Gender Differences in Recognition of Toy Faces Suggest a Contribution of Experience

By

Kaitlin F. Ryan

Thesis
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE

in

Psychology

August, 2016

Nashville, Tennessee

Approved:

Isabel Gauthier, Ph.D.

Geoffrey Woodman, Ph.D.

TABLE OF CONTENTS

Page

LIST OF TABLES iii

LIST OF FIGURES iv

Introduction 1
Methods 4
Participants 4
Materials and procedures 6
VFET 6
VET-car 8
Procedure 8
Results 9
Gender effects 12
Self-reported experience 12
Performance 14
Correlations between self-reported experience across categories 21
Correlations between performance across categories 21
Correlations between self-reported experience and performance within category 22
Discussion 22
Conclusion 27

REFERENCES 28

LIST OF TABLES

Table Page

1. Summary statistics for all participants, separated by sex and source 5

2. Means, standard deviations, and Cronbach's α on performance accuracy and self-reported experience (all participants).....10

3. Means, standard deviations, and Cronbach's α on performance accuracy and self-reported experience (by gender and source) 11

4. Between- and within- category correlations for self-reported experience and performance (all participants).....18

5. Between- and within- category correlations for self-reported experience and performance (male participants) 19

6. Between- and within- category correlations for self-reported experience and performance (female participants) 20

LIST OF FIGURES

Figure Page

1. Examples of the 3 x 2 study array.....7

2. Self-reported experience rating for each test category separated by gender 13

3. Performance for each test category separated by gender.....15

Introduction

Many biases in face recognition have been reported: people are more skilled at recognizing faces within their own race (Lindsay, et al 1991), age (Yovel, et al 2012), and species (Pascalis, de Haan, & Nelson, 2002). These performance differences for face recognition are typically attributed to underlying differences in experience (Gauthier et al., 2014; Yovel, et al, 2012). Gender differences have also been reported, however their interpretation is less clear. Women outperform men on several face recognition tasks (Goldstein & Chance, 1971; Lewin & Herlitz, 2002; Lovén, et al., 2014; Rehnman & Herlitz, 2007). While there is not always a difference in performance between genders (Duchaine & Nakayama, 2006), interestingly, no advantage for men with male faces has been reported.

There are many possible explanations for the advantage that women show on face recognition tasks. This includes differences in memory and social-cognitive skills rooted in differential brain connectivity between genders (Ingalhalikar et al., 2014), differences in gaze preferences that present from a very young age (Baron-Cohen, 2002), gender differences in social interaction, including a greater concern with the attractiveness of other women in female observers, and other socially-motivated goals (Goldstein & Chance, 1970; Lewin & Herlitz, 2002; Sawada et al., 2014; Yovel et al., 2012). Evolutionary pressures or cultural influences could lead to greater efficiency in encoding, remembering, and labeling faces for women (Wolff et al., 2014; Lovén et al., 2012) – but importantly, these explanations pertain to face recognition as a single, unitary domain.

However, if differential experience is an important driver of gender differences in face recognition, it may be possible to find categories of faces for which men outperform women. The goal of this study is not to identify the specific causes of gender biases in face recognition.

Instead, we seek evidence for a crossover interaction between observer gender and face categories. Finding such a pattern would rule out any account of gender differences at the level of the entire face domain, thereby constraining theoretical explanations. An analogy in studies of object recognition illustrates the benefits of such an interaction. One study described a male advantage on a test of car recognition and suggested that this could be due to better mental rotation in men then women (Dennett et al, 2012). Even though the result was obtained with cars, the authors considered an explanation that would apply more broadly to all object categories. It is not rare that authors consider object recognition as a unitary skill, such that performance for one category is considered representative. This means that the explanation offered for a male advantage in car recognition becomes a hypothesis for a domain-general advantage where men should outperform women for any object category. However, performance with more object categories was measured in later work, women outperformed men with some of these categories (McGugin et al., 2012). A crossover interaction reveals that gender effects for objects are domain-specific, likely influenced by experience with various categories, rather than requiring one broad domain-general explanation.

Here, we set out to create reliable tests to measure individual differences in face recognition across different types of faces (in particular, toy faces) for which men and women may differ in experience. We created a task similar to the Vanderbilt Expertise Test (VET; McGugin et al., 2012) and the CFMT (Duchaine & Nakayama, 2006), in which participants learn a set of identities across four face categories (Caucasian female faces, Caucasian male faces, Barbie doll faces, and Transformer action figure faces) and later recognize these identities among distractor faces.

Foreshadowing our results, we found a cross-over interaction in performance for men vs. women with Barbie vs. Transformer faces - most significantly, we found that Transformer faces are one category of faces for which men outperform women.

Methods

The tests we describe are a subset (Caucasian female, Caucasian male, Barbie, Transformer) of a new battery of face tests, the Vanderbilt Face Expertise Test (VFET; Ryan & Gauthier, 2014). We also included the VET car sub-test (VET-car; McGugin et al., 2012) as a measure of non-face recognition performance where we expect men to outperform women based on previous findings.

Participants

297 participants completed the VFET face tasks and the VET-car. Participants completed the tasks either in the lab or via Amazon Mechanical Turk (AMT). AMT is an online crowdsourcing platform that has been used to conduct psychological studies and produces results comparable to those in the laboratory with more diverse samples (see Crump et al., 2013; Cho et al., 2015; Richler et al., 2014). All participants were compensated with course credit or a small payment. Self-reported age, gender, and ethnicity are reported in Table 1. Participants provided informed consent per the guidelines of the Institutional Review Board of Vanderbilt University. All research was carried out in compliance with the Code of Ethics of the World Medical Association (Declaration of Helsinki).

Table 1. Summary statistics for all participants, separated by sex and source. Parentheses in age indicate standard deviation.

	All Participants	Lab participants	AMT participants
N	295	179	116
Age	27(10.5)	23(6.9)	33(11.8)
Men	161	94	67
Women	134	85	49
Caucasian	196	109	87
Black	43	33	10
Asian	41	31	10
Hispanic	7	1	6
Other	8	5	3

Materials and procedures

All participants (either in the lab or via AMT) completed the VFET face tasks (Caucasian female, Caucasian male, Barbie, Transformer) and the VET-car task.

VFET. We collected images of human faces from professional photographers and used them with their permission. We collected toy faces from online collections.

For each category we collected 6 different target identities, each with 7 exemplar images (see Figure 1). This included sets for 6 male and 6 female individuals (84 images of human faces), 6 different Barbie dolls and 6 different Transformer figures (42 images of Barbie doll faces and 42 images of Transformer faces). Barbie dolls and "Barbie-friend" dolls have been released with several different face sculpts since 1959. A large number of Transformer toys have been released since 1984. Therefore, it was possible for us to collect groups of images that corresponded to unique toy identities, similar those of human faces. Exemplar images, while corresponding to a single identity, could vary across many dimensions such as expression, pose, background context, camera view, and lighting condition. For each category we also collected 102 distractor images of other identities, separate from the 6 target identities. During the tasks, each trial contained 1 exemplar image of a learned target identity and 2 previously unseen distractor images of the same face category.

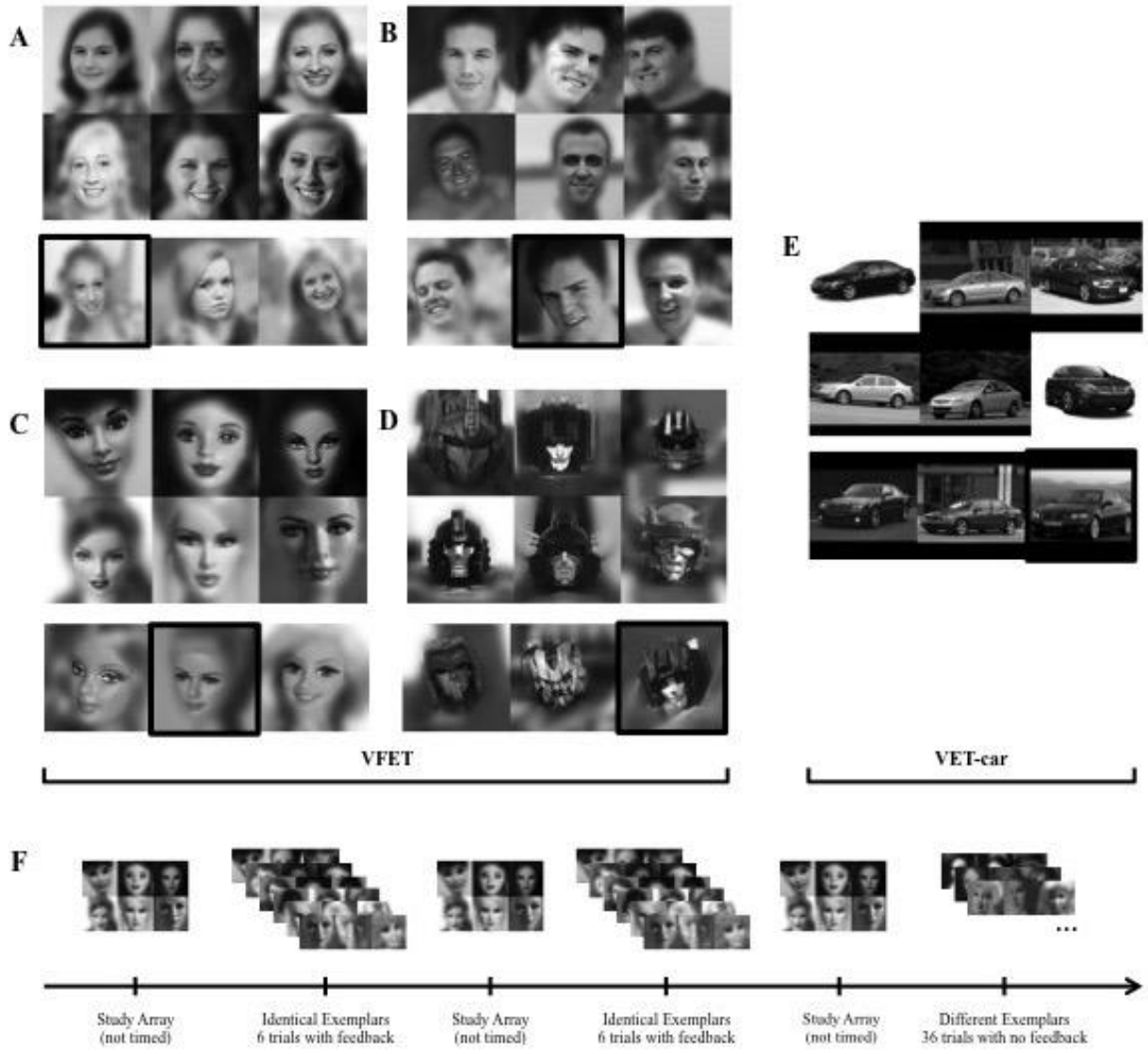

Figure 1 Examples of the 3 x 2 study array of the 6 identities for A) Caucasian adult females B) Barbie dolls C) Caucasian adult males D) Transformer action figures, and E) cars. For each category a "transfer" exemplar trial example is shown below the 6 targets array. Black boxes indicate the correct response, corresponding to a target identity in the study array. Study structure (F) shows an example of the blocked task using Barbie faces.

All images were shown in greyscale and 200 x 200 pixels. We chose to keep some level of background noise in our images, which generally works better with non-posed photos that vary in viewpoint, and is consistent with a more naturalistic context for face recognition and. We chose matching targets and distractors to minimize the diagnosticity of these cues and applied a Gaussian filter to blur backgrounds and further reduce the information from non-face features (e.g. hair, background cues).

VET-car. Participants also completed a modified version of the VET-car (McGugin et al., 2012; catch trials were added to the original test). Car images consisted of 1997-2003 sedan models commercially available in the United States.

Procedure. Participants first reported their perceived level of experience with each face category before being tested. Participants were asked to rate their experience while considering their "interest in, years of exposure to, knowledge of, and familiarity with each category compared to other individuals." Participants responded using a Likert scale with 1 meaning "very much below average" and 9 meaning "very much above average" (see Gauthier et al., 2014 and VanGulick et al., 2015, for validation of this measure). One limitation is that participants may vary in the reference group they choose to use.

Next, participants completed the VFET face tasks (Barbie, Caucasian female, Caucasian male, Transformer faces) the VET-car (see Figure 1). Each task was blocked by category. For each category, participants viewed a 3 x 2 array of the 6 target identities to study for as long as they needed. Participants then completed 6 "identical" trials with triplets consisting of one exact image of a face identity (or car model) they had previously studied and 2 distractor images of novel identities (or models). The target identity could occur in any of the 3 triplet positions and

participants indicated whether the target occurred on the left, middle, or right position. Feedback showed the correct location of the target identity. The study array was presented again, and participants completed another 6 "identical" trials (for a total of 12 "identical" trials).

After viewing the study array a third time, participants completed another 39 trials (36 "transfer" trials plus 3 "catch" trials): These showed a new exemplar of a target identity (under different viewing conditions (e.g., position of the face, lighting) and 2 distractor images. No feedback was given on these trials. The "catch" trials for each category were included to test for understanding of the task, and a constant minimal level of attention and motivation throughout the test. Catch trials showed distractors extremely different from the target (e.g. a Barbie doll target face with two stuffed animal distractors).

Results

We excluded data from 2 participants with a human face score (Caucasian male and female aggregate score) below chance (33%), leaving 295 participants for subsequent analysis.

For each test, we report Cronbach's alpha, a reliability measure of internal consistency (see Table 2). Reliability was good for each category across all participants ($\alpha > .8$) and also for each category when assessed by gender or source (lab- or AMT- collected data; $\alpha > .7$; Table 3).

Table 2 Means, standard deviations, and Cronbach's α on performance accuracy and self-reported experience for all categories across all participants

	All Participants				
	Accuracy (/48)			Self-Report (1-9)	
	Mean	SD	α	Mean	SD
C. Female	39.18	6.30	0.87	5.99	1.71
C. Male	40.34	6.16	0.89	5.86	1.64
Human	39.77	5.80	0.90	5.92	1.54
Barbie	33.46	6.27	0.82	3.57	2.01
Transformer	38.28	6.44	0.87	3.54	2.14
Car	27.42	7.94	0.85	4.28	2.02

Table 3 Means, standard deviations, and Cronbach's α on performance accuracy and self-reported experience for all categories separated by gender and source

By Gender

	Men					Women				
	Accuracy (/48)			Self-Report (1-9)		Accuracy (/48)			Self-Report (1-9)	
	Mean	SD	α	Mean	SD	Mean	SD	α	Mean	SD
C. Female	39.05	6.07	0.70	5.94	1.63	39.33	6.58	0.77	6.06	1.80
C. Male	40.32	6.23	0.88	5.96	1.65	40.42	6.09	0.81	5.73	1.63
Human	39.69	7.78	0.90	5.95	1.46	39.88	5.85	0.89	5.90	1.64
Barbie	32.29	6.03	0.88	2.48	1.50	34.86	6.28	0.91	4.87	1.77
Transformer	39.12	6.26	0.87	4.22	2.15	37.28	6.53	0.83	2.73	1.82
Car	28.24	8.47	0.89	4.82	2.02	26.44	7.15	0.79	3.63	1.81

By Source

	Lab					AMT				
	Accuracy			Self-Report		Accuracy			Self-Report	
	Mean	SD	α	Mean	SD	Mean	SD	α	Mean	SD
C. Female	39.28	5.94	0.85	6.05	1.92	39.01	6.84	0.88	5.91	1.34
C. Male	40.60	6.68	0.86	6.04	1.71	40.00	6.83	0.90	5.57	1.49
Human	39.94	5.41	0.92	6.05	1.72	39.50	6.38	0.87	5.74	1.20
Barbie	34.13	5.27	0.76	3.59	2.10	32.41	7.46	0.91	3.53	1.89
Transformer	37.90	6.28	0.86	3.43	2.30	38.87	6.66	0.83	3.72	1.86
Car	29.13	7.95	0.85	4.22	2.23	24.78	7.19	0.79	4.36	1.63

Across all participants, performance on Caucasian male and Caucasian female faces was highly correlated (r = .74, p < .001), especially since the maximum correlation possible given these measurements' reliabilities was r = .84. Additionally, there was no significant gender difference in performance with human faces (Caucasian Female faces: t_{274}= .38, p = .65; Caucasian Male faces: t_{285}= .13, p = .55). Because our predictions were mainly focused on gender differences in recognition performance for toy faces (Barbies and Transformers), we combined the Caucasian female and Caucasian male scores into an aggregate human face score for the remaining analyses.

Gender effects

Self-reported experience. Average self-reported experience (from 1 to 9) for each category is shown in Figure 2. We conducted a repeated measure ANOVA with the factors Gender and Category (human, Barbie, Transformer, cars). As expected, there was a significant main effect for Category ($F_{(3,879)}$ = 130.68, p < .001; η_p^2 = .34) and a significant Category x Gender interaction ($F_{(3,879)}$ = 65.46, p < .001, η_p^2 = .12). LSD post-hoc tests (all p < .001) revealed no gender difference in self-reported experience for human faces, but women reporting more experience with Barbies and men with cars and Transformers.

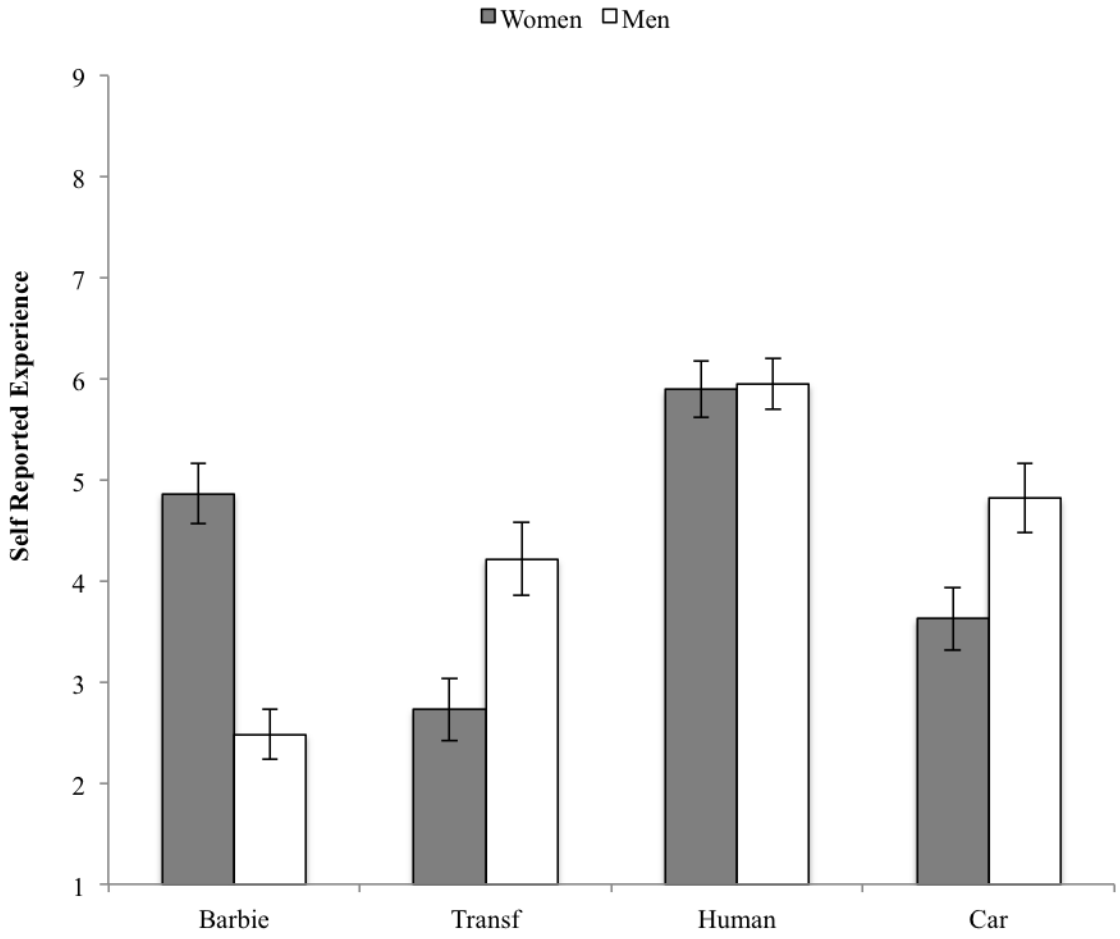

Figure 2 Self-reported experience rating for each test category separated by gender. Error bars represent 95% CI.

Performance. Average accuracy for each category is shown in Figure 3. Overall, performance was better on human faces compared to the other categories and worst with cars. We conducted an ANOVA with the factors Gender and Category (human, Barbie, Transformer, cars). There was no significant main effect for Gender ($F_{(1,293)} = .14$, $p = .71$, $\eta_p^2 = 0$) and there was a significant main effect for Category ($F_{(3,879)} = 378.92$, $p < .001$, $\eta_p^2 = .564$). The Category x Gender interaction was significant ($F_{(3,879)} = 13.22$, $p < .001$, $\eta_p^2 = .043$); LSD post hoc tests revealed that women performed better than men with Barbie faces ($p = .001$), whereas men performed better than women with Transformer faces ($p = .001$) and cars ($p = .002$).

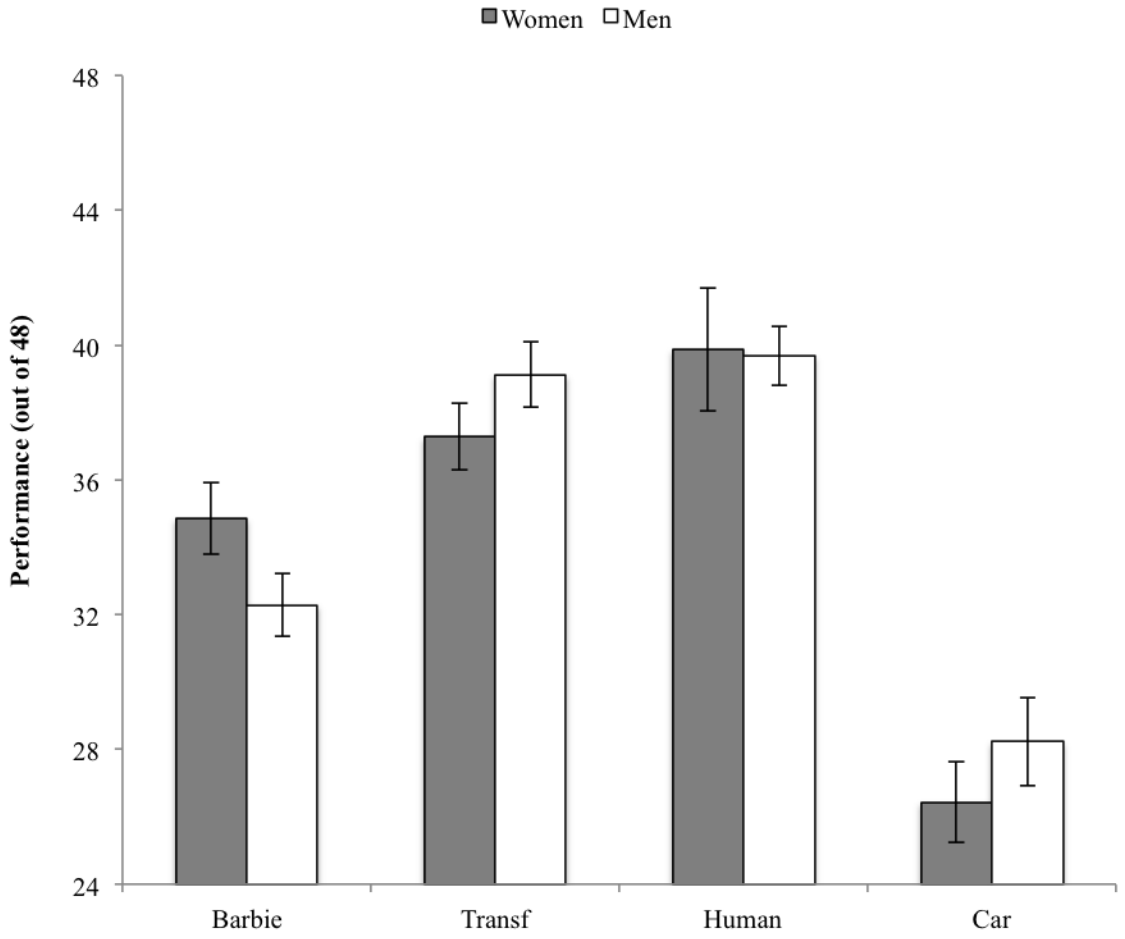

Figure 3 Performance for each test category separated by gender. Error bars represent 95% CI

These initial findings suggest that differential experience with toys may account for the gender differences in performance with these face categories. Of course, such differences can be associated with differences in motivation, which are difficult to rule out. Here, we take advantage of the fact that the VFET (just like early versions of the VET, McGugin et al;, 2012) does not use a fixed study time for the study arrays. We therefore assessed whether performance was affected by variability in the time used to study the 6 target identities. This is one indirect measure of motivation as we expect that participants will study those categories they want to perform best with the longest.

Due to technical errors, we only collected study time on the target arrays for participants who completed the study via AMT. Therefore, analysis of study time only includes these participants, and not those from our lab sample. Consistent with a role for study time, correlations between task performance and the log of study time (summed over all three study episodes) were significant across all categories (r ranged from .16 to .48, $p < .002$ for all categories). To see if differential study time accounted for gender differences in performance, we regressed study time out of performance and again conducted an ANOVA with factors Gender and Category. There were no main effects for Gender ($F_{(1,114)} = .72$, $p = .40$, $\eta_p^2 = .006$) or Category ($F_{(3,342)} = .19$, $p = .90$, $\eta_p^2 = .002$), but importantly, the Category x Sex interaction remained intact ($F_{(3,342)} = .6.53$, $p < .001$, $\eta_p^2 = .054$). LSD posthoc tests confirmed that women still outperformed men recognizing Barbies ($p = .009$), and men outperformed women recognizing Transformers ($p < .001$). Men no longer outperformed women when recognizing cars in this analysis ($p = .20$). However, the male advantage in performance for cars is also not significant when study time is not regressed out but only the AMT subjects are included in the ANOVA ($p = .22$). This may reflect the age difference between AMT and lab subjects (See

Table 1), especially since prior work has reported an influence of age on the measurement of car recognition ability (Lee et al., 2015) but not faces (Cho et al., 2015). Whatever the reason, this should not be taken to suggest that the gender difference in performance with cars is due to differences in study time.

Table 4 Between- and within- category correlations for self-reported experience (SRE) and performance (PERF) across all participants from the lab and AMT. Asterisks denote significance at the $p < .05$ level; double asterisks denote significance at the $p < .01$ level.

All Participants (N = 295)

	1	2	3	4	5	6	7
1. Human-SRE	**A**						
2. Barbie-SRE	.02						
3. Transf-SRE	-.05	.03					
4. Car-SRE	**.15****	-.04	**.25****				
				B			
5. Human-PERF	**.22****	.03	.01	**.15***	**C**		
6. Barbie-PERF	.04	**.16****	-.06	.07	**.60****		
7. Transf-PERF	.07	**-.11***	**.11***	**.22****	**.57****	**.52****	
8. Car-PERF	-.03	.02	**.17****	**.46****	**.31****	**.35****	**.39****

Table 5 Between- and within- category correlations for self-reported experience (SRE) and performance (PERF) for male participants from the lab and AMT. Asterisks denote significance at the $p < .05$ level; double asterisks denote significance at the $p < .01$ level.

Male Participants (n = 161)	1	2	3	4	5	6	7
1. Human-SRE	**A**						
2. Barbie-SRE	.04						
3. Transf-SRE	-.03	**.33****					
4. Car-SRE	**.16***	**.21****	**.20****				
				B			
5. Human-PERF	**.16***	-.01	-.01	.14	**C**		
6. Barbie-PERF	.04	-.02	-.02	**.16***	**.60****		
7. Transf-PERF	.05	.03	**.15***	**.18****	**.58****	**.58****	
8. Car-PERF	.02	**.17***	**.15***	**.51****	**.38****	**.41****	**.39****

Table 6 Between- and within- category correlations for self-reported experience (SRE) and performance (PERF) for female participants from the lab and AMT. Asterisk denotes significance at the $p < .05$ level; double asterisk denotes significance at the $p < .01$ level.

Female Participants (n = 134)	1	2	3	4	5	6	7
1. Human-SRE	**A**						
2. Barbie-SRE	.04						
3. Transf-SRE	-.11	.31					
4. Car-SRE	**.15***	.14	.12				
				B			
5. Human-PERF	**.28****	.06	.05	**.19***	**C**		
6. Barbie-PERF	.05	.12	.06	.12	**.62****		
7. Transf-PERF	.08	**-**.1	-.06	**.21****	**.58****	**.54****	
8. Car-PERF	-.09	.03	.13	**.36****	**.23****	**.37****	**.36****

Correlations between self-reported experience across categories

Although this was not a main concern of our study, for completeness, we include correlations between self-reported experience across categories (for all subjects, Table 4A, and for each gender separately, Tables 5A and 6A). Cases where self-reported experience for one category predicts a small amount of variance for another could be explained by domain-general component of experience (see Gauthier et al., 2014). There were at most small correlations across categories, with the highest being the correlation between experience with Transformers and Barbies in men (r=.33, or only 11% shared variance).

Correlations between performance across categories

With all participants, accuracy was significantly correlated across all test categories (Table 4C). The same pattern held for men (Table 5C) and women (Table 6C) separately, with one exception discussed below. Higher correlations between face categories were expected because all the face categories likely all tap into participants' general ability to recognize faces. Correlations between all pairs of face categories were numerically higher than all those involving cars. The correlation for Barbie and human faces was significantly stronger than the correlation of Barbie faces with cars ($r_{B,H}$ = .52, $r_{B,C}$ = .21, Steiger's Z = 4.53, p < .001). While we might expect the same for Transformer faces, which might seem to be less "face-like" than Barbie faces, but we found no evidence for this ($r_{T,H}$ = .50, $r_{B,C}$ = .35, Steiger's Z = 2.27, p = .02).

The correlation between performance for human faces and cars was mediated by gender, with this relationship stronger in men (r=.38) than in women (r=.28, Steiger's Z = 1.77, p = .04). This replicates previous findings from McGugin et al. (2012).

Correlations between self-reported experience and performance within category

Self-reported experience with a category is generally a poor predictor of performance on object recognition tasks (Barton et al., 2009; McGugin et al., 2012; Gauthier et al., 2014). However, in prior work (e.g., McGugin et al., 2012; Van Gulick et al., 2015), self-reports of experience with cars have been a slightly better predictor of performance than observed for other categories, though it is unclear why. We expected to replicate this here.

Across most categories, the correlation between self-reported experience and performance with that category was significant but small, and it was higher for cars (r = .46 – See Table 4B – results for each gender shown in Tables 5B and 6B). When we regress self-reported experience from performance, an ANOVA between Category, Sex, and Age Group still presents a significant Category x Sex interaction ($F_{(3,855)}$ = 4.45, p =.004) and a significant Category x Age Group interaction ($F_{(12,855)}$ = 3.87, p < .001). Therefore, while the mean results for self-report showed a pattern of gender differences similar to what we found in performance (compare figures 5B and 6B), self-report does not correlate with individual performance. Together with our finding that study time does not account for gender differences in performance with toy faces, this result argues against a motivational explanation for the gender effects in performance, to the extent that we would expect someone to be more motivated to perform with a category for which they report more experience.

Discussion

Past research finds that in the presence of a gender advantage in face recognition, women generally outperform men. We speculated that we could observe a male advantage in a face domain for which men had more experience then women. To offer this proof of concept, we

created new tests to measure face recognition abilities with human and toy faces, specifically predicting that while women would do better then men with Barbie faces, men would outperform women with Transformer faces. Women self-reported more experience with Barbies than Transformers, while men reported the reverse. Performance followed the same pattern, with women performing better with Barbie faces and men outperforming women with Transformer faces. To offer evidence that this male advantage truly occurred in a face domain (i.e., that Transformer faces are processed like faces and not like objects), we considered correlations between performance for toy faces and human faces. We found that the ability to recognize Transformer and Barbie faces was more related to the ability used to recognize human faces then to the ability to recognize cars.

Our finding that recognition of toy faces shared more variance with the recognition of human faces than with cars is consistent with research with patient CK, who has visual agnosia but very good performance with faces, which extended to faces of cartoon characters (Moscovitch, Winocur, & Behrmann, 1997). We found no evidence that recognition of Barbie faces shared more variance with human face recognition than the recognition of Transformer faces did. However, Transformer shared more variance with cars than Barbies shared with cars (mainly for men). This may appear to suggest two distinct abilities – a face recognition ability and an object recognition ability. Alternatively, a single underlying visual recognition ability could interact with domain-specific experience to yield performance differences on these various tasks (Gauthier et al., 2014). Further work with a larger number of face categories and improved measures of experience could help address these questions.

The male advantage we observed for recognizing Transformer faces provides evidence against a general female advantage for any type of faces, for instance one that could reflect an

innate, female-specific bias for learning or remembering faces (Baron-Cohen, 2002; Bowles et al., 2009; Ingalhalikar et al., 2014; Sawada et al., 2014). One alternative interpretation of the male advantage for recognizing Transformer faces is that both men and women use their face recognition skill with these faces, but only men see an additional contribution from their object recognition skill (which, according to Dennett et al., 2012, may be stronger due to a male advantage for tasks like mental rotation). Our results do not align with this account: a multiple regression of human face recognition explained 30.7% of the variance on the Transformers task for men, but explained only 20.2% of the variance for women. Car recognition (partialing out the contribution from human face recognition), added only 3.9% to Transformer face recognition for men and 6.9% for women. This is inconsistent with a greater contribution of object recognition ability to performance with Transformers for men. Attributing performance on the car recognition task with an "object recognition ability" may be too simplistic, given that car recognition shared significantly more variance with human faces in men than in women (replicating previous work, McGugin et al., 2012; VanGulick et al., 2015). Such a pattern of results could reflect a single domain-general recognition ability that is relevant to all our tasks, but that is better measured with categories for which experience is higher (see Gauthier et al., 2014).

We initially chose toy faces because we conjectured that they would be associated with different amounts of experience in men and women, and our participants confirmed this conjecture. However, the only measure of experience that we have is based on one self-report question for each domain, a question that was used in prior work (McGugin et al., 2012; Gauthier et al., 2014; VanGulick et al., 2015). This question has shown acceptable test-retest in object domains (Gauthier et al., 2014) and by itself, was as informative in predicting semantic

and visual performance as an aggregate of several questions about distinct aspects of experience (VanGulick et al., 2015). Nonetheless, the correlation between self-report and visual performance in any given domain is generally fairly low (McGugin et al., 2012; Gauthier et al., 2014; VanGulick et al., 2015), as is typical of self-reports across a wide range of domains (Zell & Krizan, 2014), although it is generally higher in the case of cars (why is currently unknown). In the present work, self-reports were not good predictors of performance within each category (again, with the exception of cars), but they did produce a clear Gender x Category interaction that paralleled the pattern of performance on our tasks. This is likely because it is much easier for individuals to provide information about their own relative experience for different categories then it is to compare themselves to others on perceptual skills. Therefore, it is important to acknowledge that we did not manipulate experience, and that there are important limitations in the measurement of experience using self-report. While we conclude that differential experience accounts for these gender effects, this is to some extent a conjecture similar to that which others have made to explain the same-race, same-species or same-age advantages in face recognition. In addition, it does not speak to the underlying causes of differential experience, which for gender effects could have either an evolutionary or a cultural basis.

A possible account of the crossover interaction we obtained between gender and toy face category is differential motivation for men and women with different toy faces. We have no direct measure of motivation but we considered two indirect markers of motivation to try to rule out this account. First, we argue that if motivation varied greatly across subjects, with some more motivated to perform well then other people with a category, say Transformers, the self-report measure would likely reflect such differences. Here, self-reported experience did not predict more than 1.7% of the variance in performance with Barbie faces, and less than 1% for

Transformer faces, suggesting either that motivation did not contribute much to performance or did not influence self-reports at all. Second, we argue that higher motivation would translate in more effort and therefore longer study times, which were here left up to subjects. We found that while longer study times predicted better performance, regressing out study time did not remove the gender effect for toy faces.

Finally, in a spirit similar to recent work in object recognition (McGugin et al., 2012), the VFET battery uses images that vary considerably in background, pose and expression. In this regard, it is more similar to the VET than to other face tasks like the CFMT. Because its format may be easier to replicate with new types of faces, it allows for more flexibility when testing with several face and object categories is desired. Despite these differences, the VFET behaves similarly to the CFMT in several ways and may therefore provide an interesting alternative (or complement in approaches that use multiple tests to derive latent factors). The correlation between the VFET human face task the VET-car task (r = .31) was similar to that reported between CFMT and car recognition tasks in prior work (Dennett et al., (2012): r = .37; Gauthier et al., (2014): r = .24). Therefore, the presence of backgrounds does not appear to render the VFET less face-specific then the CFMT. In comparison to previous findings with the CFMT that show a small female advantage (Cho et al., 2015), we observed no significant gender effect for the male and female portions of the VFET, despite our sample showing gender effects for other face and object categories. However, the CFMT does not always show a gender effect and it would be desirable to compare both tests with a sufficiently large sample to detect even small gender effects.

Conclusion

Our face recognition measures provide useful tools for continued exploration of individual differences, including gender effects, in face recognition. It will be useful to develop reliable measures of individual differences in performance for a larger variety of face categories. Experience likely plays an important role as a determinant of performance, and experience is generally domain-specific. The limitations of asking if a person has more "object" experience then another may be obvious (it seems more appropriate to ask about experience within specific object domains like cars, birds or mushrooms). Likewise, the present results suggest that domain-specific experience also influences performance with sub-categories of faces, and that measuring face recognition ability as if it was a single monolithic domain may present limitations. Extending the framework we have presented here, we may eventually be able to characterize complex individual profiles of performance across face categories that reflect our varied experience with different kinds of faces.

REFERENCES

Baron-Cohen, S. (2002). The extreme male brain theory of autism. Trends in sciences, 6(6), 248-254.

Barton, J. J., Hanif, H., & Ashraf, S. (2009). Relating visual to verbal semantic knowledge: the evaluation of object recognition in prosopagnosia. Brain,132(12), 3456-3466.

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

Cho, S., Wilmer, J., Herzmann, G., McGugin, R., Fiset, D., Van Gulick, A., Ryan, K., Gauthier, I. (2015). Item response theory analyses of the Cambridge Face Memory Test (CFMT), Psychological Assessment.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. PloS one, 8(3), e57410.

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.

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.

Gauthier, I., McGugin, R. W., Richler, J. J., Herzmann, G., Speegle, M., & Van Gulick, A. E. (2014). Experience moderates overlap between object and face recognition, suggesting a common ability. Journal of vision, 14(8), 7.

Goldstein, A. G., & Chance, J. E. (1970). Visual recognition memory for complex configurations. Perception & Psychophysics, 9(2), 237-241.

Ingalhalikar, M., Smith, A., Parker, D., Satterthwaite, T. D., Elliott, M. A., Ruparel, K., ... & Verma, R. (2014). Sex differences in the structural connectome of the human brain. Proceedings of the National Academy of Sciences, 111(2), 823-828.

Lee, W.-Y., Cho, S.-J., McGugin, R.W., Van Gulick, A.B. & Gauthier, I. (2015). Differential Item Functioning Analysis of the Vanderbilt Expertise Test for Cars (VETcar). Journal of Vision, 15(13):23

Lewin, C., & Herlitz, A. (2002). Sex differences in face recognition—Women's faces make the difference. Brain and cognition, 50(1), 121-128.

Lindsay, D. S., Jack, P. C., & Christian, M. A. (1991). Other-race face perception. Journal of Applied Psychology, 76(4), 587-589.

Lovén, J., Svärd, J., Ebner, N. C., Herlitz, A., & Fischer, H. (2013). Face gender modulates women's brain activity during face encoding. Social cognitive and affective neuroscience, nst073.

McGugin, R. W., Richler, J. J., Herzmann, G., Speegle, M., & Gauthier, I. (2012). The Vanderbilt Expertise Test reveals domain-general and domain-specific sex effects in object recognition. Vision Research.

Moscovitch, M., Winocur, G., & Behrmann, M. (1997). What is special about face recognition? Nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition. Journal of Cognitive Neuroscience, 9(5), 555-604.

Pascalis, O., de Haan, M., & Nelson, C. A. (2002). Is face processing species-specific during the first year of life? Science, 296(5571), 1321-1323.

Rehnman, J., & Herlitz, A. (2007). Women remember more faces then men do. Acta Psychol., 124, 344-355.

Richler, J.J., Floyd, R.J., & Gauthier, I. (2014). The Vanderbilt Holistic Face Processing Test: a short and reliable measure of holistic face processing. Journal of Vision, Vol.14, 10. doi:10.1167/14.11.10

Ryan, K., & Gauthier, I. (2014). Gender effects for toy faces: quantitative differences in face processing strategies. Poster presented at the Vision Sciences Society Annual Meeting, St. Pete Beach, FL, USA.

Sawada, R., Sato, W., Kochiyama, T., Uono, S., Kubota, Y., Yoshimura, S., & Toichi, M. (2014). Sex Differences in the Rapid Detection of Emotional Facial Expressions. PloS One, 9(4), e94747.

VanGulick, A.E., McGugin, R.W., & Gauthier, I. (2015). Measuring non-visual knowledge about object categories: The Semantic Vanderbilt Expertise Test. Behavior research methods, 1-19.

Wolff, N., Kemter, K., Schweinberger, S. R., & Wiese, H. (2014). What drives social in-group biases in face recognition memory? ERP evidence from the own-gender bias. Social cognitive and affective neuroscience, 9(5), 580-590.

Yovel, G., Halsband, K., Pelleg, M., Farkash, N., Gal, B., & Goshen-Gottstein, Y. (2012). Can massive but passive exposure to faces contribute to face recognition abilities. Journal of Experimental Psychology: Human Perception and Performance, 38(2), 285-289.

Zell, E., & Krizan, Z. (2014). Do people have insight into their abilities? A metasynthesis. Perspectives on Psychological Science 9(2), 111-125.