scientific reports

Check for updates

OPEN

Investigating the stability of individual differences in face recognition behavior

Myles N. Arrington[1] & K. Suzanne Scherf[1,2]✉

Individual differences in face recognition abilities are characterized as heritable and resilient to change. However, this work is largely based on inter-individual differences, tends to include participants with extreme behavior (e.g., prosopagnosia, super-recognizers), and does not accommodate patterns of bias in intra-individual recognition behavior. Here, we investigated the continuity and stability of intra-individual differences in face recognition behavior among emerging adults using two tasks of unfamiliar face recognition that differ in the gender of the faces to be recognized. Although the estimate of stability is high (0.71) across the sample, there are instabilities in the behavior of many individual participants. For example, approximately 16.7% of the sample exhibited a discrepancy between tasks that was larger than 1 SD. Also, stability was more characteristic of extreme behavior. This is a bit surprising given the potential for close generalization of performance across these two tasks (identical structure and similar stimuli). Inter-individual differences in participant characteristics (i.e., gender, age, social skills) do not explain this variability. These findings are difficult to accommodate into current models of individual differences in face recognition behavior.

Keywords Face recognition, Face processing, Individual differences, Autism, Gender differences, Own-gender bias, Sex differences, Social behavior

Individuals vary in their ability to recognize faces[1], which is especially apparent for those who exhibit extreme faculties. For example, those diagnosed with developmental prosopagnosia[2] (i.e., face blindness) often cannot recognize their own children, and those who never seem to forget a face (i.e., the super-recognizers[3]) excel far beyond typical recognizers. Researchers argue that these individual differences in face recognition abilities are heritable and resilient to change[4–6]. They have searched for social antecedents of these individual differences, like extraversion[7,8] and hometown size[9,10] and have generated important questions about social sequalae of these individual differences, making face recognition skills targets for intervention[11–13].

There are three important caveats regarding our understanding of individual differences in face recognition abilities based on the current literature. First, many studies of individual differences include participants with extreme behavior (e.g., super-recognizers, prosopagnosics) together with those in the more typical range of performance[11,14–16]. This approach assumes that individual differences within the typical range parallel individual differences in the atypical range. However, this is an empirical question[17].

Second, most studies of individual differences in face recognition abilities are cross-sectional, evaluating between-subjects data. In other words, findings overwhelmingly reflect *inter-individual* differences (e.g., heritability differences between monozygotic and dizygotic twins[5], association between face recognition behavior and social network size[18]). Importantly, claims about the strength and stability of individual differences in behavior need to consider *intra-individual* differences as well (i.e., how the same individual performs on multiple tests). This can be addressed by distinguishing between *continuity* and *stability* in behavior over time[19]. *Continuity* reflects the consistency or inconsistency (i.e., relative change) in a group mean-level characteristic over time[19]. A relevant example of continuity in this field is that individuals diagnosed with prosopagnosia *consistently* perform worse as a group than "normal" face recognizers on multiple tests of unfamiliar face recognition over time[20,21]. On the other hand, *stability* measures the consistency in *relative ordering* of individuals on a characteristic over time[19]. Stability in face recognition abilities of individual prosopagnosic participants can be observed when their scores consistently fall in the lowest tail of the distribution of scores across multiple tests[22]. Instability is reflected in a change in the relative order, standing, or rank of individuals in a group on a characteristic over time[19]. Importantly, across tasks of face recognition, there is often *instability* in the rank

[1]Department of Psychology, Pennsylvania State University, 113 Moore Building, University Park, PA 16802, USA. [2]Social Science Research Institute, Pennsylvania State University, 113 Moore Building, University Park, PA 16802, USA. ✉email: suzyscherf@psu.edu

ordering of individuals' performance[23,24]. Therefore, it is an open question whether intra-individual differences in face recognition are stable.

Finally, the Cambridge Face Memory Test (CFMT[20]) is typically used to assess individual differences in face recognition abilities. Critically, the CFMT only tests recognition of unfamiliar *young adult White male faces*. There are alternative forms of this task, including the CFMT-Australian (CFMT-Aus[25]), CFMT-Chinese (CFMT-C[26]), and most recently, the CFMT-Malaysian (CFMT-MY[27]). These tasks also exclusively use male faces and have been developed, in part, to evaluate potential *discontinuities* in face recognition behavior (e.g., own-race bias). Importantly, using multiple versions of the same task that test recognition for different kinds of faces provides a critical opportunity to understand whether and how *continuity* and *stability* in face recognition behavior interact. For example, one study[28] employed the standard CFMT and the CFMT-C to evaluate whether White participants who score at the lower end of the CFMT are more likely to score in the prosopagnosia range on the CFMT-C. They evaluated the relative stability of low face recognition scores in the context of two face processing tasks that produce discontinuous behavior (i.e., White > Chinese recognition in White participants).

In this work, we begin to address these issues by evaluating the stability and continuity of intra-individual differences in face recognition abilities for White male *and female faces*. We do so across the full spectrum of face recognition abilities and in the absence of extreme behavior. We also evaluate how well participant characteristics (gender, social skills, age) contribute to discontinuities in face recognition behavior. What follows is a brief overview of the literature on intra-individual differences in face recognition behavior and findings relating participant characteristics to face recognition behavior.

Intra-individual differences in face recognition behavior

Table 1 provides a representative sampling of studies that evaluate intra-individual differences in face recognition abilities using some version of the CFMT. The standard test of intra-individual differences in performance involves a test–retest reliability analysis, which is typically measured with a Pearson's correlation. Across the existing literature, studies employing versions of the CFMT have reported a wide range of test–retest reliability correlations that vary from 0.67 to 0.92[5,22,25,29,30]. A handful of studies compared performance of the same

Study	Task(s)	*N*	Estimate Test	Estimate	Distribution M (*SD*), [Range]	Notes
Current study	M-CFMT, F-CFMT	126	Pearson's *r* Spearman's rho	0.70 [0.60, 0.78] 0.68 [0.56, 0.77]	M-CFMT: 55.7 (*8.8*) items [29–71] F-CFMT: 63.0 (*7.4*) items [38–72]	
	M-CFMT +, F-CFMT +	126	Pearson's *r* Spearman's rho	0.71 [0.61, 0.79] 0.72 [0.61, 0.80]	M-CFMT +: 68.4 (*11.2*) items [36–95] F-CFMT +: 81.6 (*11.4*) items [48–100]	
Test–retest						
Wilmer et al. (2010)	CFMT	389	Pearson's *r*	0.70 [0.64, 0.74]	Test: 76.9% (*12.9*) Retest: 83.2% (*12.9*)	
Stantić et al. (2022)	CFMT	69	Pearson's *r*	0.67 [0.52, 0.78]	Test: 56.7 (*10.2*) items Retest: 55.5 (*10.6*) items	61% women
McKone et al. (2011)	CFMT-Aus	75	Pearson's *r*	0.84 [0.76, 0.90]	Test: 57.7 (*7.34*) items [40–70]	Baseline + short delay, ages 18–66 years, 55% women
		75	Pearson's *r*	0.84 [0.76, 0.90]		Baseline with long delay, ages 18–66 years, 55% women
Murray & Bate (2020)	CFMT	70	Spearman's rho	0.68		DP participants, 76% women
Petersen & Leue (2022)	CFMT +	89	Spearman's rho	0.89	Test: 84.2 (*11.06*) items [52–101] Retest: 88.7 (*11.30*) items [50–101]	White participants, ages 21-69 years, 68% women
Alternate-forms						
Wilmer et al. (2010)	CFMT, CFMT3	42	Pearson's *r*	0.76 [0.59, 0.86]	CFMT: 72.0% (*16.4*) CFMT3: 78.6% (*16.3*)	
McKone et al. (2011)	CFMT, CFMT-Aus	74	Pearson's *r*	0.61 [0.44, 0.74]		White participants, ages 18–66 years, 55% women
Robertson et al. (2020)	CFMT, CFMT-C	111	Pearson's *r*	0.65 [0.53, 0.75]	CFMT: 76.0% (*12.0*) [49–100] CFMT-C: 71.0% (*11.0*) [43–100]	White participants, ages 18–53 years, 84% women
DeGutis et al. (2013)	CFMT, CFMT-C	53	Pearson's *r*	0.79 [0.66, 0.87]	CFMT: 81.6% (*11.8*) CFMT-C: 75.6% (*12.1*)	White participants, 54% women
Wan et al. (2017)	CFMT-Aus, CFMT-C	268	Pearson's *r*	0.65 [0.58, 0.71]		White participants, ages 17–49 years, 66% women
		176	Pearson's *r*	0.63 [0.53, 0.71]		Asian participants, ages 17-32 years, 70% women
Kho et al. (2024)	CFMT-C, CFMT-MY	124	Pearson's *r*	0.59 [0.46, 0.69]	CFMT-C: 79% (*12*) CFMT-MY: 83% (*10*)	Chinese Malaysian participants, ages 18–66 years, 69% women
Murray & Bate (2020)	CFMT, CFMT-Aus	46	Spearman's rho	0.58		DP participants

Table 1. Representative sample of findings evaluating reliability of the CFMT. *DP* developmental prosopagnosia participants, *CFMT* Cambridge Face Memory Test with 3 blocks, *CFMT +* Cambridge Face Memory Test with 4 blocks, *CFMT-Aus* Cambridge Face Memory Test-Australian, *CFMT-C* Cambridge Face Memory Test-Chinese, *CFMT-MY* Cambridge Face Memory Test-Malaysian, *CFMT3* Cambridge Face Memory Test with computer generated faces. Demographics for each study are reported as available.

participants on different versions of the CFMT to generate an alternative-forms reliability estimate (see Table 1). Across the existing studies, the alternative-forms reliability estimates range from 0.58 to 0.79[5,22,25,27,28,31,32], which is overlapping but notably lower than the range of test–retest reliability estimates.

Critically, it is important to note that reliability estimates quantify the reproducibility of the *measurement when participants are stable* in performance[33]. Furthermore, Pearson's correlation is influenced by outlier scores. The less homogenous the sample, the larger the magnitude of Pearson's correlation[33]. Therefore, it is important to evaluate how these reliability estimates are influenced by extreme scores and to specifically measure the *stability* of participant scores across tests. There are only two studies that measured the stability of participant scores over time/measures using with the CFMT paradigm and Spearman's rho[22,29], which measures stability via continuity in rank ordering. These studies revealed vastly different estimates (see Table 1).

Therefore, rather than assume that there is (or is not) stability of both typical and extreme face recognition performance, we argue that it is important to investigate this as an empirical question. For example, people with prosopagnosia could have consistently poor face recognition skills across all contexts and tests[34], while super-recognizers' abilities might be fine-tuned towards the faces that they experience most often (i.e., in-group faces[23,35]). Therefore, we evaluated the stability of performance in the presence and absence of these extreme scores and separately for the extreme performers.

Personal characteristics that may influence stability in face recognition behavior

In addition to extreme behavior, there are many inter-individual characteristics that potentially contribute to instabilities in face identity recognition behavior. Some of the most prominent characteristics include participant gender, social skills, and age. The original CFMT was designed with male faces to avoid potential gender differences in recognition behavior[20] given reports that women may excel at recognizing female compared to male faces via the own gender bias[36] (OGB). The OGB is less consistently reported in male participants[37,38]. There are numerous confounds involving the task/stimulus design and inclusion criteria for participants in much of the existing work evaluating the OGB[39]. In recent work that addressed these confounds, there is no evidence of an OGB in the behavior or neural network activation for women or men[39]. With the availability of the new F-CFMT+, we can evaluate the potential contribution of stimulus gender and observer gender to potential instabilities in face recognition skills across individuals.

Inter-individual differences in face identity recognition are also associated with social behavior. For example, face recognition abilities are reportedly positively associated with extraversion[7,8], empathy[40,41], and social network properties[42], and negatively associated with narcissism[43]. There are also reports that people with more autism-like social traits in their personality perform worse on the CFMT[44–46] and other tasks of face identity recognition[47]. Therefore, we can evaluate the potential contribution of variations in social skills to the stability of behavior across multiple tasks of face recognition behavior (e.g., maybe people with worse social skills exhibit instability across face recognition tasks).

Finally, research has also indicated that individual differences in face recognition abilities are related to participant age, even in early adulthood[48,49]. Indeed, performance in the original CFMT reportedly peaks at age 30[48] and declines across established adulthood[50]. Therefore, we can evaluate the potential contribution of variations in age to the stability of face recognition behavior (e.g., maybe older individuals have more stable performance across tasks).

Current study

Here, we explore these questions by investigating the stability of face recognition behavior in a large sample of typically developing emerging adults[51] (18–25 years of age) across the M-CFMT+[3] and F-CFMT+[52]. These tasks include a fourth task block that was specifically designed to elicit inter-individual differences in performance. The M-CFMT + is particularly good for identifying extreme recognition behavior on the superior end of the continuum (i.e., super-recognizers[3,53]). The recently developed F-CFMT + is comparable to the M-CFMT + in all task parameters, but uses young adult, White, *female* faces as stimuli[52]. The psychometric properties of these tasks are highly similar, including the internal reliability, convergent and divergent validity, making them good tests for evaluating the stability of intra-individual differences in face recognition behavior[52].

This work was guided by five goals. First, to connect with the existing literature, we determined the alternative-forms reliability for face recognition when the face stimuli differ by *gender*. Second, we determined the stability of performance between the two tasks by computing the Spearman's rho rank coefficient. Third, we assessed whether stability of performance is contingent upon extreme scores. Fourth, we tested patterns of stability for extreme performance. Lastly, we evaluated whether person characteristics (gender, age, social skills) predict instabilities in face recognition behavior in response to male and female faces (e.g., more consistent performance in male than female participants).

Methods

Study design

This study analyzed data that were previously used to evaluate the psychometric properties of the M-CFMT + and F-CFMT +[52]. Importantly, none of the previously published analyses investigated individual differences in performance. The study uses a within-subjects design, such that each participant completed both the M-CFMT +[3] and the F-CFMT +[52] in the same testing session. Participants also completed two additional recognition tasks that have been described in our previous work[52,54] but are not the focus of the current study.

Participants

Recruitment and screening

Participants were recruited via the Psychology Department Subject Pool and from flyers posted around campus. Recruitment materials did not advertise a study about face recognition. Instead, they described the goal of the study "to learn more about how perception and brain function vary in typically developing adults, particularly in the way they look at pictures of human faces." Participants indicated their interest in the study by sending an email to the research team, who then provided an email link to the screening materials, or by scanning a QR code that linked them to the online Qualtrics screening materials.

The online materials included a screening consent form and several questions to assess the likelihood that the potential participant would meet the initial inclusion criteria. This recruitment strategy allowed us to quickly and efficiently eliminate potential participants who were unlikely to meet the eligibility criteria of the study. Approximately 50% of potential participants were screened out due to being outside the target age range, having active psychiatric or neurological conditions, being non-native English speakers, and providing incomplete forms. Participants whose responses indicated that they were likely to be eligible for the study were invited to the Laboratory of Developmental Neuroscience for consenting, assessment of full eligibility, and behavioral testing.

Inclusion and exclusion criteria

Participants had to be (1) between the ages of 18–25 years; (2) capable of cooperating with the study procedures; (3) native English speakers; (4) free of neurologic disorders currently and in the past based on a neurologic history obtained with a questionnaire; (5) free of psychiatric disorders currently and in the past on the basis of a semi-structured interview designed to ascertain present episode and lifetime history of psychiatric illness according to DSM-IV criteria[55]; and to have (6) a negative family history in first degree relatives of affective and anxiety disorders and other major psychiatric disorder; (7) no historical evidence of significant difficulty during the pregnancy, labor, delivery, or immediate neonatal period or abnormal developmental milestones as determined by questionnaire; (8) no history of loss of consciousness; (9) no sensory impairments such as vision or hearing loss that persist without correction; and (10) Autism Quotient[56] total score < 145.

Autism Quotient[56] scores greater than 145 and ongoing psychopathology may indicate conditions that are associated with atypical face processing behavior[57,58]. These exclusion criteria were designed to help improve accuracy of estimating the potential influence of gender on face recognition abilities by reducing unrelated variability (e.g., due to mental illness symptoms) that likely exists between men and women in the population at large[59,60]. Notably, participants were neither recruited nor included in the final sample based on their perceived or actual face recognition performance.

The final sample included 126 typically developing emerging adults (range 18–25 years, see Table 2). Written informed consent was obtained using procedures approved by the Internal Review Board of the Pennsylvania State University. Assessments for full eligibility were only administered after consent was obtained. Participants who met the final eligibility criteria were invited to continue with the behavioral testing session. The data were collected between January 2018 and March 2020.

All actors in stimuli images and videos signed an informed consent to use the photos.

Power analysis

The study was originally designed to include 150 participants based on an a priori power analysis[52]; however, the COVID-19 pandemic forced us to truncate data collection in March 2020. The original analysis was designed to power a moderately sized within-subjects fixed effect of task at $\alpha = 0.01$ separately in male and female participants. For these analyses, we computed a priori power to detect a large correlation ($r = 0.70$, see Table 1)

	Total sample	Men	Women
Demographics			
N	126	59	67
Age in years (*SD*)	19.6 (*1.6*)	19.6 (*1.6*)	19.6 (*1.6*)
Proportion white	81.0%	86.4%	76.1%
AQ total (*SD*)	104.3 (*12.1*)	107.4 (*14.4*)*	101.5 (*8.8*)
Performance			
M-CFMT	77.3% (*12.2*)	75.3% (*12.8*)	79.1% (*11.5*)
M-CFMT +	67.0% (*11.0*)	67.0% (*11.5*)	68.8% (*10.3*)
F-CFMT	87.6% (*10.3*)	84.5% (*11.5*)*	90.3% (*8.2*)
F-CFMT +	80.0% (*11.2*)	77.0% (*12.4*)*	82.7% (*9.4*)
Own gender bias	0.1 (*0.8*)	0.1 (*0.8*)	0.1 (*0.7*)

Table 2. Demographic and performance characteristics of sample. Cells contain *M (SD)*. Scores for the M-CFMT + and F-CFMT + are reported in accuracy (percent correct). The own-gender bias is calculated by first z-scoring the M-CFMT + and F-CFMT + distributions, then subtracting other-gender recognition from own-gender recognition for each participant. *Significant difference between gender groups at the level of $p < 0.01$.

for the reliability and sensitivity analyses. The sample size of $N = 126$ was sufficient to detect a large correlation with a power of 0.95, and an $\alpha = 0.05$.

Measures and assessments

Questionnaires

Adult Self-Report Inventory-4 (ASRI-4) The ASRI-4 is a 136-item, DSM-IV-referenced rating scale[55]. Participants rate how frequently they experience a behavior described within multiple symptom scales. In this study, participants rated behaviors in the anxiety, mood disorders, ADHD, and personality scales. Scales were scored using the Symptom Severity method[55]. Participants who reached a screening cut-off score[55] for any of these scales were excluded from participating in the study.

Autism-spectrum quotient The Autism-Spectrum Quotient Test[56] (AQ) is a 50-item self-report questionnaire that measures autistic-like traits (ALT) on a 4-point Likert scale (1 = Definitely Agree, 4 = Definitely Disagree). We used the Hoekstra scoring method[61] to compute a total range of scores between 50 and 200 such that higher scores indicate the presence of more ALTs. Any individual who failed to answer more than four items was excluded from the study because the AQ score could not be computed. In this sample, Cronbach's alpha indicated good reliability for the test ($\alpha = 0.75$), which is comparable to existing reports in the literature[57].

Face recognition tasks

Male Cambridge face memory test-long form (M-CFMT+) The M-CFMT+ is a test of unfamiliar face recognition using White young adult male faces[3,20]. It captures a broad range of face recognition abilities[3,62]. The task is divided into four blocks that increase in difficulty (Fig. 1a). In Block 1, participants study six target faces with no hair and neutral expressions in each of three viewpoints. During recognition trials, participants identify target faces by button press in a three-alternative forced choice paradigm under conditions of increasing difficulty. Participants have an unlimited amount of time to respond on each trial; however, they were instructed to "go as fast as you can without making mistakes." The blocks increase in difficulty by including noise, altered lighting, novel images of the target and distractor faces, and the introduction of emotional expressions and varied context. There are a total of 102 trials that are presented in a fixed order for all participants. The task can take ~ 20–30 min to complete. There are formal breaks built in between blocks. There is strong reliability and internal consistency (Cronbach's $\alpha = 0.87$) as well as convergent and divergent task validity[52] for this task.

Female Cambridge face memory test (F-CFMT+) The F-CFMT+ was developed to match all the parameters of the M-CFMT+ at both the block and trial levels[52] (Fig. 1b). The stimuli include high-resolution images of individual White women in the age range of 20–30 years old. The instructions, task structure, and task timing are the same as the M-CFMT+. The reliability and internal consistency (Cronbach's $\alpha = 0.91$) are strong, as are convergent and divergent task validity[52] for this task.

Procedure

The lab visit took between 1.5 and 2 h, including the eligibility evaluation, behavioral testing, and breaks. Participants were invited to take a break (~ 10 min) after completing the consenting and eligibility assessment so that the measures could be scored and the eligibly decision could be determined. Eligible participants completed both Cambridge face recognition tasks (FCFMT+, M-CFMT+) on a laptop computer in a single lab session.

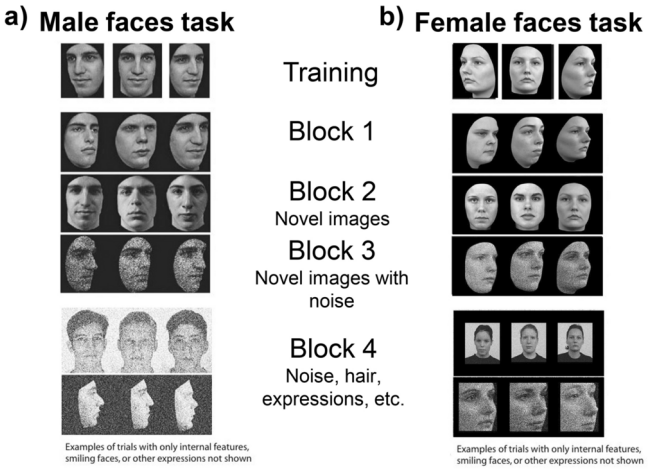

Fig. 1. Male and female face recognition tasks. Task outlines of the (**a**) male (figure adapted from Russell et al., 2009) and (**b**) female (Arrington et al., 2022) versions of the CFMT+ (images of female faces are published with permission from the Rafd and KDEF databases and include images AF16NES, AF19NES, and AF29NES). In these tasks, participants view target identities at multiple viewing angles and then must recognize the target faces among distractors with increasing levels of difficulty across blocks, which add noise with changes in lighting and viewpoint (Block 2), visual noise (Block 3), hair, affect, and repeating distractors (Block 4).

The order of the tasks was counterbalanced separately for male and female participants. A researcher sat in the lab next to the participant as they completed the tasks. This procedure allowed the researcher to address any questions or technical difficulties and to ensure that participants attended to the task. Participants were permitted to take breaks as needed between blocks of the tasks and between tasks. Participants were either paid $20 or received course credit if they were part of the Psychology Department Subject Pool.

Data analysis

All analyses were conducted using R software using the R Studio interface[63,64]. R packages included *tidyverse*[65], *lme4*[66], *lmerTest*[67], *psych*[68], *rstatix*[69], *interactions*[70], *MASS*[71], *car*[72], *ggpubr*[73], *emmeans*[74], *effectsize*[75], *ordinal*[76], *jmv*[77], and *statspsych*[78]. Accuracy was the dependent measure in the analyses. Prior to analyses, the raw scores were examined for violations of normality and used to assess potential differences in variance across tasks. No participants performed at or below chance (i.e., 33% accuracy) on either task. Based on prior recommendations[52], we z-scored raw accuracy when comparing performance across tasks.

Results

Sample characteristics

The demographic characteristics and performance of the sample are provided in Table 2 as a function of gender. Men and women were matched in age, $t(124) = 0.03$, $p = 0.978$, $d = 0.00$. Men reported more autistic-like traits than women, $t(124) = -2.80$, $p = 0.006$, $d = 0.12$, which is consistent with prior research[56,61]. The sample was predominantly White, not Hispanic or Latino (81.0%). In total, 96 individuals self-identified as "White", 4 as "Black", 15 as "Asian", 2 as "Hispanic", 7 as "Multiracial", and 2 as "Other". Because of concerns that non-White participants might exhibit a disproportionate reduction in recognition of White faces due to an own-race bias[79], we compared performance of the White and non-White participants across both tasks; these groups did not differ in face recognition performance on either test, or in demographic variables (all p's > 0.050). This is consistent with reports in the literature that in the United States, non-White participants often do not exhibit an own race bias[80].

Task characteristics

The distributions of accuracy scores on both face recognition tests exhibited normal skew [− 1, 1] and kurtosis [− 3, 3]. Raw scores from the M-CFMT+ were normally distributed as determined by the Shapiro-Wilkes test of normality, $W = 0.99$, $p = 0.325$. However, the scores from the F-CFMT+ were not, $W = 0.95$, $p < 0.001$ (Fig. 2a). This is likely related to slight negative skew in these data, which is common across studies[29,81] using these tasks. Therefore, we used the Fligner-Killeen test to evaluate homogeneity of variances, which is robust against deviations from normality. The variances from the two tasks were not different, X^2 (1) = 0.17, $p = 0.678$.

To evaluate possible task order effects, we submitted raw accuracy to a repeated-measures ANOVA with task (within-persons) and task order (between-persons) as predictors. There was no main effect of order, $F(1,122) = 0.48$, $p = 0.489$, $\eta^2_G = 0.003$, and no task × order interaction, $F(1,122) = 0.02$, $p = 0.893$, $\eta^2_G < 0.001$.

Although reaction time (RT) is not typically reported for these tasks because of the unlimited response time for each trial, we determined that there were no differences in RT between the tasks, $t(125) = 1.86$, $p = 0.065$, using a two-way paired-samples t-test. Also, average RT did not predict accuracy for either task (p's > 0.80: see Supplementary Materials).

What is the alternative-forms reliability of these two face recognition tasks?

To connect with the existing literature, we evaluated the *reliability* in performance across the M-CFMT + and F + CFMT + . Importantly, reliability analyses evaluate the *consistency of the measurement tool over time/forms*. We used Pearson's correlation to assess alternate-forms reliability. We interpreted this result in the context of the upper bound correlation for the two tasks[25,82]. In addition, we evaluated potential differences between the observed correlation and those reported in the prior literature (see Table 1).

The Pearson's correlation analysis indicated that there is strong reliability in performance across these tasks, $r(126) = 0.71$, $p < 0.001$, 95% CI [0.61, 0.79] . However, the upper bound correlation for these two tasks is 0.89, which indicates that the observed estimate of alternate-forms reliability is not at ceiling. Also, there is ~ 49% ($1 - r^2$) variance in scores unaccounted for by this estimate. We compared this correlation to those reported in previous work by using Fisher's transformation to conduct two-tailed t-tests. Our reliability estimate was not different from the alternate-forms reliability of the CFMT as reported in previous work[5,22,25,27,28,31,32], all $p > 0.10$.

What is the stability for face recognition abilities when the face stimuli differ by gender?

We evaluated the *stability of performance* across the M-CFMT + and F-CFMT + . Stability analyses determine whether consistencies in performance across these two tasks *reflect persistence of the underlying characteristic within participants*—namely, their relative ability to recognize faces (not of the tasks to elicit such behavior as in reliability). Instability in scores across these two tasks would indicate that face recognition abilities vary as a function of the gender of faces.

The Spearman's rho analysis indicated stability across the full sample: $\rho(126) = 0.72$, $p < 0.001$, 95% CI [0.61, 0.80]. However, a closer look reveals instabilities in scores. To visualize these instabilities, we computed difference scores between the two standardized recognition scores for each participant ($z_{M\text{-}CFMT+} - z_{F\text{-}CFMT+}$). Figure 2b plots the difference scores from − 2 to + 2. A score of 0 represents perfect stability in performance across the two tasks. A score of + 2 indicates that the participant scored 2 *SD* better on the M-CFMT + . Of note, 16.67% of the sample exhibited a difference of more than 1.0 *SD*, and only 31.74% had a difference of 0.25 *SD* or less (Fig. 3). The range of difference scores between tasks was broad [− 1.93, 2.13]. Given that differences of 1.0

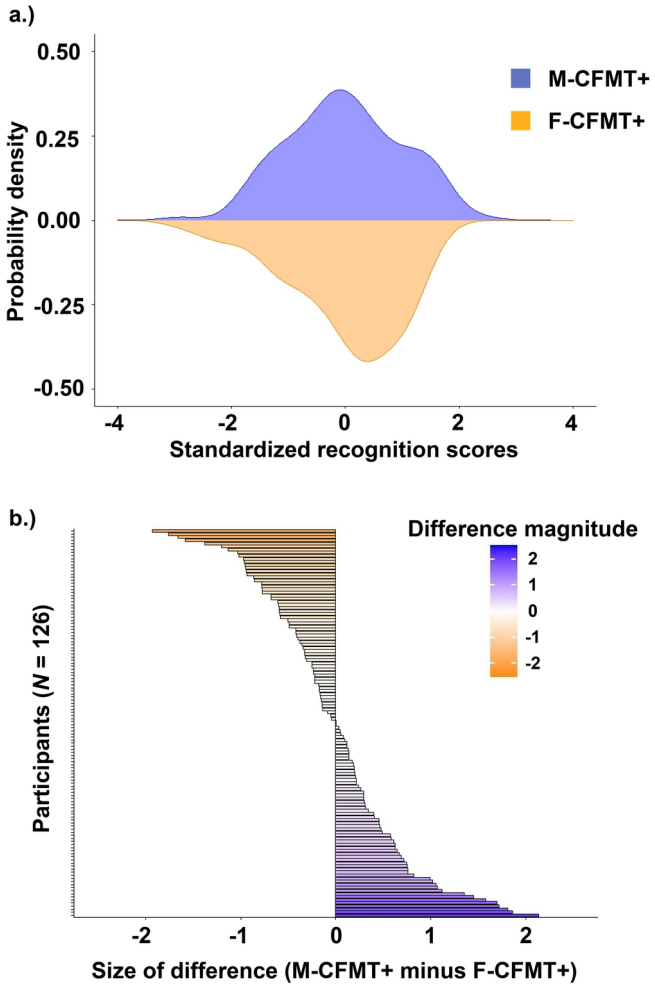

Fig. 2. Score distribution across tasks. (**a**) Kernel-smoothed probability density plot for each task. Y-axis values represent the probability density function of the standardized scores. The individual task distributions are plotted on top in blue (M-CFMT +) and on bottom in gold (F-CFMT +). (**b**) Difference scores between performance on each task. Each bar represents one of the 126 participants in this sample. The magnitude and direction of the bar corresponds to the size of the difference between standardized scores on the two tests. Positive differences (i.e., participants who performed better on the M-CFMT +) are increasingly blue, and negative differences (i.e., participants who performed better on the F-CFMT +) are increasingly orange. Note that there are many participants who exhibit difference scores at or beyond 1 full SD.

SD between participants (within tasks) are clinically meaningful[57], we pursued a more fine-grained analysis to evaluate stability (or lack thereof) in performance.

Next, we determined the quartile ranges of scores separately for each task. Using this approach allowed us to determine *sample-specific criteria* for extreme performance, rather than evaluating several different criteria used to identify face blind and super-recognizer behavior[3,20,23,83–87]. While these existing criteria have been used in several studies, there is no consensus or standardized criterion, and none of the criteria have been tested in the F-CFMT +. Therefore, we used a quartile approach to divide all participants scores into one of four groups (low, low-medium, high-medium, and high ability) for each task (see Table 3). The boundaries of the quartile ranges overlapped especially for the lowest and highest quartiles. At the same time, there were differences in these boundaries that we evaluated more rigorously.

Next, we evaluated how scores on one task predict quartile membership on the other task using an ordinal regression. We extracted predicted probabilities of membership in each quartile from the regression, allowing us to determine stability in quartile placement across tasks. We predicted that if participants are largely stable in performance across the full spectrum of scores, the quartile placements would closely align across all four quartiles. Results indicated that F-CFMT + scores significantly predicted quartile placement on the M-CFMT +, $OR = 5.82$, $p < 0.001$, 95% CI [3.66, 9.73]). Similarly, M-CFMT + scores significantly predicted quartile placement in the F-CFMT +, $OR = 7.40$, $p < 0.001$, 95% CI [4.59, 12.56].

To understand whether different parts of the spectrum of scores are contributing to the overall stability in performance, we identified nine standardized scores across the full spectrum of responses (i.e., between − 2 and 2) at intervals of 0.5. We examined the predicted probabilities of quartile placement for each score on

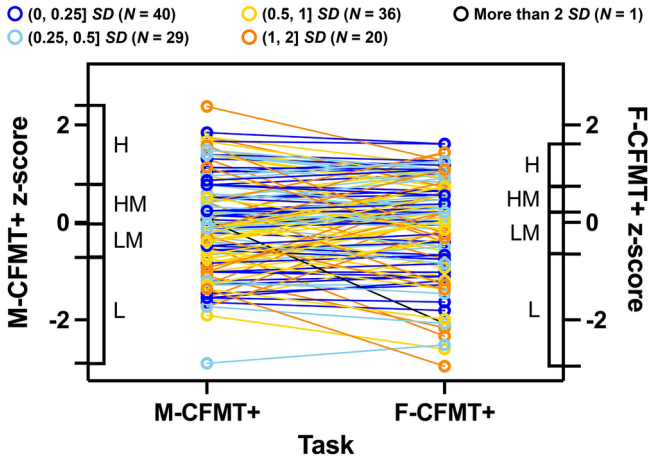

Fig. 3. Stability in participant performance by task. Figure shows a spaghetti plot of standardized scores in the M-CFMT + and the F-CFMT + . Lines in dark blue represent individuals with difference scores less than or equal to 0.25 *SD*, lines in light blue are individuals with a difference score less than or equal to 0.5 *SD*, lines in gold are individuals with a difference less than or equal to 1 *SD*, lines in orange are individuals with a difference less than or equal to 2 *SD*, finally, lines in black are individuals with a difference greater than 2 *SD*. Brackets indicate the ranges of scores for each quartile in the M-CFMT + (left axis) and the F-CFMT + (right axis). Note that 45% of the sample exhibited difference scores greater than 0.5 SD, which is a clinically relevant difference (Griffin et al., 2021).

	M-CFMT +			F-CFMT +		
	Z-range	Items-range	*N*	*Z*-range	Items-range	*N*
Low	[− 2.89, − 0.72]	[36, 60]	32	[− 2.95, − 0.65]	[48, 74]	32
Low–medium	[− 0.71, − 0.03]	[60, 68]	33	[− 0.64, 0.21]	[74, 84]	36
High–medium	[− 0.02, 0.78]	[68, 77]	33	[0.22, 0.74]	[84, 90]	28
High	[0.79, 2.40]	[77, 95]	28	[0.75, 1.61]	[90, 100]	30

Table 3. Quartile ranges of performance for each task. Ranges reported in z-scores and in the number of items scored as correct.

the alternative task (see Table 4). Note that chance levels of stability between quartiles is 25%. For both tasks, scores that fall in the lowest or highest quartiles predict placement in the corresponding quartile of the other task extremely well ($M_{\text{Probability}} = 71.38\%$). However, scores that fall in the two medium quartiles are much less predictive of quartile placement in the alternative task ($M_{\text{Probability}} = 39.11\%$). If a participant produces a score that falls in one of the medium quartiles on one task, their score on the other task is relatively unlikely to be in the corresponding quartile on the other task. A simulation study confirmed that these results were unexpected based on the properties of the two distributions (see Supplemental Materials).

To follow up on this result, we assessed whether participants in the extreme quartiles produced more stable performance than participants closer to the mean. To do so, we identified the individuals whose performance was in the lowest quartiles for *both* tasks ($N = 20$) or in the highest quartiles for *both* tasks ($N = 21$). When excluding the lowest performers, stability in the remaining sample was reduced, $\rho(106) = 0.59$, $p < 0.001$, 95% CI [0.43, 0.71]. Stability was also reduced when excluding only the highest performers, $\rho(105) = 0.58$, $p < 0.001$, 95% CI [0.42, 0.70]. Importantly, the largest reduction in stability came when excluding both sets of extreme performers to evaluate stability in the typical range of performance, $\rho(85) = 0.33$, $p = 0.002$, 95% CI [0.12, 0.51]. In contrast, stability among all the extreme performers (high and low) was especially strong, $\rho(41) = 0.82$, $p < 0.001$, 95% CI [0.66, 0.91]. In other words, stability is much weaker for participants who exhibit face recognition behavior in the more typical range.

Do person characteristics predict stability in performance across tasks?

Finally, we evaluated the possibility that person characteristics influence stability in task performance. Recall that stability reflects the relative rank-ordering across multiple measures[19]. Given previous work, we focused on participant gender[88,89], AQ Total score[44,45], and age[49]. We ran separate linear mixed effect models that examined task × characteristic interactions. We included participant as a random effect and task as a repeated measure. The interaction term indicated whether each characteristic predicted instabilities in recognition behavior. We evaluated interactions by testing the simple slopes. All analyses used the z-scores to assess rank order between participants. Scores for women in the F-CFMT + were treated as the reference.

(a)		M-CFMT + Quartile			
F-CFMT + Score (Quartile)		Low	Low–medium	High–medium	High
− 2.0	L	**88.05% [73.36, 95.17]**	9.61% [3.98, 21.42]	1.91% [0.64, 5.58]	0.44% [0.12, 1.62]
− 1.5	L	**75.33% [58.26, 86.98]**	19.18% [10.38, 32.73]	4.43% [1.88, 10.10]	1.05% [0.36, 3.08]
− 1.0	L	**55.87% [40.67, 70.04]**	31.85% [21.50, 44.37]	9.77% [5.22, 17.56]	2.50% [1.05, 5.82]
− 0.5	LM	34.42% [23.98, 46.63]	**40.34% [29.65, 52.03]**	19.41% [12.66, 28.59]	5.83% [3.02, 10.95]
0.0	LM	17.87% [11.55, 26.63]	**37.25% [27.32, 48.38]**	31.89% [23.15, 42.13]	12.99% [8.03, 20.33]
0.5	HM	8.28% [4.64, 14.34]	25.46% [17.85, 34.94]	**39.79% [29.53, 51.04]**	26.47% [18.33, 36.60]
1.0	H	3.61% [1.67, 7.61]	13.83% [8.38, 21.97]	36.09% [26.04, 47.53]	**46.47% [33.89, 59.52]**
1.5	H	1.53% [0.57, 4.01]	6.52% [3.19, 12.86]	24.27% [14.92, 36.93]	**67.68% [51.69, 80.39]**
2.0	H	0.64% [0.19, 2.11]	2.86% [1.11, 7.17]	13.03% [6.21, 25.27]	**83.47% [68.07, 92.29]**

(b)		F-CFMT + Quartile			
M-CFMT + Score (Quartile)		Low	Low–medium	High–medium	High
− 2.0	L	**90.05% [78.35, 95.77]**	8.48% [3.67, 18.41]	1.19% [0.41, 3.46]	0.28% [0.07, 1.05]
− 1.5	L	**76.86% [61.88, 87.21]**	19.21% [10.77, 31.90]	3.15% [1.35, 7.17]	0.75% [0.25, 2.25]
− 1.0	L	**55.01% [41.07, 68.21]**	35.04% [24.39, 47.43]	7.93% [4.24, 14.33]	2.02% [0.84, 4.80]
− 0.5	LM	31.01% [21.54, 42.40]	**45.88% [34.61, 57.59]**	17.79% [11.42, 26.64]	5.31% [2.70, 10.18]
0.0	HM	14.18% [8.74, 22.18]	40.84% [30.42, 52.17]	**31.73% [22.32, 42.91]**	13.24% [8.02, 21.09]
0.5	HM	5.73% [2.96, 10.81]	25.30% [17.20, 35.58]	**39.64% [28.44, 52.04]**	29.34% [19.95, 40.88]
1.0	H	2.19% [0.91, 5.13]	12.01% [6.73, 20.50]	32.77% [22.26, 45.35]	**53.04% [38.82, 66.78]**
1.5	H	0.81% [0.27, 2.41]	4.92% [2.19, 10.69]	18.82% [10.48, 31.49]	**75.44% [59.71, 86.42]**
2.0	H	0.30% [0.08, 1.12]	1.89% [0.66, 5.27]	8.50% [3.66, 18.54]	**89.31% [76.82, 95.47]**

Table 4. Predicted probabilities of standardized scores from ordinal regression. Leftmost column includes scores from the standardized distribution of the respective task and their corresponding quartile (*L* low; *LM* Low-medium; *HM* High-medium; *H* High). Percentages represent predicted probabilities of the score in each of the four quartiles for the other task, with confidence interval. (a) Predicting quartiles in the M-CFMT + based on score in the F-CFMT + . (b) Predicting quartiles in the F-CFMT + based on score in the M-CFMT + . Bolded cells indicate the predicted probabilities for overlapping quartile groups (e.g., low in F-CFMT + and low in M-CFMT +). Note that the most extreme scores (Low, High) on each task have the highest predicted probabilities for the corresponding quartile on the opposite task.

Gender

We found a significant main effect of participant gender for the F-CFMT + , $b = -0.50$, $p = 0.005$, 95% CI [− 0.85, − 0.16]. Women consistently performed better in the task ($M = 0.24$, $SD = 0.84$) compared to men ($M = -0.27$, $SD = 1.10$). There was no main effect of task for women, $b = -0.07$, $p = 0.427$. There was no gender × task interaction, $b = 0.16$, $p = 0.246$. Simple slope analysis conducted separately for each group indicated that men were consistent across tasks, $b = 0.08$, $p = 0.397$, 95% CI [− 0.11, 0.28].

Next, we examined evidence for an OGB using a "bias" score[39]. The bias score was computed by subtracting each participants' accuracy for other-gender faces from their accuracy for own-gender faces ($z_{own} - z_{other}$). A positive score indicates an OGB. We submitted the bias score to a linear regression model to evaluate the potential main effect of gender. The analysis of the OGB bias score indicated no main effect of gender, $b = 0.01$, $p = 0.941$ (see Fig. 4a). There was no evidence of an own-gender bias in either men ($M = 0.08$, $SD = 0.84$) or women ($M = 0.07$, $SD = 0.68$). Therefore, participant gender did not contribute to instability in performance.

Autism-like traits (ALTs)

Figure 4b plots the standardized face recognition scores as a function of task and AQ Total scores. AQ Total score was negatively associated with F-CFMT + scores, $b = 0.01$, $p = 0.046$, 95% CI [− 0.03, − 0.00], indicating that participants who self-reported more ALTs performed worse in F-CFMT + . Importantly, there was no task × AQ Total score interaction, $t(124) = -0.44$, $p = 0.658$, indicating that the negative association between AQ Total scores and face recognition performance was *consistent* across both tasks. This was confirmed by the simple slope analysis of AQ Total scores on the M-CFMT + , $b = -0.02$, $p = 0.020$, 95% CI [− 0.03, − 0.00].

Given the gender differences in both the AQ Total and face recognition scores in this sample, we investigated whether the relation between AQ Total score and face recognition holds when accounting for the effects of participant gender. We used a mixed effects linear model, with participant gender and AQ Total score as between-participants variables, and task as a repeated measure. In this analysis, the relation between AQ Total scores and performance on the F-CFMT + was *not* significant, $b = -0.01$, $p = 0.390$, 95% CI [− 0.04, 0.01]. Also, there were no two- or three-way interactions: AQ Total scores × task, $b = 0.00$, $p = 0.950$, 95% CI [− 0.02, 0.02], gender × AQ Total score × task, $b = -0.01$, $p = 0.577$, 95% CI [− 0.03, 0.02]. Therefore, when we control for gender, social skills as indexed by AQ Total score do not explain inconsistencies in performance across tasks.

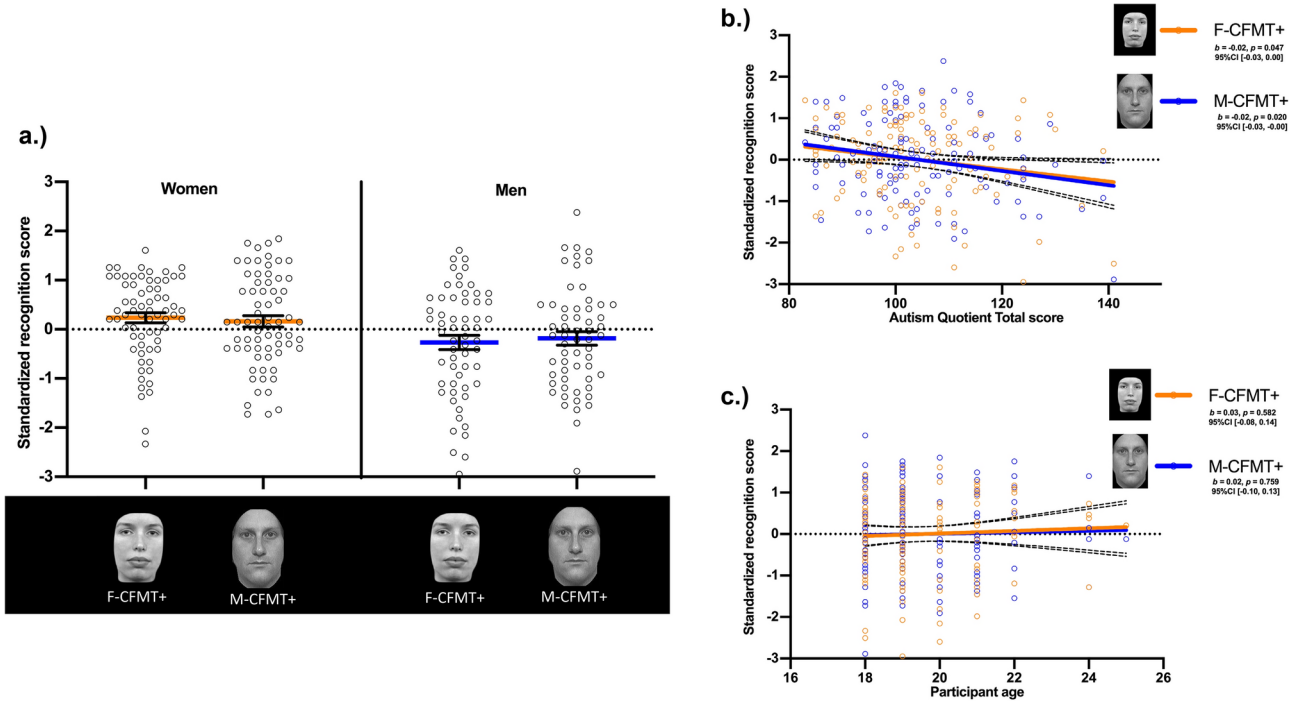

Fig. 4. Stability of individual differences in face recognition as a function of participant characteristics. (**a**) Recognition performance as a function of participant gender and stimulus gender. Standardized performance scores are plotted separately for women (orange) and men (blue). Error bars represent ± 1 standard error. Overall, women performed more accurately than men. Neither group evinced an own gender bias in face recognition behavior; both men and women performed equally well on both tasks. Participant gender did not explain instabilities in task performance. (**b**) Association between Autism Quotient (AQ) Total scores and face recognition performance. AQ Total scores are plotted on the x-axis. Standardized face recognition performance is plotted on the y-axis. Regression lines are plotted in blue (M-CFMT +) and orange (F-CFMT +) with the respective 95% confidence intervals. For both tasks, AQ Total scores were negatively related to face recognition performance. There was no difference in the slopes. AQ scores did not explain instabilities in task performance. (**c**) Relation between age and face recognition performance. Age (in years) is plotted on the x-axis. Standardized face recognition scores are plotted on the y-axis. Regression lines are plotted in blue (M-CFMT +) and orange (F-CFMT +) with the respective 95% confidence intervals. Age did not explain instabilities in task performance.

Age

Figure 4c illustrates the performance on each task as a function of participant age. There was no main effect of age, $b = 0.03$, $p = 0.582$, and no age x task interaction, $t(124) = -0.32$, $p = 0.749$. This indicates that among emerging adults (18–25 years), age does not explain instability in performance across these two tasks.

Discussion

The study of individual differences in face recognition behavior has skyrocketed in the last 15 years. The field is working to identify the scope of these differences, understand specific antecedents and sequalae, and use them to elucidate broader scientific principles in the study of human cognition[1]. However, there are three limitations in the existing work. First, many studies include participants with extreme behavior together with those in the more typical range of performance[11,14–16]. This approach assumes that there is stability in individual differences throughout this full range of scores, which is an open question. Second, most of the existing work evaluates inter-individual differences in face recognition abilities; further, work that does focus on intra-individual differences has largely focused on evaluating reliability of measurement tools (i.e., test–retest reliability, alternate-forms reliability). Much less is known about the *stability* (or lack thereof) of intra-individual differences across different sets of faces. Third, most of these studies employ the Cambridge Face Memory Test[20]. All the existing CFMT versions that have been used to evaluate continuities in face recognition only test recognition of *White, male, young adult* faces, which prevents a rigorous evaluation of potential gender biases in face recognition behavior.

Here, we started to address these gaps by investigating stability of intra-individual differences on two versions of the CFMT that differ in gender (but not race or age): the M-CFMT +[3] and the F-CFMT +[52]. Our central questions are fundamentally informed by principles of developmental science, which distinguish between *continuity* and *stability* in behavior over time. *Continuity* reflects the consistency or inconsistency (i.e., relative change) in a group mean-level characteristic over time[19]. By contrast, *stability* reflects the consistency in *relative ordering* of individuals on a characteristic over time[19]. We investigated stability of individual differences across

the full range of behavior and in the absence of extreme behavior. Finally, we investigated whether participant characteristics like gender, autism-like traits, and age influence stability in performance across tasks.

Stability of intra-individual differences in face recognition ability

To connect with the prior literature, we assessed the alternative forms reliability across the tasks. Even when the face stimuli vary by gender across the two tasks, there is strong reliability in performance. In other words, the tasks are similarly strong in their ability to test face recognition skills. This finding converges with other evaluations of alternative-forms reliability in the literature. However, our observed reliability estimate was much lower than the upper bound correlation, with ~ 49% of the variance left unexplained. These results led us to investigate stability in performance and whether it varies as a function of the extremity of the score.

Is there stability in behavior across the full spectrum of scores, or does it vary?

Stability in the full sample was of similar magnitude as the reliability estimate. However, we observed some interesting patterns in the scores. For example, there were 21 participants who exhibited differences in performance between tasks that approximated clinically relevant criteria (i.e., > 1 *SD*). These results motivated a more methodical analysis of stability between the tasks.

To investigate stability in a more fine-grained way, we tested stability within quartile ranges of scores for each task, across a range of scores within the quartiles, and in the absence of extreme scores. We observed two results. First, when the participants who exhibited stable behavior across the two tasks *in these extreme quartiles* were removed from the sample, the stability estimate was dramatically reduced (from 0.72 to 0.33). However, the stability among this group of extreme performers remained quite high (0.82).

Second, the ordinal regressions revealed that there was strong stability in scores in the highest and lowest quartiles of the tasks. Consider that the chance of a score aligning in a matched quartile across tasks is 25%. Scores from the F-CFMT + that fell in the lowest quartile were overwhelmingly likely (i.e., 2–3 times higher than chance, 55–90%) to align in the lowest quartile of the M-CFMT + as well. The same pattern was observed for scores from the M-CFMT + that fell in the lowest quartile. Similarly, scores from the highest quartiles of both tasks were also more likely (i.e., 46–89%) to align in the highest quartile for the corresponding task than any other quartile. However, stability was less characteristic of the behavior of individuals with the more typical range of performance (i.e., in the middle quartiles). For example, a score of -0.5 *SD* fell in the low medium quartile on the F-CFMT + had a predicted probability of 40% for the corresponding low medium quartile of the M-CFMT + , but also a 34% predicted probability of falling in the lowest quartile (i.e., higher than chance).

These findings suggest that stability in face recognition performance varies as a function of its position in the distribution of scores. The scores in the lowest and highest quartiles represent the most extreme behavior and are also most stable. The scores in the middle two quartiles fall in the more typical range of face recognition behavior and they are less stable. As a result, we suggest that considering the stability of behavior (i.e., rank ordering) across tasks and the full spectrum of inter-individual differences in performance (normal and abnormal ranges) is essential for determining the extent to which a behavior functions in a "trait-like" way. Specifically, the stability in behavior among poor recognizers might represent disorganized and perhaps inflexible face recognition systems given the levels of difficulty they experience. On the other hand, the relatively less stable behavior within the "normal" range might reflect face recognition systems that are dynamic and flexibly responsive to situational factors (e.g., kinds of faces to remember, social context of testing environment, level of anxiety experienced during testing).

Individual differences in person characteristics and stability in behavior

Given the relative instability of behavior in individual participants, we investigated whether multiple participant characteristics contribute to these instabilities in behavior across these face recognition tasks.

Participant and stimulus gender

There are mixed findings in the literature about the extent to which the gender of the observer and/or of the target faces influences recognition performance. Some work has argued for the presence of an own-gender bias (OGB), particularly for women[20,36,88–93]. Therefore, we investigated whether the gender of the observer or of the faces influenced the stability of performance. Importantly, both the M-CFMT + and the F-CFMT + train participants to recognize faces in the absence of hair or makeup, which can impact human face perception[94] and recognition abilities[95]. Hair and facial expressions are added in Block 4, which is particularly hard and therefore, good at identifying super recognizers. Finally, the psychometric properties of the two tasks are similar[52], which makes the evaluation of OGB here especially rigorous.

Women outperformed men across both tasks. This finding contributes to an inconsistent literature regarding the presence of observer gender differences in face recognition using the Cambridge tasks[39,93,96–101]. Critically, gender did not predict instabilities in performance across the two tasks. We found no evidence for the presence of an OGB in the face recognition performance of either men or women. While this result converges with existing findings using an earlier version of the F-CFMT +[39] and an early meta-analysis[37] of the literature, it does not replicate the finding of an OGB in women that has been reported previously[36]. In the current work, task-related factors on the tests using male and females faces are matched. This is evident in the similar levels of reliability, internal consistency, and convergent and divergent validity[52]. Also, the stimuli are well matched across the tasks. However, male *and* female participants performed better on the F-CFMT + than on the M-CFMT +, which we previously attributed mostly to performance in Block 4. Performance on Block 4 of the M-CFMT + is less reliable, and much more variable, than in the F-CFMT +[52]. Note that is a task design related issue, not instability in task performance related to gender. Also, we rigorously screened out participants who have a history of concussions and/or subclinical behaviors indicative of a potential psychiatric diagnosis. Concussions cause

widespread visual dysfunction[102] and multiple aspects of face perception are disrupted in every social-emotional disorder (e.g., anxiety, depression, bipolar, schizophrenia, autism). Previous studies reporting gender differences in face recognition behavior do not screen participants for these conditions (as far as we can tell). Therefore, some of the effects that have been reported previously as gender differences may instead reflect differences in health histories or task design.

Autism-like traits (ATLs)

We also investigated whether variations in socioemotional behavior, as indexed by ALTs, contribute to instability in face recognition performance. Although we observed a negative association between face recognition behavior and ALTs, which replicates previous findings using the CFMT[44–46], this effect was consistent across tasks. In other words, the magnitude of ALTs did not contribute to differences in stability of face recognition behavior.

Age

There is a large foundation of work that has reported age-related changes in face recognition abilities among youth and even older age groups using the CFMT[48–50,103]. In particular, previous research indicates that face recognition abilities increase throughout early adulthood, up to ~ 30[48,103]. Here, we evaluated whether age differences contributed to instabilities in face recognition performance. We did not observe an age-related change in performance overall or on either task (i.e., M-CFMT + , F-CFMT +) among the emerging adults (ages 18–25 years). Therefore, differences in age of participants did not magnify or reduce stability in performance among participants. It is important to note the distribution of age within the sample was positively skewed (18–19 years old). Therefore, our ability to detect increasing stability with age may be underpowered in this sample. This finding may also reflect the demographic characteristics of the stimuli. The faces to be recognized in these tasks are peer-aged individuals relative to the sample (emerging adults 18–25 years). Given findings of peer biases in face recognition (i.e., superior recognition of peer-aged compared to other-aged faces) among this age group in the literature[54,104], it is not surprising that we did not observe age-related changes in performance. In the future, evaluations of the role of age on the stability of performance might consider how both the age of the observer and the age of the faces to be recognized might influence behavior.

Limitations and future directions

Our study is not without limitations. Our sample consisted of a narrowly defined participant pool of largely White, cisgender, emerging adult, university students in the US. Although this sampling approach reduces "noise" and converges with existing research in individual differences of face recognition behavior, we think it is important for future research to investigate how participant characteristics (e.g., age, race/ethnicity, gender-identity) and social contexts (e.g., alone, with social partner) influence the stability of behavior. In addition, the age-range included in this sample is narrow within the emerging adult range. Therefore, more work is needed to thoroughly test potential age and/or race/ethnicity effects using these tasks.

Although these Cambridge tasks have become "standardized", we acknowledge that they are not optimized for recognition abilities of many people (especially children, older adults, potentially non-White, non-American individuals). Going forward, it is important to continue developing more Cambridge tests that include a broader range of face stimuli (i.e., gender, race, age, ethnicity) (e.g., CFMT-Aus, CFMT-C, CFMT-MY) especially those that include the most difficult Block 4 (only M-CFMT + and F-CFMT +). Testing participants with multiple tasks will allow for stronger conclusions about how these characteristics of face stimuli interact with participant characteristics to shape the stability (or lack thereof) of intra-individual differences in face recognition abilities. In addition, it is crucial to conduct these studies in samples that vary along the same parameters (age, gender, race/ethnicity).

It is also worth noting that we characterized extreme performance in this study by defining the quartile ranges of scores for each task. This allowed us to apply a sample-specific criterion to identifying extreme behavior (i.e., membership in both high quartiles, membership in both low quartiles). However, we acknowledge ongoing work[3,20,23,83–87] seeking to evaluate specific criteria to identify extreme behavior. Importantly, none of these criteria have been applied to the F-CFMT + and they are not typically applied to standardized scores on the CFMT. The quartile analysis allowed us to focus our investigation on the characteristics of this sample rather than engaging in ongoing debate about how best to identify these conditions. Therefore, we propose that this should be the focus of future work highlighting the importance of using multiple tests to identify extreme behavior[83].

Finally, it is important to consider that these conclusions may not generalize to characterize stability in familiar face recognition, which can be procedurally distinguished from unfamiliar face recognition[105].

Conclusions

We evaluated stability of intra-individual differences in face recognition abilities of emerging adults who are primarily White and live in the United States. We used two tests with the same task parameters but different kinds of face stimuli (i.e., White young male and female faces). Despite the similar task parameters and psychometric properties, stability in performance varies. Extreme performance, particularly deficient performance, is more stable across tasks. However, stability varies in the "normal" range (i.e., ± 1 *SD* of the mean) of performance. This is a bit surprising given the potential for generalization of performance across these two tasks (identical structure and similar stimuli). These findings are difficult to accommodate into current models of individual differences in face recognition behavior that do not currently account for potential patterns of bias (i.e., gender, race, developmental group) in face recognition behavior.

Given these findings, we suggest that there are important discontinuities in the stability of face recognition behavior between normal and abnormal behavior. The relatively more stable behavior of the abnormal range might reflect inflexible systems that are less responsive to situational factors and participant characteristics. In

contrast, relatively less stable performance within the normal range may be more flexible to the goals of social information processing systems[106]. This hypothesis aligns with previous findings that centers contextual factors, like social motivation, as predictors of face recognition ability[104,107]. Going forward, we suggest that future studies of face recognition research evaluate intra-individual differences within samples that vary on meaningful social dimensions (e.g., gender, race, developmental group) as they recognize multiple kinds of faces that vary in social relevance (e.g., peers, potential romantic partners, in-group, out-group) to test these hypotheses.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request. The F-CFMT + task is available for experimental purposes on Testable.org (tstbl.co/174-523) and via download on Databrary (Scherf, 2021; http://doi.org/https://doi.org/10.17910/b7.1396).

Code availability

The code used in this study is freely available on GitHub upon reasonable request to enable reproduction.

Received: 9 August 2024; Accepted: 12 February 2025
Published online: 19 March 2025

References

1. Wilmer, J. B. Individual differences in face recognition: A decade of discovery. *Curr. Dir. Psychol. Sci.* **26**(3), 225–230. https://doi.org/10.1177/0963721417710693 (2017).
2. Susilo, T. & Duchaine, B. Advances in developmental prosopagnosia research. *Curr. Opin. Neurobiol.* **23**(3), 423–429 (2013).
3. Russell, R., Duchaine, B. & Nakayama, K. Super-recognizers: People with extraordinary face recognition ability. *Psychon. Bull. Rev.* **16**(2), 252–257. https://doi.org/10.3758/PBR.16.2.252 (2009).
4. Wilmer, J. B. et al. Capturing specific abilities as a window into human individuality: The example of face recognition. *Cogn. Neuropsychol.* **29**(5–6), 360–392. https://doi.org/10.1080/02643294.2012.753433 (2012).
5. Wilmer, J. B. et al. Human face recognition ability is specific and highly heritable. *PNAS Proc. Natl. Acad. Sci. U. S. A.* **107**(11), 5238–5241. https://doi.org/10.1073/pnas.0913053107 (2010).
6. Shakeshaft, N. G. & Plomin, R. Genetic specificity of face recognition. *PNAS Proc. Natl. Acad. Sci. U. S. A.* **112**(41), 12887–12892. https://doi.org/10.1073/pnas.1421881112 (2015).
7. Li, J. et al. Extraversion predicts individual differences in face recognition. *Commun. Integr. Biol.* **3**, 295–298 (2010).
8. Lander, K. & Poyarekar, S. Famous face recognition, face matching, and extraversion. *Q. J. Exp. Psychol.* **68**(9), 1769–1776. https://doi.org/10.1080/17470218.2014.988737 (2015).
9. Balas, B. & Saville, A. N170 face specificity and face memory depend on hometown size. *Neuropsychologia* **69**, 211–217. https://doi.org/10.1016/j.neuropsychologia.2015.02.005 (2015).
10. Balas, B. & Saville, A. Hometown size affects the processing of naturalistic face variability. *Vision. Res.* **141**, 228–236. https://doi.org/10.1016/j.visres.2016.12.005 (2017).
11. Berger, A., Fry, R., Bobak, A. K., Juliano, A. & DeGutis, J. Distinct abilities associated with matching same identity faces versus discriminating different faces: Evidence from individual differences in prosopagnosics and controls. *Q. J. Exp. Psychol.* **75**(12), 2256–2271. https://doi.org/10.1177/17470218221076817 (2022).
12. Faja, S. et al. The effects of face expertise training on the behavioral performance and brain activity of adults with high functioning autism spectrum disorders. *J. Autism Dev. Disord.* **42**(2), 278–293. https://doi.org/10.1007/s10803-011-12438 (2012).
13. Tanaka, J. W. et al. Using computerized games to teach face recognition skills to children with autism spectrum disorder: the Let's Face It! program. *J. Child Psychol. Psychiatry* **51**(8), 944–952. https://doi.org/10.1111/j.14697610.2010.02258.x (2010).
14. Shah, P., Gaule, A., Sowden, S., Bird, G. & Cook, R. The 20-item prosopagnosia index (PI20): a self-report instrument for identifying developmental prosopagnosia. *R. Soc. Open Sci.* **2**(6), 140343. https://doi.org/10.1098/rsos.140343 (2015).
15. Noyes, E., Davis, J. P., Petrov, N., Gray, K. L. & Ritchie, K. L. The effect of face masks and sunglasses on identity and expression recognition with super-recognizers and typical observers. *R. Soc. Open Sci.* **8**(3), 201169 (2021).
16. Tardif, J. et al. Use of face information varies systematically from developmental prosopagnosics to super-recognizers. *Psychol. Sci.* **30**(2), 300–308. https://doi.org/10.1177/0956797618811338 (2019).
17. Rutter, M. Epidemiological approaches to developmental psychopathology. *Arch. Gen. Psychiatry* **45**(5), 486–495 (1988).
18. Wang, X. et al. Behavioral and neural correlates of social network size: The unique and common contributions of face recognition and extraversion. *J. Pers.* **90**(2), 294–305. https://doi.org/10.1111/jopy.12666 (2022).
19. Bornstein, M. H., Putnick, D. L. & Esposito, G. Continuity and stability in development. *Child Dev. Perspect.* **11**(2), 113–119. https://doi.org/10.1111/cdep.12221 (2017).
20. Duchaine, B. & Nakayama, K. 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. https://doi.org/10.1016/j.neuropsychologia.2005.07.001 (2006).
21. Behrmann, M., Avidan, G., Marotta, J. J. & Kimchi, R. Detailed exploration of face-related processing in congenital prosopagnosia: 1. behavioral findings. *J. Cogn. Neurosci.* **17**(7), 1130–1149. https://doi.org/10.1162/0898929054475154 (2005).
22. Murray, E. & Bate, S. Diagnosing developmental prosopagnosia: repeat assessment using the Cambridge Face Memory Test. *R. Soc. Open Sci.* **7**(9), 200884 (2020).
23. Bate, S. et al. The limits of super recognition: An other-ethnicity effect in individuals with extraordinary face recognition skills. *J. Exp. Psychol. Hum. Percept. Perform.* **45**(3), 363–377. https://doi.org/10.1037/xhp0000607 (2019).
24. Fysh, M. C., Stacchi, L. & Ramon, M. Differences between and within individuals, and subprocesses of face cognition: Implications for theory, research and personnel selection. *R. Soc. Open Sci.* **7**(9), 200233 (2020).
25. McKone, E. et al. Face ethnicity and measurement reliability affect face recognition performance in developmental prosopagnosia: Evidence from the cambridge face memory Test-Australian. *Cogn. Neuropsychol.* **28**(2), 109–146. https://doi.org/10.1080/02643294.2011.616880 (2011).
26. McKone, E. et al. A robust method of measuring other-race and other-ethnicity effects: The Cambridge Face Memory Test format. *PLoS One* **7**(10), e47956 (2012).
27. Kho, S. K., Leong, B. Q. Z., Keeble, D. R., Wong, H. K. & Estudillo, A. J. A new Asian version of the CFMT: The Cambridge Face Memory Test-Chinese Malaysian (CFMT-MY). *Behav. Res. Methods* **56**(3), 1192–1206 (2024).
28. Wan, L. et al. Face-blind for other-race faces: Individual differences in other-race recognition impairments. *J. Exp. Psychol. Gen.* **146**(1), 102–122. https://doi.org/10.1037/xge0000249 (2017).
29. Petersen, L. A. & Leue, A. Face memory and face matching: Internal consistency and test-retest reliability for the CFMT+ and the GFMT-S. *J. Individ. Differ.* **43**(3), 152–159. https://doi.org/10.1027/1614-0001/a000361 (2022).

30. Stantić, M. et al. The Oxford Face Matching Test: A non-biased test of the full range of individual differences in face perception. *Behav. Res. Methods* **54**(1), 158–173 (2022).
31. Robertson, D. J., Black, J., Chamberlain, B., Megreya, A. M. & Davis, J. P. Super-recognisers show an advantage for other race face identification. *Appl. Cogn. Psychol.* **34**(1), 205–216. https://doi.org/10.1002/acp.3608 (2020).
32. DeGutis, J., Mercado, R. J., Wilmer, J. & Rosenblatt, A. Individual differences in holistic processing predict the own-race advantage in recognition memory. *PLoS One* **8**(4), e58253 (2013).
33. Vaz, S., Falkmer, T., Passmore, A. E., Parsons, R. & Andreou, P. The case for using the repeatability coefficient when calculating test–retest reliability. *PLoS One* **8**(9), e73990 (2013).
34. Cenac, Z., Biotti, F., Gray, K. L. & Cook, R. Does developmental prosopagnosia impair identification of other-ethnicity faces?. *Cortex* **119**, 12–19. https://doi.org/10.1016/j.cortex.2019.04.007 (2019).
35. Bate, S., Bennetts, R., Murray, E. & Portch, E. Enhanced matching of children's faces in "super-recognisers" but not high-contact controls. *i-Perception* https://doi.org/10.1177/2041669520944420 (2020).
36. Herlitz, A. & Lovén, J. Sex differences and the own-gender bias in face recognition: A meta-analytic review. *Visual Cogn.* **21**(9–10), 1306–1336. https://doi.org/10.1080/13506285.2013.823140 (2013).
37. Shapiro, P. N. & Penrod, S. Meta-analysis of facial identification studies. *Psychol. Bull.* **100**(2), 139 (1986).
38. Wright, D. B. & Sladden, B. An own gender bias and the importance of hair in face recognition. *Acta Psychol.* **114**(1), 101–114 (2003).
39. Scherf, K. S., Elbich, D. B., & Motta-Mena, N. V. Investigating the influence of biological sex on the behavioral and neural basis of face recognition. *Eneuro*. **4**(3) (2017).
40. Bate, S., Parris, B., Haslam, C. & Kay, J. Socio-emotional functioning and face recognition ability in the normal population. *Pers. Individ. Differ.* **48**(2), 239–242. https://doi.org/10.1016/j.paid.2009.10.005 (2010).
41. Turano, M. T. & Viggiano, M. P. The relationship between face recognition ability and socioemotional functioning throughout adulthood. *Aging Neuropsychol. Cogn.* **24**(6), 613–630. https://doi.org/10.1080/13825585.2016.1244247 (2017).
42. Engfors, L. M. et al. Face recognition's practical relevance: Social bonds, not social butterflies. *Cognition* **250**, 105816 (2024).
43. Giacomin, M., Brinton, C. & Rule, N. O. Narcissistic individuals exhibit poor recognition memory. *J. Pers.* **90**(5), 675–689 (2022).
44. Lewis, G. J., Shakeshaft, N. G. & Plomin, R. Face identity recognition and the social difficulties component of the autism-like phenotype: Evidence for phenotypic and genetic links. *J. Autism Dev. Disord.* **48**(8), 2758–2765. https://doi.org/10.1007/s10803-018-3539-4 (2018).
45. Rhodes, G., Jeffery, L., Taylor, L. & Ewing, L. Autistic traits are linked to reduced adaptive coding of face identity and selectively poorer face recognition in men but not women. *Neuropsychologia* **51**(13), 2702–2708. https://doi.org/10.1016/j.neuropsychologia.2013.08.016 (2013).
46. Hedley, D., Brewer, N. & Young, R. Face recognition performance of individuals with asperger syndrome on the Cambridge face memory test. *Autism Res.* **4**(6), 449–455. https://doi.org/10.1002/aur.214 (2011).
47. Tanaka, J., Halliday, D., MacDonald, S. & Scherf, S. A reciprocal model of face recognition and the autism condition: Evidence from an individual differences perspective. *J. Vis.* **14**(10), 1443–1443 (2014).
48. Germine, L. T., Duchaine, B. & Nakayama, K. Where cognitive development and aging meet: Face learning ability peaks after age 30. *Cognition* **118**(2), 201–210. https://doi.org/10.1016/j.cognition.2010.11.002 (2011).
49. Susilo, T., Germine, L. & Duchaine, B. Face recognition ability matures late: evidence from individual differences in young adults. *J. Exp. Psychol. Hum. Percept. Perform.* **39**(5), 1212 (2013).
50. Stantić, M., Hearne, B., Catmur, C. & Bird, G. Use of the oxford face matching test reveals an effect of ageing on face perception but not face memory. *Cortex. J. Devot. Study Nervous Syst. Behav.* **145**, 226–235. https://doi.org/10.1016/j.cortex.2021.08.016 (2021).
51. Arnett, J. J. Emerging adulthood: A theory of development from the late teens through the twenties. *Am. Psychol.* **55**(5), 469–480. https://doi.org/10.1037/0003-066X.55.5.469 (2000).
52. Arrington, M., Elbich, D., Dai, J., Duchaine, B. & Scherf, K. S. Introducing the female Cambridge face memory test–long form (F-CFMT+). *Behav. Res. Methods* **54**(6), 3071–3084 (2022).
53. Ramon, M., Bobak, A. K. & White, D. Towards a 'manifesto' for super recognizer research. *Br. J. Psychol.* **110**(3), 495–498. https://doi.org/10.1111/bjop.12411 (2019).
54. Dai, J., & Scherf, K. S. The privileged status of peer faces: subordinate-level neural representations of faces in emerging adults. *J. Cogn. Neurosci.* 1–21 (2023).
55. Gadow, K. D., Sprafkin, J., & Weiss, M. *Adult Self-Report Inventory-4 Manual*. (Checkmate Plus, 2004).
56. Baron-Cohen, S., Wheelwright, S., Skinner, M., Martin, J. & Clubley, E. "The autism-spectrum quotient (QA): Evidence from asperger syndrome/high functioning autism, males and females, scientists and mathematicians": Errata. *J. Autism Dev. Disord.* **31**(6), 603. https://doi.org/10.1023/A:1017455213300 (2001).
57. Griffin, J. W., Bauer, R. & Scherf, K. S. A quantitative meta-analysis of face recognition deficits in autism: 40 years of research. *Psychol. Bull.* **147**(3), 268292. https://doi.org/10.1037/bul0000310 (2021).
58. Bortolon, C., Capdevielle, D. & Raffard, S. Face recognition in schizophrenia disorder: A comprehensive review of behavioral, neuroimaging and neurophysiological studies. *Neurosci. Biobehav. Rev.* **53**, 79–107 (2015).
59. Snedecor, G. W., & Cochran, W. G. Factorial experiments. *Stat. Methods*. **7** (1980).
60. Rosenthal, R., & Rosnow, R. L. *Essentials of Behavioral Research: Methods and Data Analysis* (2008).
61. Hoekstra, R. A. et al. The construction and validation of an abridged version of the autism-spectrum quotient (AQ-short). *J. Autism Dev. Disord.* **41**(5), 589–596. https://doi.org/10.1007/s10803-010-1073-0 (2011).
62. Elbich, D. B. & Scherf, S. Beyond the FFA: Brain-behavior correspondences in face recognition abilities. *NeuroImage* **147**, 409–422. https://doi.org/10.1016/j.neuroimage.2016.12.042 (2017).
63. R Core Team. *R: A Language and Environment for Statistical Computing*. 54–64 (R Foundation for Statistical Computing, 2019). https://www.R-project.org/.
64. RStudio Team. RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. http://www.rstudio.com/ (2020).
65. Wickham et al. Welcome to the tidyverse. *J. Open Source Softw.* **4**(43), 1686. https://doi.org/10.21105/joss.01686 (2019).
66. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**(1), 1–48 (2015).
67. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest Package: Tests in linear mixed effects models. *J. Stat. Softw.* **82**(13), 1–26 (2017).
68. Revelle, W. psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA. https://CRAN.Rproject.org/package=psych Version = 2.0.12 (2020).
69. Kassambara, A. rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R package version 0.7.0. https://CRAN.R-project.org/package=rstatix (2021).
70. Long, J.A. interactions: Comprehensive, User-Friendly Toolkit for Probing Interactions_ R package version 1.1.0. https://cran.rproject.org/package=interactions (2019).
71. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S-PLUS* 4th edn. (Springer, 2002).
72. Fox, J., & Weisberg, S. *An {R} Companion to Applied Regression*, 3rd edn. https://socialsciences.mcmaster.ca/jfox/Books/Companion/ (Sage, 2019).
73. Kassambara, A. _ggpubr: 'ggplot2' Based Publication Ready Plots_. R package version 0.6.0. https://CRAN.R-project.org/package=ggpubr (2023).

74. Lenth, R. _emmeans: Estimated Marginal Means, aka Least-Squares Means_. R package version 1.10.4. https://CRAN.R-project.org/package=emmeans (2024).
75. Ben-Shachar, M., Lüdecke, D. & Makowski, D. effectsize: Estimation of effect size indices and standardized parameters. *J. Open Source Softw.* **5**(56), 2815. https://doi.org/10.21105/joss.02815 (2020).
76. Christensen, R. _ordinal-Regression Models for Ordinal Data_. R package version 2023.12-4.1. https://CRAN.R-project.org/package=ordinal (2023).
77. Selker, R., Love, J., Dropmann, D. _jmv: The 'jamovi' Analyses_. R package version 2.5.6. https://CRAN.R-project.org/package=jmv (2024).
78. Bonett, D. _statpsych: Statistical Methods for Psychologists_. R package version 1.6.0. https://CRAN.R-project.org/package=statpsych (2024).
79. Meissner, C. A. & Brigham, J. C. Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychol. Public Policy Law* **7**(1), 3 (2001).
80. Dai, J., Griffin, J. W. & Scherf, K. S. How is race perceived during adolescence? A meta-analysis of the own-race bias. *Dev. Psychol.* **60**(4), 649–664. https://doi.org/10.1037/dev0001721 (2024).
81. Corrow, S. L., Albonico, A. & Barton, J. J. Diagnosing prosopagnosia: The utility of visual noise in the Cambridge Face Recognition Test. *Perception.* **47**(3), 330–343 (2018).
82. Spearman, C. The proof and measurement of association between two things. *Am. J. Psychol.* **15**, 72–101. https://doi.org/10.2307/1412159 (1904).
83. Ramon, M. Super-recognizers—A novel diagnostic framework, 70 cases, and guidelines for future work. *Neuropsychologia* **158**, 11. https://doi.org/10.1016/j.neuropsychologia.2021.107809 (2021).
84. Bate, S., Dalrymple, K. & Bennetts, R. J. Face recognition improvements in adults and children with face recognition difficulties. *Brain Commun.* **4**(2), fcac068 (2022).
85. DeGutis, J. et al. What is the prevalence of developmental prosopagnosia? An empirical assessment of different diagnostic cutoffs. *Cortex. J. Devot. Study Nervous Syst. Behav.* **161**, 51–64. https://doi.org/10.1016/j.cortex.2022.12.014 (2023).
86. Bate, S. et al. Applied screening tests for the detection of superior face recognition. *Cogn. Res. Princ. Implic.* **3**, 1–19 (2018).
87. Bobak, A. K., Pampoulov, P. & Bate, S. Detecting superior face recognition skills in a large sample of young british adults. *Front. Psychol.* **7**, 11 (2016).
88. Rehnman, J. & Herlitz, A. Higher face recognition ability in girls: Magnified by own-sex and own-ethnicity bias. *Memory* **14**(3), 289–296 (2006).
89. Rehnman, J. & Herlitz, A. Women remember more faces than men do. *Acta Psychol.* **124**(3), 344–355 (2007).
90. Hills, P. J., Pake, J. M., Dempsey, J. R. & Lewis, M. B. Exploring the contribution of motivation and experience in the postpubescent own-gender bias in face recognition. *J. Exp. Psychol.* **44**(9), 1426 (2018).
91. Morgan, M. & Hills, P. J. Correlations between holistic processing, autism quotient, extraversion, and experience and the own-gender bias in face recognition. *PLoS One* **14**(7), 19 (2019).
92. Mukudi, P. B. L. & Hills, P. J. The combined influence of the own-age, -gender, and -ethnicity biases on face recognition. *Acta Psychol.* **194**, 1–6 (2019).
93. Sunday, M. A., Patel, P. A., Dodd, M. D. & Gauthier, I. Gender and hometown population density interact to predict face recognition ability. *Vis. Res.* **163**, 1423 (2019).
94. Jones, A. L., Porcheron, A. & Russell, R. Makeup changes the apparent size of facial features. *Psychol. Aesthet. Creat. Arts* **12**(3), 359 (2018).
95. Toseeb, U., Keeble, D. R. & Bryant, E. J. The significance of hair for face recognition. *PLoS One* **7**(3), e34144 (2012).
96. Petersen, L. A. & Leue, A. Extraordinary face recognition performance in laboratory and online testing. *Appl. Cogn. Psychol.* **35**(3), 579–589 (2021).
97. Tagliente, S., Passarelli, M., D'Elia, V., Palmisano, A., Dunn, J. D., Masini, M., et al. Self-reported face recognition abilities moderately predict facelearning skills: Evidence from Italian samples. *Heliyon*. **9**(3) (2023).
98. Boutet, I. & Meinhardt-Injac, B. Measurement of individual differences in faceidentity processing abilities in older adults. *Cogn. Res. Princ. Implic.* **6**(1), 1–11 (2021).
99. Lowes, J., Hancock, P. J., & Bobak, A. K. Balanced integration score: a new way of classifying developmental prosopagnosia. https://doi.org/10.31234/osf.io/g85k7 (2023).
100. Østergaard Knudsen, C., Winther Rasmussen, K. & Gerlach, C. Gender differences in face recognition: The role of holistic processing. *Vis. Cogn.* **29**(6), 379–385. https://doi.org/10.1080/13506285.2021.1930312 (2021).
101. Turbett, K., Palermo, R., Bell, J., Burton, J. & Jeffery, L. Individual differences in serial dependence of facial identity are associated with face recognition abilities. *Sci. Rep.* **9**(1), 1–12 (2019).
102. Barnett, B. P. & Singman, E. L. Vision concerns after mild traumatic brain injury. *Curr. Treat. Opt. Neurol.* **17**, 1–14 (2015).
103. DeGutis, J. et al. The rise and fall of face recognition awareness across the life span. *J. Exp. Psychol. Hum. Percept. Perform.* **49**(1), 22–33. https://doi.org/10.1037/xhp0001069 (2023).
104. Picci, G. & Scherf, K. S. From caregivers to peers: Puberty shapes human face perception. *Psychol. Sci.* **27**(11), 1461–1473 (2016).
105. Gobbini, M. I. & Haxby, J. V. Neural systems for recognition of familiar faces. *Neuropsychologia* **45**(1), 32–41 (2007).
106. Scherf, K. S. & Scott, L. S. Connecting developmental trajectories: Biases in face processing from infancy to adulthood. *Dev. Psychobiol.* **54**(6), 643663 (2012).
107. Van Bavel, J. J., Swencionis, J. K., O'Connor, R. C. & Cunningham, W. A. Motivated social memory: Belonging needs moderate the own-group bias in face recognition. *J. Exp. Soc. Psychol.* **48**(3), 707–713. https://doi.org/10.1016/j.jesp.2012.01.006 (2012).

Acknowledgements

The research reported in this paper was supported by supported by the Department of Psychology and the Social Science Research Institute at Pennsylvania State University. Content is the responsibility of the authors and does not represent the views of the SSRI. M.A. conducted this work as a graduate student at Pennsylvania State University and is currently affiliated with the Center for Mind and Brain at the University of California, Davis, as a postdoctoral scholar.

Author contributions

K.S. conceptualized the study and designed the methodology. K.S. and M.A. collected the data from participants. M.A. performed the data analyses. K.S. and M.A. wrote the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Inclusion and 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

Informed consent to be a part of the study was obtained from all individual participants included in the study. No identifiable participant information has been included in this manuscript. All actors in stimuli images and videos signed an informed consent to use the photos.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-90317-4.

Correspondence and requests for materials should be addressed to K.S.S.

Reprints and permissions information is available at www.nature.com/reprints.

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