict the hit rate in the old-new recognition task, $r(112) = 0.05$, $p = 0.608$, *95% CI* [−0.14, 0.23], but it did negatively predict the false alarm rate, $r(112) = -0.21$, $p = 0.023$, *95% CI* [−0.38, −0.03]. This relation was consistent across the first three blocks of the task but did not hold in the fourth block (see Table 7). In other words, better performance in the first three blocks of the M-CFMT+ predicted fewer false alarms in the old/new face recognition memory task, but this was not evident in the fourth block of the M-CFMT+.

In contrast, accuracy across all four blocks, $r(112) = 0.19$, $p = 0.042$, *95% CI* [0.01, 0.36], and on just the first three blocks, $r(112) = 0.19$, $p = 0.049$, *95% CI* [0.00, 0.36], of the F-CFMT+ was significantly associated with performance in the old-new face recognition task. The task was not significantly related to the hit rate in the old-new recognition memory task, $r(112) = 0.10$, $p = 0.296$, *95% CI* [−0.09, 0.28], but was negatively associated with the false alarm rate, $r(112) = -0.22$, $p = 0.022$, *95% CI* [−0.39, −0.03]. This relation was evident in Blocks 1, 2, and 4 (see Table 7). In other words, fewer false alarms in these three blocks of the F-CFMT+ predicted better performance in the old-new face recognition memory task, but this was not evident in the third block of the F-CFMT+[3].

Discussion

Our primary goal was to present the newly developed Female Cambridge Face Memory Task – Long Form (F-CFMT+) to assesses unfamiliar face recognition. We created the F-CFMT+ to match the design of the M-CFMT+ at the block and trial levels. We identified multiple psychometric properties of the M-CFMT+ and the F-CFMT+, including reliability, internal consistency, convergent validity, divergent validity, target-face accuracy, and item-level sensitivity, and compared them by evaluating performance *in both tasks in the same participants*. These analyses for the M-CFMT+ have not been reported in the literature even though both tasks have been used in prior studies (Bate et al., 2018; Bobak et al., 2016; Elbich & Scherf, 2017; Ramon et al., 2016; Russell et al., 2009). Briefly, the F-CFMT+ has high reliability and internal consistency and a very similar internal task structure to the M-CFMT+; strong convergent validity with the M-CFMT+ and moderate convergent validity with an old-new task of unfamiliar face recognition; and divergent validity from the CCMT.

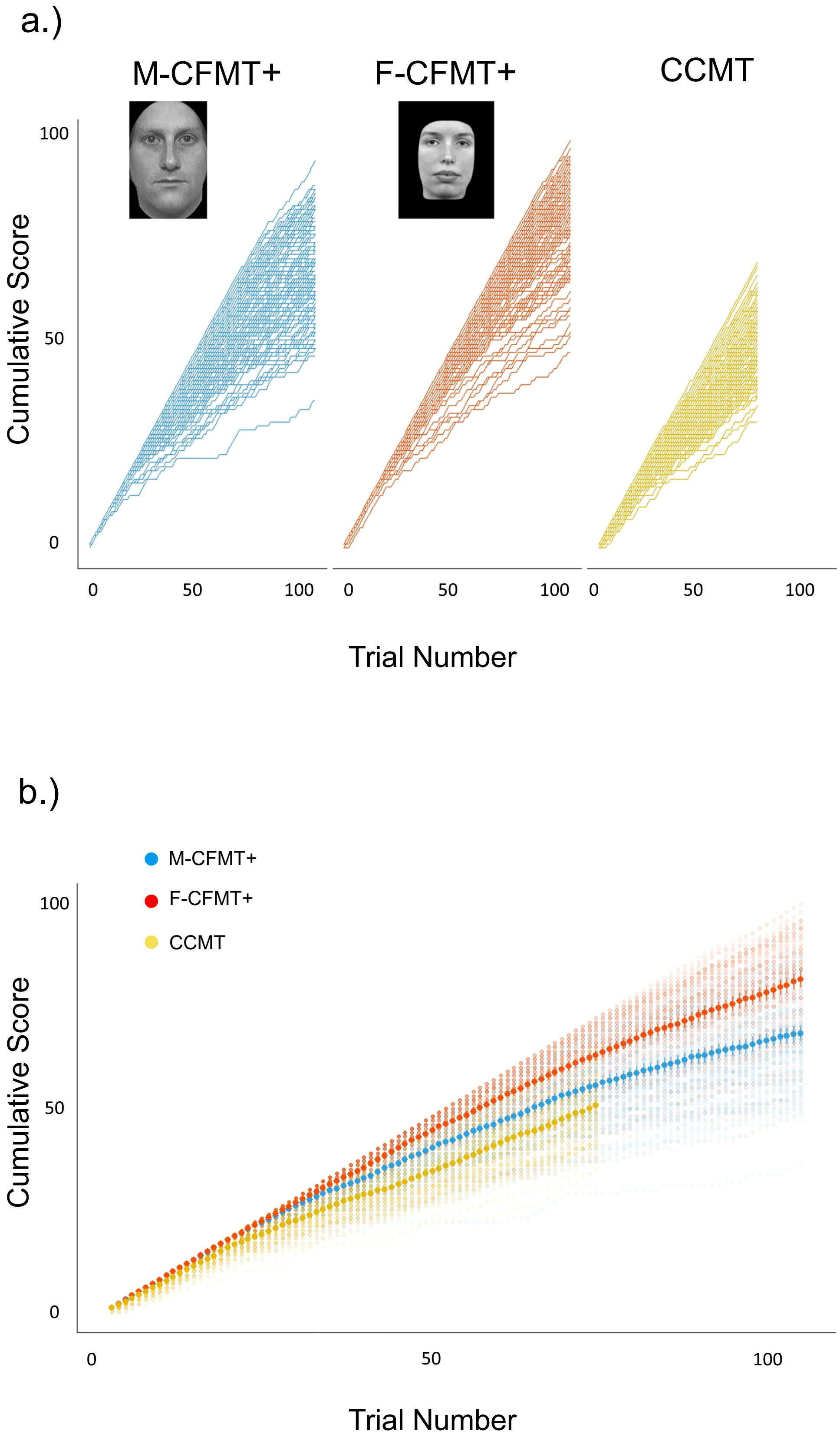
Psychometric properties of the M-CFMT+ and F-CFMT+

To validate the psychometric properties of the F-CFMT+, we needed to measure these properties for the M-CFMT+ as well. Although the authors of the original CFMT provided a clear reporting of the psychometric properties of the task (Duchaine & Nakayama, 2006), there are no empirical reports of the psychometric properties of the M-CFMT+ with the fourth block of trials that was designed to identify superior face recognition abilities (see Russell et al., 2009).

For the M-CFMT+, we found that average total performance decreases significantly (~10%) when including Block 4, indicating the relative difficulty of the trials in this block. This decrement in performance impacted recognition of each of the six target faces similarly. Importantly, there is strong internal consistency, but weak reliability, across the recognition memory blocks even with Block 4. At the

[2] Importantly, accuracy in the CCMT was not related to performance in the old-new face recognition memory task, $r(112) = 0.08$, $p = 0.376$, *95% CI* [− 0.10, 0.27].

[3] Unlike in the M-CFMT+ and F-CFMT+, accuracy in the CCMT did not consistently predict hit rate or false alarm rate in the old-new face recognition memory task.


a.)
M-CFMT+
F-CFMT+
CCMT
100
50
0
Cumulative Score
0
50
100
0
50
100
0
50
100
Trial Number
b.)
● M-CFMT+
● F-CFMT+
● CCMT
100
50
0
Cumulative Score
0
50
100
Trial Number

◂**Fig. 4** Cumulative score for each task. Cumulative score is plotted as a function of individual participant (**a**) for each task to reflect the range of individual differences in this metric of performance across tasks. Note that the car recognition task (CCMT) only has three blocks for a total of 72 trials, whereas the two face tasks have four blocks and a total of 104 trials. In (**b**), the average cumulative score is plotted for each trial ± 1 SEM for each task. The multilevel models revealed that the slope of cumulative scores was steeper in the F-CFMT+ than the M-CFMT+

item-level, there is also consistency particularly within each block, which was largest for Blocks 2–3. In both Blocks 1 and 4, correct performance on any one item is only weakly associated with overall performance, which likely reflects the near-ceiling performance in Block 1 and relatively poor performance in Block 4.

We observed similar psychometric properties in the F-CFMT+. Average total performance was lower by ~7% with the addition of the difficult Block 4. Recognition performance on each of the six female target faces was similarly reduced in Block 4. As in the M-CFMT+, the internal consistency among the blocks was strong. In contrast to the M-CFMT+, reliability was also strong. At the item level, consistency within each block is present and most evident in Blocks 2–4.

Comparing properties of the F-CFMT+ and M-CFMT+

Because we tested the same participants with both Cambridge face recognition tasks, we were able to directly compare the psychometric properties of these tasks to evaluate convergent and divergent validity. Critically, both average total performance and block level performance are highly correlated for the M-CFMT+ and the F-CFMT+. In addition, the reliability and internal consistency of the tasks at the block level was similar but was higher in the F-CFMT+, particularly in Block 4. These findings reflect the strong convergent validity of these two tasks with each other, even though participants were more accurate overall on the F-CFMT+ than the M-CFMT+.

To further assess convergent validity of the F-CFMT+, we compared performance in this task (and the M-CFMT+) to that in a subset of participants who also completed an old-new task of unfamiliar face recognition of male and female emerging adult faces. We found that performance

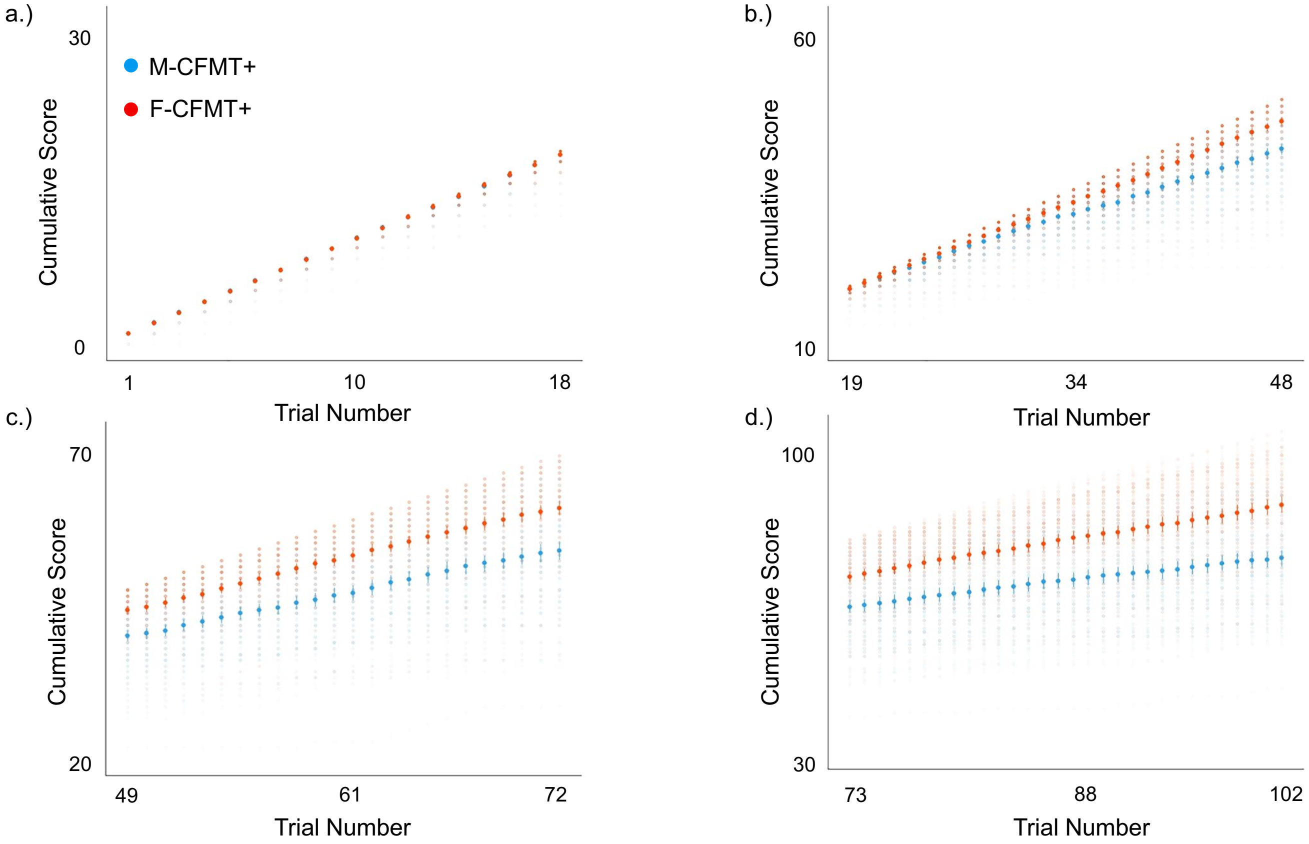


Fig. 5 Comparing changes in average cumulative scores across blocks in the face recognition tasks. In Block 1 (**a**), the increase in the slope of cumulative score is comparable across both tasks. Recall that this is the task instruction and encode faces block. In Blocks 2–4 (**b**–**d**), the increasing slope of cumulative score is slightly higher in the F-CFMT+ task than in the M-CFMT+. This reflects more consistent and successful performance across participants in the F-CFMT+ than the M-CFMT+ across these blocks

Table 6 Assessment of convergent and divergent validity across Cambridge tasks

	M-CFMT+	F-CFMT+	CCMT
D-prime	0.16 [−0.03, 0.34]	**0.19* [0.01, 0.36]**	0.08 [−0.10, 0.27]
Hit rate	0.05 [−0.14, 0.23]	0.10 [−0.09, 0.28]	0.11 [−0.07, 0.30]
False alarm	**−0.21* [−0.38, −0.03]**	**−0.22* [−0.39, −0.03]**	−0.02 [−0.21, −0.16]

Cells represent Pearson product correlation values between total performance on each Cambridge task and the total performance on a task of old/new face recognition memory (i.e., d-prime) and the signal (hit rate) and noise (false alarm rate) trials separately with 95% confidence intervals; *$p < .05$

in the F-CFMT+ was more strongly associated with performance in the old-new recognition memory task than was performance in the M-CFMT+, particularly when scores included Block 4. Also, performance in both Cambridge tasks failed to predict hit rates in the old-new face recognition memory task but did predict false positive rates. Specifically, individuals with better performance in the M-CFMT+ and/or the F-CFMT+ exhibited fewer false positives (i.e., better recall) in the old-new face recognition memory task. Importantly, performance in the car task was unrelated to performance in the old-new face recognition memory task and failed to predict both hit and false positive rates in the unfamiliar face recognition memory task. These results reflect both the convergent and divergent validity of the F-CFMT+.

In sum, the F-CFMT+ and the M-CFMT+ exhibit strong convergent validity and internal consistency at the block and item levels. The two tasks also exhibit good divergent validity with a task of car recognition. The notable differences between the F-CFMT+ and the M-CFMT+ are more reliable responses in the F-CFMT+ and stronger convergent validity between the F-CFMT+ and an old-new face recognition memory task, particularly in terms of performance in the most difficult Block 4.

Limitations and caveats

There are some subtle differences between the M-CFMT+ and the F-CFMT+ to note. First, because of the limitations in digital face stimuli during the development of the F-CFMT+, there are fewer total distractor identities in the F-CFMT+ (38) than in the M-CFMT+ (46). This influenced the relative frequency with which distractor faces occurred in each task. In the M-CFMT+, distractors occurred 1 in every 23 trials during the first three blocks and 1 in every 6.5 trials in the last block. In the F-CFMT+, distractors occurred 1 in every 20 trials during the first three blocks and 1 in every 7 trials in the last block. Recall that there are total of 102 trials, so that distractors repeated approximately 4.4 times in the M-CFMT+ and 5.1 times in the F-CFMT+. This difference in the frequency of distractors may have made them easier to reject as targets in the F-CFMT+, which may have contributed to the task differences in overall performance.

Second, across the entire task, participants performed more accurately in the F-CFMT+ than in the M-CFMT+. Our findings indicate that these differences largely originate from performance in the more difficult blocks, especially Block 4. For the F-CFMT+, performance in Block 4 is reliable and exhibits external validity by predicting low false positive rates in an old-new face recognition task. This is not the case for performance in Block 4 of the M-CFMT+. This is the first time that the psychometric properties of the M-CFMT+ have been investigated, and performance in Block 4 is inconsistent within and across individuals. This varied performance in Block 4 may explain why the task has been useful for identifying individual differences in face recognition abilities, particularly when used in combination with other strong measures of unfamiliar face recognition (see Elbich & Scherf, 2019). Other alternative measures of the CFMT, as in the CFMT-Aus (McKone et al., 2011), only developed tasks that were matched on the first three blocks of trials, which is more consistent within and between participants.

Table 7 Assessment of convergent and divergent validity across Cambridge tasks as a function of block

	M-B1	M-B2	M-B3	M-B4	F-B1	F-B2	F-B3	F-B4	C-B1	C-B2	C-B3
D-prime	0.17	0.18	**0.20***	−0.02	**0.29****	0.16	0.14	0.17	**0.22***	−0.03	0.10
Hit rate	0.06	0.12	0.07	−0.11	**0.21***	0.01	0.14	0.08	**0.20***	0.06	0.02
False alarm rate	**−0.21***	**−0.19***	**−0.24***	−0.09	**−0.26****	**−0.26****	−0.10	**−0.19***	−0.16	0.07	−0.05

Cells represent Pearson product correlation values between total performance on each Cambridge task block and the total performance on a task of old/new face recognition memory (i.e., d-prime) and the signal (hit rate) and noise (false alarm rate) trials separately with 95% confidence intervals; *$p < .05$; **$p < .01$. M = M-CFMT+, F = F-CFMT+, C = CCMT; B1 = Block 1, B2 = Block 2, B3 = Block 3, B4 = Block 4

Therefore, our recommendation is that it is essential to normalize scores when comparing performance on the F-CFMT+ and M-CFMT+ directly as in evaluating gender differences in performance or in the own-gender bias in face recognition.

Third, it is essential to note that these two tasks of face recognition specifically measure recognition of unfamiliar, young adult White faces. Familiar face recognition reportedly involves an overlapping, but different set of processes and underlying neural architecture than unfamiliar face recognition (Gobbini & Haxby, 2007); future research may do well to incorporate the use of a familiar recognition measure. In addition, when characterizing the results generated from these tasks, they should be contextualized in this way and not as a ubiquitous reflection of face recognition abilities in general. Research with children and adolescents reflects that caregiver and peer biases in face recognition abilities will differentially influence the way individuals from these developmental groups recognize the faces in these tasks (e.g., Picci & Scherf, 2016). Also, the ethnic and racial identification of participants may influence the way individuals recognize faces in these tasks (McKone et al., 2012, 2017), which is associated with real-world social experiences (Estudillo et al., 2020; McKone et al., 2021).

Conclusion

The new F-CFMT+ is a valid, internally consistent measure of unfamiliar face recognition that can be used alone or in tandem with the M-CFMT+ to assess recognition abilities for young adult White faces. When used together, performance on the F-CFMT+ and M-CFMT+ can be directly compared, adding to the ability to extrapolate about face recognition abilities across different sets of faces. By making the F-CFMT+ available to researchers, this will support the study of the behavioral and neural basis of face identity recognition. This task will enable researchers to have a standardized task for studying the perception and recognition of unfamiliar female faces, which is underrepresented in the current literature. Future researchers can use the F-CFMT+ to develop a better understanding about gender differences in face identity recognition and gender biases in face recognition (e.g., own-gender bias). This will enable researchers to fill important gaps in the literature evaluating a broad range of questions about face recognition abilities in both typically developing individuals as well as those with atypical social information processing abilities.

Availability data The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Funding The research reported in this paper was funded by the Department of Psychology and the Social Science Research Institute at Pennsylvania State University. We would like to thank Natalie Motta-Mena for her contributions to this project in its earlier stages of task development and pilot data collection.

Declarations

Conflicts of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

Ethics Approval was obtained from the Institutional Review Board at Pennsylvania State University – University Park. The procedures used in this study adhere to the tenets of the Declaration of Helsinki.

Consent statement Informed consent was obtained from all individual participants included in the study.

References

Baron-Cohen, S., Wheelwright, S., Skinner, M., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (QA): Evidence from Asperger syndrome/high functioning autism, males and females, scientists and mathematicians: Errata. *Journal of Autism and Developmental Disorders*, 31(6), 603.

Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Murray, E., Bobak, A. K., ... Richards, S. (2018). Applied screening tests for the detection of superior face recognition. *Cognitive Research: Principles and Implications, 3*, 19.

Bates, D., Mächler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bobak, A. K., Pampoulov, P., & Bate, S. (2016). Detecting superior face recognition skills in a large sample of young British adults. *Frontiers in Psychology, 7*, 11.

Bowles, D. C., McKone, E., Dawel, A., Duchaine, B., Palermo, R., … & Yovel, G. (2009). Diagnosing prosopagnosia: Effects of ageing, sex, and participant-simulus ethnic match on the Cambridge Face Memory Test and Cambridge Face Perception Test. Cognitive *Neurospsychology, 26*(5), 423-455.

Dennett, H. W., McKone, E., Tavashmi, R., Hall, A., Pidcock, M., Edwards, M., & Duchaine, B. (2012). 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*, *44*(2), 587-605.

Devue, C., Wride, A., & Grimshaw, G. M. (2019). New insights on real-world human face recognition. *Journal of Experimental Psychology: General*, 148(6), 994-1007. https://doi.org/10.1037/xge0000493.

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.

Elbich, D. B., & Scherf, S. (2017). Beyond the FFA: Brain-behavior correspondences in face recognition abilities. *NeuroImage*, 147, 409-422.

Estudillo, A. J., Lee, J. K. W., Mennie, N., & Burns, E. (2020). No evidence of other-race effect for Chinese faces in Malaysian non-Chinese population. *Applied Cognitive Psychology, 34*(1), 270-276.

Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Sciences, 3*(4), 484-501.

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

Gobbini, M. I., & Haxby, J. V. (2007). Neural systems for recognition of familiar faces. *Neuropsychologia, 45*(1), 32-41.

Herlitz, A., & Lovén, J. (2013). Sex differences and the own-gender bias in face recognition: A meta-analytic review. *Visual Cognition, 21*(9-10), 1306-1336.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2021). semTools: Useful tools for structural equation modeling. R package version 0.5-5. Retrieved November 8, 2021, from https://CRAN.R-project.org/package=semTools.

Kassambara, A. (2021). rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R package version 0.7.0. Retrieved February 26, 2021, from https://CRAN.R-project.org/package=rstatix.

Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud faces database. *Cognition and Emotion*, 24(8), 1377-1388.

Lundqvist D, Flykt A, Öhman A (1998). *The Karolinska directed emotional faces (KDEF)* [Database record]. APA PsycTests. https://doi.org/10.1037/t27732-000.

McKone, E., Hall, A., Pidcock, M., Palermo, R., Wilkinson, R. B., Rivolta, D., ... O'Connor, K. B. (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.

McKone, E., Stokes, S., Cohan, S., Fiorentini, C., Pidcock, M., Yovel, G., Broughton, M., & Pelleg, M. (2012). A robust method of measuring other-race and other-ethnicity effects: The Cambridge Face Memory Test format. *Plos One, 7(10)*, 1-6.

McKone, E., Wan, L., Robbins, R., Crookes, K., & Liu, J. (2017). Diagnosing prosopagnosia in East Asian individuals: Norms for the Cambridge Face Memory Test–Chinese. Cognitive Neuropsychology, 34(5), 253-268.

McKone, E., Dawel, A., Robbins, R. A., Shou, Y., Chen, N., & Crookes, K. (2021). Why the other-race effect matters: Poor recognition of other-race faces impacts everyday social interactions. *British Journal of Psychology*, https://doi.org/10.1111/bjop.12508.

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*, 3955-3960.

Picci, G., & Scherf, K. S. (2016). From caregivers to peers: Puberty shapes human face perception. *Psychological Science, 27*(11), 1-13.

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved February 26, 2021, from https://www.R-project.org/.

Ramon, M., Miellet, S., Dzieciol, A. M., Konrad, B. N., Dresler, M., & Caldara, R. (2016). Super-memorizers are not super-recognizers. *PLoS ONE, 11*(3), 24.

Revelle, W. (2020). psych: Procedures for Personality and Psychological Research, Northwestern University. Retrieved February 26, 2021, from https://CRAN.R-project.org/package=psych.Version=2.0.12.

Rosseel, Yves (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 1-36. Retrieved August 26, 2021, from https://www.jstatsoft.org/v48/i02/.

RStudio Team (2020). *RStudio: Integrated Development for R*. RStudio, PBC. Retrieved February 26, 2021, from http://www.rstudio.com/.

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

Scherf, K.S. (2021). Introducing the Female Cambridge Face Memory Test – Long Form (F-CFMT+). *Databrary*. https://doi.org/10.17910/b7.1396.

Scherf, K. S., Elbich, D. B., & Motta-Mena, N. V. (2017). Investigating the influence of biological sex on the behavioral and neural basis of face recognition. *Eneuro*, 4(3).

Shakeshaft, N. G., & Plomin, R. (2015). Genetic specificity of face recognition. PNAS *Proceedings of the National Academy of Sciences of the United States of America, 112*(41), 12887-12892.

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

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., … & Nelson, C. (2009). The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research, 168*, 242-249.

Whyte, E. M., Behrmann, M. Minshew, N. J., Garcia, N. V., & Scherf, K. S. (2016). Animal, but not human, faces engage the distributed face network in adolescents with autism. *Developmental Science, 19*(2), 306-317.

Wickham et al., (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686.

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., ... & Duchaine, B. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of sciences*, *107*(11), 5238-5241.

Open practices statement The task is available for experimental purposes on Testable.org (tstbl.co/174-523) and via download on Databrary (Scherf, 2021; https://doi.org/10.17910/b7.1396).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.