Vision Research 69 (2012) 10–22

Contents lists available at SciVerse ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

The Vanderbilt Expertise Test reveals domain-general and domain-specific sex effects in object recognition

Rankin W. McGugin [a,*], Jennifer J. Richler [a], Grit Herzmann [b], Magen Speegle [a], Isabel Gauthier [a]

[a] Department of Psychology, Vanderbilt University, Nashville, TN 37203, USA
[b] Department of Psychology and Neuroscience, University of Colorado Boulder, Boulder, CO, USA

ARTICLE INFO

Article history:
Received 21 January 2012
Received in revised form 20 July 2012
Available online 2 August 2012

Keywords:
Perceptual expertise
Face
Object recognition
Sex effects

ABSTRACT

Individual differences in face recognition are often contrasted with differences in object recognition using a single object category. Likewise, individual differences in perceptual expertise for a given object domain have typically been measured relative to only a single category baseline. In Experiment 1, we present a new test of object recognition, the Vanderbilt Expertise Test (VET), which is comparable in methods to the Cambridge Face Memory Task (CFMT) but uses eight different object categories. Principal component analysis reveals that the underlying structure of the VET can be largely explained by two independent factors, which demonstrate good reliability and capture interesting sex differences inherent in the VET structure. In Experiment 2, we show how the VET can be used to separate domain-specific from domain-general contributions to a standard measure of perceptual expertise. While domain-specific contributions are found for car matching for both men and women and for plane matching in men, women in this sample appear to use more domain-general strategies to match planes. In Experiment 3, we use the VET to demonstrate that holistic processing of faces predicts face recognition independently of general object recognition ability, which has a sex-specific contribution to face recognition. Overall, the results suggest that the VET is a reliable and valid measure of object recognition abilities and can measure both domain-general skills and domain-specific expertise, which were both found to depend on the sex of observers.

1. Introduction

Most efforts aimed at understanding how faces are perceived and recognized have relied on comparing performance with faces to performance with other objects. Research in psychology, neuropsychology, and cognitive neuroscience has seen much debate regarding whether face processing is qualitatively special, with evidence coming from studies in which faces are compared to a single category of non-face objects (Carey, 1992; Diamond & Carey, 1986; Gauthier & Tarr, 1997; Gauthier et al., 1999, 2000; Kanwisher, McDermott, & Chun, 1997; Moscovitch, Winocur, & Behrmann, 1997; Tanaka & Farah, 1993; Young, Hellawell, & Hay, 1987). Given that the specific contrast category varies from study to study, a review of this literature can appear to support strong claims about a given behavioral or neural response associated with faces, concluding that faces elicit phenomenon X more than category A, and more than category B, and more than category C, and so on, such that phenomenon X may be called face-selective. However, there are logical problems with an approach that relies on a qualitative summary of multiple single paired quantitative contrasts, in that potential differences between non-face object categories are not revealed and their theoretical significance is thereby ignored (Gauthier & Nelson, 2001).

For instance, consider the Thatcher Illusion for faces – the relative difficulty in perceiving the local inversion of face parts when the whole face is inverted (Thompson, 1980). When this effect for faces was compared to the same effect for a variety of non-face categories (instead of contrasted to one category at a time), the magnitude of the illusion for faces was not found to be exceptionally large compared to the distribution obtained for non-face objects (Wong et al., 2010). Whenever the goal is to compare faces to non-face objects, performance for multiple categories should be obtained.

Likewise, similar problems arise in studies that aim to quantify perceptual expertise in a specific domain. For instance, car expertise has often been quantified using performance on a matching task for pairs of cars relative to matching for another category (typically birds), so that general visual performance and motivation may be factored out (Gauthier et al., 2000, 2003; Harel et al., 2010; McGugin & Gauthier, 2010; McGugin et al., 2010; McKeeff et al., 2010; Rossion & Curran, 2010; Rossion & Gauthier, 2002; Xu, 2005). However, Williams, Willenbockel, and Gauthier (2009) noted that this measure of perceptual expertise can be difficult

* Corresponding author.
E-mail address: rankin.mcgugin@vanderbilt.edu (R.W. McGugin).

0042-6989/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.visres.2012.07.014

to interpret for individuals who do not claim to be bird experts but nonetheless perform very well with birds. In that study, such participants appeared to be outliers in the relationship between expertise for cars relative to birds and perceptual sensitivity to spatial frequency content in car images. It is simply suboptimal to measure expertise through the comparison of one category of interest to a single control category. Instead it would be more compelling to deem someone a car expert (for instance) based not only on above average performance with cars, but also on markedly better performance for cars relative to several other non-car categories. While it is possible to be an expert in more than one domain, evidence of domain-specific expertise for any one category is weakened by equally good performance for multiple other categories. Presumably, a test with multiple categories is better apt to distinguish between general object recognition ability and expertise factors. There have been efforts in the study of patients with agnosia to compare performance (typically on an old/new recognition test) for faces with that for a number of other categories (e.g., Duchaine, Germine, & Nakayama, 2007; Duchaine et al., 2003, 2006; Germine et al., 2011). However, such studies use fairly small samples and individual differences among controls was not the focus of the work.

In Experiment 1, we will describe a new test of object recognition, the Vanderbilt Expertise Test (VET), which measures the ability to recognize examples from eight categories of visually similar objects: leaves, owls, butterflies, wading birds, mushrooms, cars, planes and motorcycles. The goal of Experiment 1 is to demonstrate the reliability of the VET in a large sample of participants and to explore the structure of object recognition skills as measured by the VET using principal component analysis. To preview one aspect of our results, we find that increasing the number of tested categories had an interesting side-effect: to reveal large sex differences in object recognition.

In Experiment 2, we demonstrate how the VET can be used to dissociate domain-general and domain-specific contributions to a standard perceptual matching measure of expertise (Bukach, Phillips, & Gauthier, 2010; Gauthier et al., 1999; Harel et al., 2010; Rossion & Curran, 2010; Xu, 2005). Interestingly, these two contributions were similar for men and women in one domain (cars) but quite different for another (planes).

In Experiment 3, we use the VET to assess the contribution of general object recognition ability to face recognition and test the hypothesis that holistic processing of faces predicts face recognition independently of general object recognition ability. Here again, we show that taking sex into account combined with a variety of categories is critical, since object recognition reveals a sex-specific contribution to face recognition.

2. Experiment 1

In Experiment 1, our goals are similar to those of Dennett et al. (2011), who published the Cambridge Car Memory Task (CCMT) aiming to assess object recognition in a manner matched to the well-established Cambridge Face Memory Task (CFMT; Duchaine & Nakayama, 2006), which effectively measures face recognition using a 3-alternative forced choice memory test. But here, we sought to create a test for eight different categories to capture both domain-general object recognition skills as well as domain-specific expertise. A test with a single category (Dennett et al., 2011) would not be able to dissociate these two factors. Even a test with two categories (e.g., a comparison of cars to birds; Gauthier et al., 2000) could not achieve this goal, because it cannot distinguish between someone with expertise in both categories and someone who simply has very good general object recognition skills. Like Dennett et al., we modeled our test after the CFMT (with some differences that will be mentioned below), because this measure has become a standard test of face recognition, has very good reliability (Bowles et al., 2009; Herzmann et al., 2008; Wilmer et al., 2010), and has demonstrated validity in a number of different studies and for several populations (e.g., Bowles et al., 2009; Germine, Duchaine, & Nakayama, 2010; Richler, Cheung, & Gauthier, 2011b; Woolley et al., 2008), including a large sample study that demonstrated both convergent and discriminant validity (Wilmer et al., 2010).

We chose eight categories of objects from visually homogeneous categories that included cars, planes, motorcycles, butterflies, wading birds, owls, mushrooms and leaves. "Visually homogenous" here is a fairly intuitive criterion that requires a common configuration of parts (and would thus exclude subordinate categories such as "toys" or "jewelry"). We selected some categories that have been used in studies of real-world experts (cars, planes, butterflies, birds; Bukach, Phillips, & Gauthier, 2010; Dennett et al., 2011; Gauthier et al., 2000; McGugin et al., submitted for publication; Rhodes et al., 2004) or in lab-training studies (owls, wading birds; Tanaka, Curran, & Sheinberg, 2005; cars, Scott et al., 2006). We acknowledge that the selection of these categories was relatively arbitrary, aiming to include a variety of different categories and influenced by convenience (categories for which we were able to find a sufficient number of exemplars). This seemed appropriate given that our main goal was to demonstrate the advantage of measuring individual differences in a broader context (see Wong et al., 2010 for a similar approach to category selection). The selection of these categories is not meant as a theoretical statement about the kinds and number of categories required to fully sample object recognition abilities.

We chose to include both owls and wading birds to explore whether two sub-categories that are highly related might be more associated than any of the other categories in a sample of undergraduates not selected for special interest in birds. Indeed, prior work has shown that performance on these two categories (Tanaka, Curran, & Sheinberg, 2005) and other related sub-categories (such as modern and antique cars; Bukach, Phillips, & Gauthier, 2010) can be dissociated through real-world experience or lab training, but these studies did not include a range of other categories as a comparison.

We applied principal component analysis (PCA) to the mean performance with different categories in the VET. PCA combines categories that are correlated with one another into factors that are largely independent from other factors that represent different, uncorrelated categories. A PCA of the VET can help reveal the latent factors that account for performance in a learning-to-individuate task with different object categories. One possibility is that most of the variance is accounted for by a single common factor of general object-recognition ability (at least in a sample of participants who are not selected for special interests for any of these categories). Another possibility is that these eight categories will cluster into meaningful factors, such as living vs. non-living groupings. We also explored how any factors we extract from the VET relate to face recognition as measured by the CFMT.

2.1. Participants

Two hundred and twenty-seven individuals participated for a small honorarium or course credit: Caucasian (76 male, 82 female, mean age 23.3 ± 4.2), African American (11 male, 20 female, mean age 22.8 ± 3.3), Asian (10 male, 18 female, mean age 20.7 ± 3.6), and Other (5 male, 1 female, mean age 19.3 ± 1.4). All participants had normal or corrected-to-normal visual acuity. The experiment was approved by the Institutional Review Board at Vanderbilt University, and all participants provided written informed consent.

2.2. Vanderbilt Expertise Test

Stimuli Eight different subordinate object categories were included: leaves, owls, butterflies, wading birds, mushrooms, cars, planes and motorcycles. For each category, target images consisted of four exemplars from six unique species/models, while distractor images showed 48 exemplars from novel species/models (Supplemental Table 1). Stimuli were digitized, eight-bit greyscale images presented on a 20-in. Samsung LCD monitor (refresh rate = 100 Hz) with a Macintosh Mini computer using Matlab and Psychophysics Toolbox extension (Brainard, 1997; Pelli, 1997). All images subtended a 5.2° visual angle.

The 72 images for each category, as well as the target–distractor pairings for each trial, were selected based on the results of a pilot experiment ($N = 32$). We used item and trial analyses to assess the reliability of specific trials and of individual exemplars within a category. We considered averaged accuracy for each of the six items in each category, as well as the correlation between mean item performance and mean overall performance for each category. Twelve percent of target items were replaced for being too difficult or not predictive of overall performance. They were replaced with images that were judged to be more comparable to items that fared better. Next, we considered mean trial performance, and asked how well performance on a given triplet-item predicted overall behavior. In triplets with insufficient subject variability, low predictive power, and/or simply accuracy that was too low, distractor objects were replaced to make the target more distinguishable (6% trials), as it seemed to be the most obvious cause of these problems.

2.3. Procedure

The task was roughly modeled after the CFMT (Duchaine & Nakayama, 2006). While the CFMT includes trials that have noise added to the images, pilot testing showed that performance with object categories was unlikely to be high enough to require this manipulation. In addition, CFMT trials (Duchaine & Nakayama, 2006) or CCMT trials (Dennett et al., 2011) with and without noise tend to be highly correlated.

At the start of the experiment, participants rated themselves on their expertise with all tested categories (leaves, owls, butterflies, wading birds, mushrooms, cars, planes and motorcycles), and also with faces, considering "interest in, years exposure to, knowledge of, and familiarity with each category", where 1 represented the lowest assumed skill level and 9 represented the highest.

Participants began with six practice trials where they studied three cartoon faces or three owls, followed by three recognition memory trials of either an identical study image or an image showing another viewpoint of a studied cartoon/owl. Categories were tested in separate blocks. Before the start of an experimental category block, participants viewed a study screen with one exemplar from each of six species/models. The target images were arranged in two rows of three images with target names presented above or below each image (Fig. 1, Supplemental Table 1). We chose to include labels because experts often know labels that novices do not have for individual exemplars, and even though this task does not require labels, they may be useful to some participants (e.g., Lupyan, Rakison, & McClelland, 2007). We hoped that making labels available might provide novices with at least some of the advantage that experts may experience for this reason. The VET could also be used easily without the labels.

Participants studied these images for as long as they needed. For each of the first twelve trials, one of the studied exemplars was presented with two distractors from another species/model in a forced-choice paradigm ('Same exemplar' trials). The target image could occur in any of the three positions, and participants indicated which image of the triplet was the studied target: the left, middle and right positions were specified by 1, 2 and 3, respectively, on the number pad. Participants were instructed to be as accurate as possible and triplets were shown until participants made a response. On these trials, performance could depend on any aspect of the image, backgrounds included, since the target image was identical to the studied image. Feedback was provided on each Same exemplar trial, indicating the correct image and image name. The study screen appeared for review at the end of the twelve image-match trials.

After the first 12 trials, participants were warned that the subsequent target images would be different exemplars of the studied species/models, and that they would be required to generalize across viewpoint, size and backgrounds ('Different exemplar' trials). For the remaining 36 Different exemplar trials, no feedback was provided. Participants viewed image triplets, indicating which

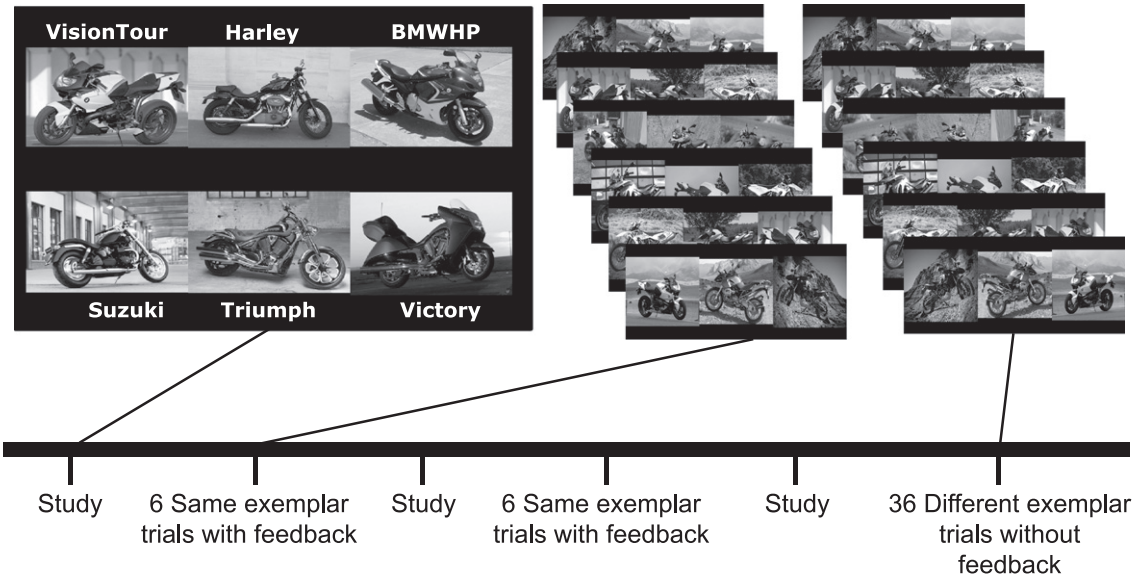

Fig. 1. VET trial structure. One representative category block is shown (Motorcycles). Participants studied six exemplars with labels, then performed a forced-choice matching task on triplets containing either the identical image that was studied (Same exemplar trials) or a different view or year from the same make and model (Different exemplar trials). Six examples are shown for the Same and Different exemplar trial types.

exemplar corresponded to one of the target species/models studied. For each category, there were three transfer images for each of the six targets; transfer images represented unique exemplars of the same subordinate category; e.g., 3 different images of the Pipevine Swallowtail served as transfer images for the 1 studied image of a Pipevine Swallowtail. All transfer images were shown twice, each time with unique distractors.

Participants performed 12 identical and 36 transfer trials for each object category. The trial structure – including block order (alphabetical: butterflies, cars, leaves, motorcycles, mushrooms, owls, planes and wading birds[1]), trial order and distractor selection – was fixed for all participants to provide more reliable across-category comparisons. The total experiment for eight categories consisted of 384 trials, lasting between 35 and 45 min depending on individual speed.

2.4. Cambridge Face Memory Test (CFMT)

At the start of the CFMT, participants studied frontal views of six target faces for a total of 20 s. Then, they completed an 18-trial introductory learning phase, after which they were presented with 30 forced-choice test displays. Each display contained one target face and two distractor faces. Participants were told to select the face that matched one of the original six target faces. The matching faces varied from their original presentation in terms of lighting condition, pose, or both. Next, participants were again presented with the six target faces to study, followed by 24 test displays presented in Gaussian noise. For a complete description of the CFMT, see Duchaine and Nakayama (2006).

3. Results

3.1. Accuracy and self-report

First, individual data were inspected for outliers. Four participants were excluded because more than 40% of the trials in at least 3 out of 8 object categories showed reaction times below 200 ms, suggesting invalid response patterns for these participants. Subsequent analyses represent data from the remaining 223 participants.

We then examined performance for each category. For all categories, average performance was significantly above chance (.333) (Table 1; Supplemental Table 2). As expected, accuracy rates were higher during Same exemplar trials (i.e., the image tested was identical to the image studied) relative to Different exemplar trials (i.e., the image tested represented a Different exemplar of the studied make/model/species).

The boxplots in Fig. 2 provide a quick visual comparison across all VET categories and faces from the CFMT, showing the central tendency, dispersion and skewness of individual accuracy scores for each category. (See Supplemental Fig. 1 for equivalent boxplots separated by sex.) Non-parametric Kruskal–Wallis H test revealed a significant difference between categories ($H_{(9)} = 368.83$, $p < 0.0001$). Nemenyi post hoc tests comparing pairwise mean ranks found that the performance on CFMT no-noise trials was significantly greater than performance for all other categories (all $ps < .001$).

Indicated by the position of the box within the whiskers (taking into consideration the tagged extreme values) and representing the degree and direction of asymmetry, most category distributions are skewed towards high accuracy (Table 1), reflecting a ceiling effect for certain categories (e.g., faces without noise) and not others (e.g., wading birds). Relative to the other face conditions, the CFMT trials *with noise* appear most comparable to the non-face trials. The CFMT trials with and without noise are correlated ($r = .72$, $p < .001$), consistent with earlier work showing a correlation of $r = .74$ for the CFMT (Duchaine & Nakayama, 2006) and $r = .61$ for the CCMT (Dennett et al., 2011). The box length, or the interquartile range (IQR), indicates the sample variability, revealing the largest spread for faces (IQR = .25 noise and .23 no noise) and cars (IQR = .23), with all other categories less than .20.

Table 1
Accuracy from CFMT and VET for all subjects ($N = 223$) separated by sex. Columns represent the Mean, 95% Confidence Interval (CI), Median, Interquartile Range (IQR), and Skewness.

	Mean	95% CI	Median	IQR	Skewness
Males					
CFMT no noise	0.77	(.74, .80)	0.8	0.27	−0.58
Owls	0.68	(.66, .70)	0.69	0.15	−0.13
Planes	0.7	(.67, .72)	0.71	0.19	−0.11
CFMT noise	0.66	(.62, .70)	0.67	0.27	−0.30
Cars	0.67	(.64, .70)	0.69	0.27	−0.24
Butterflies	0.58	(.55, .06)	0.6	0.21	−0.31
Wading birds	0.6	(.58, .62)	0.6	0.14	0.24
Mushrooms	0.6	(.57, .62)	0.6	0.17	−0.58
Leaves	0.57	(.55, .59)	0.58	0.17	−0.17
Motorcycles	0.6	(.57, .62)	0.6	0.21	−0.09
Females					
CFMT no noise	0.82	(.79, .85)	0.87	0.2	−1.09
Owls	0.71	(.69, .73)	0.73	0.16	−0.47
Planes	0.66	(.64, .68)	0.67	0.15	−0.62
CFMT noise	0.68	(.65, .71)	0.67	0.23	−0.30
Cars	0.62	(.59, .64)	0.63	0.19	0.18
Butterflies	0.64	(.62, .66)	0.65	0.17	−0.39
Wading birds	0.62	(.60, .64)	0.63	0.16	0.17
Mushrooms	0.61	(.60, .63)	0.63	0.15	−0.61
Leaves	0.6	(.58, .62)	0.6	0.16	−0.21
Motorcycles	0.57	(.55, .59)	0.58	0.15	−0.32

Similarly, self-report scores of expertise plotted to the right reveal the highest self-reported abilities for faces and cars (Supplemental Table 3). A Kruskal–Wallis H test revealed a significant difference in self-reported expertise between categories ($H_{(9)} = 546.56$, $p < 0.0001$), and Nemenyi post hoc tests showed higher self-ratings for faces relative to other categories. Self-report for all categories showed a positive skew, except faces (−.758).

3.2. Sex and age effects

Interestingly, age was significantly correlated with performance for cars and planes, but no other categories (Table 2). Moreover, we also found significant relationships between performance and sex for most categories (except wading birds and mushrooms; Table 2). These effects will be explored in more depth later, but first we wanted to remove their influence to explore relationships between accuracy and self-reported expertise. Thus, we considered pairwise correlations across accuracy rates for all categories, and correlations between accuracy and self-report, while partialing out the contributions of age and sex.

3.3. Within and between category correlations for accuracy

First, we consider the partial correlations in performance among all nine categories, including faces (Fig. 3a). Most correlations are positive and significant, suggesting an expected common factor in the memory task. Faces and cars are the least related to other categories, and interestingly, the face–car correlation is also below average. Faces and cars are also the two categories for which participants report the most expertise (Supplemental Table 3), suggesting that category-specific experience is likely to make perfor-

[1] Results for the separate categories are presented in an order that will be relevant for subsequent factor analyses, rather than the order in which the category blocks were shown in the experiment.

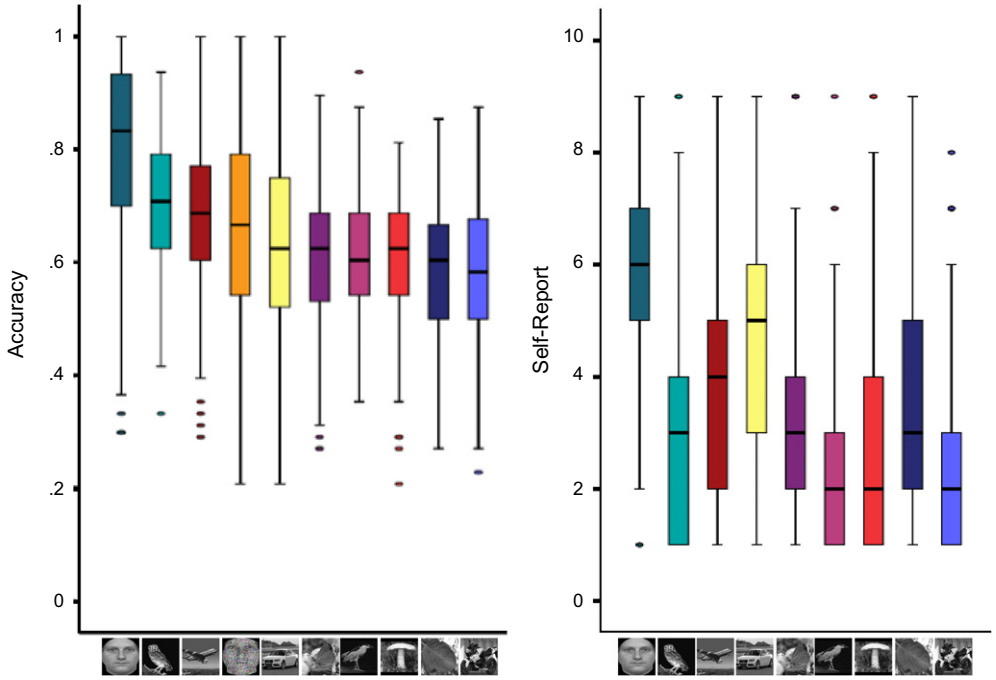

Fig. 2. Parallel boxplots depicting the distribution of accuracy scores (*left*) or self-report scores (*right*) for all categories in descending order of mean accuracy (left to right: CFMT-no noise, owls, planes, CFMT-noise, cars, butterflies, wading birds, mushrooms, leaves and motorcycles) in all participants ($N = 223$). The top and bottom of the box represent the upper and lower quartiles (75th and 25th percentiles, respectively), and the central brand signifies the median (50th percentile). Points outside the whiskers indicate extreme values that are over one and a half times beyond the interquartile range.

Table 2
Correlations with category accuracy. Zero-order Pearson's correlations are shown for correlations with age for correlations with Sex, Spearman's correlation was used and male and female were coded as 1 and −1, respectively. Ninety five percentage confidence intervals are shown in brackets.

	Pearson's correlation, *r*, with age	Spearman's correlation, *r*, with sex
CFMT	.109*	−0.108
	(−.022, .236)	(−.236, .023)
Leaves	.014	−0.141*
	(−.117, .145)	(−.267, −.01)
Owls	−.010	−0.139*
	(−.141, .121)	(−.265, −.008)
Butterflies	.069	−0.227*
	(−.062, .198)	(−.347, −.099)
Wading birds	.093	−0.06
	(−.038, .221)	(−.189, .071)
Mushrooms	.087	−0.073
	(−.044, .215)	(−.202, .058)
Cars	.223*	0.173*
	(.095, .344)	(.043, .297)
Planes	.214	0.109
	(.086, .335)	(−.022, .236)
Motorcycles	.091	0.113
	(−.04, .219)	(−.018, .24)

* $p < .0 < .05$.

mance on a category diverge from that for other categories. When sex and age were not partialed out, car performance was still only significantly correlated with that for faces, planes and motorcycles, suggesting that age and sex alone cannot account for this effect (see Supplemental Table 4 for zero-order correlations). Furthermore, considering partial correlations between self-reported expertise and performance (age and sex partialed out; Supplemental Fig. 2a), here again, cars and faces stand out: people who perform well with cars/faces report greater knowledge of cars/faces ($r = .322$ and $r = .167$, respectively) (see Supplemental Fig. 3 for heatmaps split by sex.).

3.4. Reliability

We measured reliability as an estimate of the internal consistency of test items within a category using Cronbach's alpha (Table 3). Reliability across all participants was acceptable for all categories ($\alpha > .7$) except mushrooms ($\alpha = .635$). Reliability was especially high for cars ($\alpha = .845$). Most categories (with the exception of wading birds and mushrooms) showed similar internal consistency for males and females. Importantly, the poor test reliability for these categories did not influence the principal components' extraction.[2] Thus, while wading birds and mushrooms are questionable tests for males and females, respectively, they do not alter the reported findings.

3.5. Principal component analysis

Performance for all trials and all conditions of the VET were subjected to a factor analysis. Principal component analysis (PCA) was applied to the mean accuracy data for each task and each subject in an effort to summarize patterns of correlations among observed category accuracy rates. The colinearity of the eight different object categories in the recognition memory tasks was explored through PCA with varimax rotation. The PCA matrix conducted on the accuracy data reduced a set of eight categories into two principal components (factor loadings reported in Table 4), providing an operational definition for underlying object recognition skills.

Note that the CFMT was not subjected to the PCA because the goal was to extract the structure of the object recognition captured

[2] We repeated the PCA dropping wading birds and mushrooms from the analysis. Two factors were extracted, one corresponding to leaves, owls, and butterflies (accounting for ~49% of the total variance) and the other corresponding to cars, planes, and motorcycles (accounting for ~18% of the total variance). Critically, the sex effect observed for Factors 1 and 2 when *all* categories are considered is preserved when we restrict the dataset to the categories with highest reliability.

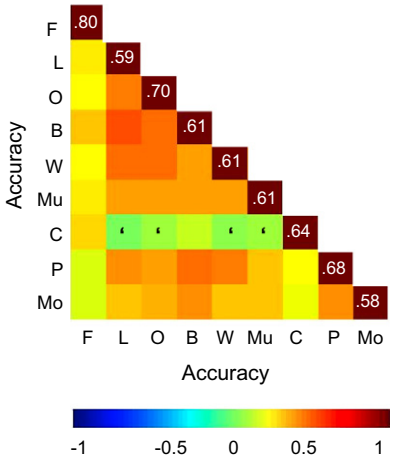

Fig. 3. Heatmap depicting partial correlations for Accuracy rates across categories (faces (F), leaves (L), owls (O), butterflies (B), wading birds (W), mushrooms (Mu), cars (C), planes (P) and motorcycles (Mo)), while partialing out the influence of age and sex variables. All correlations are significant ($p < 0.05$) except those indicated with an apostrophe ('). The values in the downward diagonal of maroon squares represent the average group accuracy for a particular category.

Table 3
Cronbach's Alpha coefficients as an estimate of internal reliability shown separately for all participants ($N = 223$), males ($N = 102$) and females ($N = 121$).

	All	Males	Females
Leaves	.702	.703	.701
Owls	.740	.731	.751
Butterflies	.778	.786	.764
Wading birds	.701	.635	.742
Mushrooms	.635	.704	.545
Cars	.845	.866	.812
Planes	.791	.807	.771
Motorcycles	.734	.740	.723

Table 4
Principal component analysis of mean accuracy data, from which factor loadings are extracted for all tasks.

	Factor 1	Factor 2
Leaves	.793	.011
Owls	.741	.086
Butterflies	.735	.224
Wading birds	.727	.134
Mushrooms	.645	.152
Cars	−.142	.880
Planes	.494	.600
Motorcycles	.431	.550

by the VET by itself. However, when included into the analysis, we find a factor loading nearly exclusively on faces. Critically, the results for object categories are qualitatively the same regardless of whether we choose to extract two factors (without the CFMT) or three factors (with the CFMT).

The first factor of the PCA without CFMT explained 47.8% of the variance in all tasks and loaded on performance for all natural categories (leaves, owls, butterflies, wading birds and mushrooms). The second factor explained an additional 13.9% of the variance. This factor was primarily associated with cars, but also planes and motorcycles. Although performance for planes and motorcycles loaded moderately on both factors, the loadings on Factor 1 did not reach the minimum .5 value, upon which tasks were classified as represented by a given factor. All other extracted zcomponents explained less than 8% of the variance. The Kaiser–Meyer–Olkin measure of sampling adequacy was high (.86) and indicated a good fit of the two-factor model with the real data.

3.6. Sex effects

We tested the influence of sex on individual factor loadings using a 2 × 2 Mixed ANOVA with the within-subjects variable Factor (2 levels: Factor 1, nature factor and Factor 2, vehicle factor) and the between-subjects variable Sex (2 levels: Male and Female). Neither main effect was significant, but there was a significant interaction between Factor and Sex ($F_{1,221} = 23.76$, $MSE = 0.907$, $p < .001$, $\eta^2 = .10$). Post-hoc analyses reveal significant differences between all relevant points: Factor 1, Males vs. Factor 1, Females ($p = .002$); Factor 2, Males vs. Factor 2, Females ($p < .001$); Factor 1, Males vs. Factor 2, Males ($p < .001$); Factor 1, Females vs. Factor 2, Females ($p = .001$) (Fig. 4). In other words, females performed better on Factor 1 than Factor 2 and better than males on Factor 1, and *vice versa* for males.

Table 5 shows the mean performance for the category components of each factor separately for males and females, along with the *t*-statistic comparing performance across sexes. These results complement the 2 × 2 ANOVA, suggesting a female advantage in performance for categories loading highest on Factor 1, and a male performance advantage for categories loading highest on Factor 2.

One possibility is that this interaction between Sex and Factor can be accounted for by differences in self-reported expertise between sexes. Indeed, when we calculate aggregate self-ratings for each factor, self-reported expertise shows a similar interaction with sex ($F_{1,221} = 18.23$, $MSE = 0.91$, $p < .001$, $\eta^2 = .05$). Because Factor 2 self-report correlated with Factor 2 loadings ($r = .208$, $p = .001$) (although self-report and loadings corresponding to

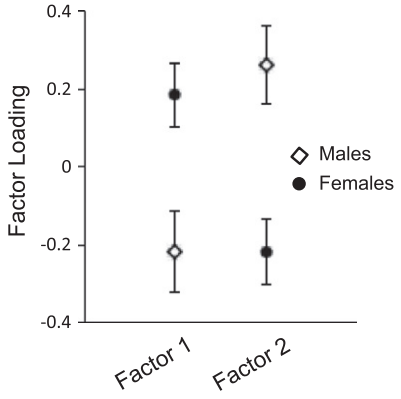

Fig. 4. Interaction effect between Factor (Factor 1, nature factor and Factor 2, vehicle factor) and Sex (Male and Female). Error bars represent the standard error of the mean.

Table 5
Performance for category components of Factors 1 and 2 for males and females.

	Males accuracy, *u* (sem)	Females accuracy, *u* (sem)	*t*-Value (*p*-value)
Factor 1			
Leaves	.57 (.011)	.60 (.01)	2.22 (.03)
Owls	.68 (.012)	.71 (.01)	−1.93 (.05)
Butterflies	.58 (.013)	.64 (.01)	−3.75 (.001)
Wading birds	.60 (.011)	.62 (.01)	−1.19 (n.s.)
Mushrooms	.60 (.012)	.61 (.008)	−1.29 (n.s.)
Factor 2			
Cars	.67 (.016)	.62 (.013)	2.55 (.01)
Planes	.69 (.014)	.66 (.011)	1.82 (n.s.)
Motorcycles	.60 (.012)	.57 (.011)	1.85 (n.s.)

Factor 1 were not correlated), it was critical to examine effects of factor loadings independent of individual differences in perceived expertise. We computed residuals for Factors 1 and 2 for each subject by regressing Self-report Factor 1 out of Factor 1 and Self-report Factor 2 out of Factor 2. We ran a 2×2 ANOVA on Factors 1 and 2 residuals with Sex (Male, Female) as a between-subjects variable and Factors (residualized Factors 1 and 2) as a within-subjects variable. The interaction effect observed with non-residualized factor loadings persists even when we account for the influence of self-report scores ($F_{1,221} = 20.16$, $MSE = 16.81$, $p < .001$, $\eta^2 = .08$).

The results of Experiment 1 demonstrate that the VET has good reliability and captures interesting sex differences suggested by Dennett et al. (2011), who found that men showed an advantage for the recognition of cars on the CCMT. Here, we replicate this finding, showing a male advantage for cars and also trends towards a male advantage for planes and motorcycles (i.e., Factor 1). Dennett et al. reported that sex had an effect on car performance over and above what could be attributed to interest or expertise. We also replicate this finding, but importantly, we can reject one account proposed by these authors for this result: they proposed a domain-general advantage for men in mental rotation since the CCMT, like the VET, requires matching across different views. This account, or any other account that relies on a domain-general advantage for one sex, fails to explain the significant advantage we observed for women with leaves, owls and butterflies. Our ability to refute this account demonstrates an important advantage of using a test with several categories. When a study compares faces to only a single category of objects, differences between faces and objects are typically interpreted as being due to face-specific factors vs. object-general mechanisms. With several object categories, it is possible to tease apart domain-general and domain-specific explanations. Therefore, while we found, like Dennett, that self-reported expertise does not appear to account for the sex differences, we prefer to interpret this result as indicating that self-reports of expertise/interest may be inadequate predictors of perceptual expertise. It is also possible that people with a lot of semantic knowledge about cars may not necessarily be very good at visual recognition of cars, and other individuals may have paid attention to cars and refined their perceptual skills without learning much about cars in other ways (but see Barton, Hanif, and Ashraf (2009) for evidence that a verbal knowledge test about cars that does not rely on images correlates with visual recognition of cars). In addition, people may not have access to much evidence about others' performance with various categories, limiting their ability to estimate their own performance relative to the average population (Barton, Hanif, & Ashraf, 2009). In sum, our results confirm the importance of quantifying perceptual expertise in perceptual tasks that can be performed by novices and experts alike.

4. Experiment 2

Because the VET is a new test, we investigated how this measure of individual variability in object recognition relates to performance on a sequential matching test that has been used extensively to predict both behavioral (Curby, Glazek, & Gauthier, 2009; Gauthier et al., 2003; McGugin & Gauthier, 2010) and neural (Gauthier et al., 2000, 2003; McGugin et al., submitted for publication; Rossion, Kung, & Tarr, 2004; Xu, 2005) effects of expertise. Given the sex effects observed in Experiment 1, we also ask whether the relationship between the two tasks depends on this factor. It is important to note that we are not interested here in the absolute role of sex in the relationship between these tasks, but rather with the fact that many domains of expertise are likely to be of more interest to some groups than others, often correlated with sex.

4.1. Participants

Seventy-six participants (42 female; mean age 23.34 years) completed a perceptual expertise test immediately following the VET. The VET data for a subset of these participants ($n = 26$) was included in Experiment 1, and they also participated in an fMRI experiment on the basis of self-report for either high or low experience with cars (McGugin et al., submitted for publication). The rest of the sample answered an ad posted on an online research system to recruit paid volunteers.

4.2. Materials and procedure

The procedure for the matching task was identical to that reported previously (e.g., Gauthier et al., 2000, 2003, 2005; McGugin & Gauthier, 2010). It included matching trials with cars, planes and passerine birds. Because we were interested in comparing the same categories across different tasks, and since the VET did not include passerine birds but only owls and wading birds, we focus here on the car and plane conditions. Participants made same-different judgments on car and plane images (at the level of make and model, regardless of year). The test included four blocks of 28 sequential matching trials each for cars and planes, using 56 grey-scale images per category. Cars were all relatively recent car models (1997–2003) and planes were modern commercial or military plane models constructed during or after WWII. All images used in this perceptual matching task were different from those employed in the VET. On each trial, the first stimulus appeared for 1000 ms, followed by a 500 ms mask. A second stimulus then appeared until a same/different response was made or 5000 ms elapsed.

Prior to this matching task, participants completed the VET as described in Experiment 1.

4.3. Results

Before we relate individual differences in matching tasks to the VET as a function of sex, we first considered sex differences in self-reports of expertise and performance in both tasks in each category (Supplemental Table 5). Men in this sample rated themselves on average higher than women for cars ($t = 2.16$, $p = .03$), motorcycles ($t = 3.27$, $p = .001$) and planes ($t = 2.68$, $p = .01$). Performance only differed by sex for the matching task for planes ($t = 3.09$, $p = .003$), and the VETcar ($t = 2.07$, $p = .042$) and VETplane ($t = 3.40$, $p = .001$), in all cases better for men.

We then consider how performance on the matching task for planes and cars correlated with performance on the VET for each category, for men and women separately (Table 6). In the matching task, performance for cars and planes was significantly correlated for both men and women, with no difference between these correlations. But the table reveals salient sex differences. First, matching performance for cars or planes was significantly correlated with VET performance for many more categories for men than for women (11 vs. 3 significant correlations), and VETplane and VETmotorcycle performance were significantly more correlated with matching performance for cars and planes for men than for women.

For car matching and for both men and women, the strongest correlation with a VET category was with VETcar, and this correlation was significantly stronger than those with all other VET categories. But for plane matching, while the highest correlation was with VETplane in men, there was no significant correlation with VETplane in women (the plane matching correlation with VETplane for *men* was significantly stronger than the correlation with leaves, owls and mushrooms, while for *women* it was stronger than no other category). At least for men, these correlations suggest that

Table 6
Between-task correlations (matching task d' and VET accuracy) for Male ($N = 34$) and Female ($N = 42$) participants in Experiment 2.

	Males		Females	
	Car d'	Plane d'	Car d'	Plane d'
Car d′				
Plane d'	.555* (<.001)		.389* (.011)	
Bird d'	0.009 (.959)	0.181 (.305)	0.294 (.059)	0.413* (.007)
Leaves	0.313 (.072)	.375* (.029)	0.041 (.796)	0.064 (.688)
Owls	0.31 (.074)	.371* (.031)	0.231 (141)	.460* (.002)
Butterflies	0.326 (.060)	.465* (.006)	0.133 (.400)	0.239 (.127)
Wading birds	.483* (.004)	.508* (.002)	0.122 (.441)	.424* (.005)
Mushrooms	0.137 (.440)	0.265 (.130)	−0.089 (.573)	0.125 (.432)
Cars	.825* (<.001)	.579* (<.001)	.714* (<.001)	0.258 (.099)
Planes	**.617* (<.001)**	**.645* (<.001)**	**0.125 (.431)**	**0.18 (.253)**
Motorcycles	**.546* (<.001)**	**.572* (<.001)**	**0.037 (.816)**	**0.081 (.609)**

* $p < .05$; shaded cells are significantly different between male and female participants at $p < .05$.

if there is a domain-specific relationship for cars or planes across the two tasks, it is also accompanied by domain-general effects. Thus, to assess the domain-specific and domain-general effects for each category and each sex, we turned to multiple regression.

We conducted multiple regression analyses with performance on car matching or on plane matching as the dependent variable, entering all predictors simultaneously, including *z*-transformed measures of VET performance for the same category as matching (i.e., VETcar for the regression on car matching), and VET performance for all other categories combined (i.e. all non-car categories for the regression on car matching). Note that we did not split VET categories by factor here, because without planes or cars the two factors would be too unbalanced in the number of categories: we therefore chose to use an aggregate of all other categories as an estimate of domain general ability. Because Experiment 1 revealed an interaction between Sex and VET Factor loadings, we also included sex (dummy coded) and the interaction between sex and each predictor. We also did not include age, since preliminary analyses showed that matching performance was not correlated with age for either cars or planes, for either sex.

When predicting Car Matching, VETcar performance was a significant predictor, independent of all other predictors, which were not significant (Table 7).

Not only was Sex *not* significant, but the same relationship was found when multiple regressions were conducted separately for males and females (Table 8).

This stands in contrast with the results of the Regression on Plane Matching. When predicting Plane Matching, VETplane performance was not a significant predictor, but there were significant effects of Sex and of VETall_except_Plane. In addition, the interaction between Sex and VETplane was near significant ($p = .055$) and

Table 7
Results of multiple regression analysis.

Model and predictor	*B*	*SE*	*t*	*p*
Car d′ (R^2 adjusted = 59.2%)				
Intercept	1.41789	0.05	31.10	<.001
Sex	−0.00844	0.05	−0.19	.854
VET car	0.481537	0.05	8.90	<.001
VET (all except cars)	−0.04006	0.05	−0.75	.457
Sex × VET car	−0.00050	0.05	−0.01	.993
Sex × VET (all except cars)	−0.02389	−0.05	−0.45	.657
Plane d′ (R^2 adjusted = 35.3%)				
Intercept	1.38599	0.05	27.00	<.001
Sex	−0.10919	0.05	−2.13	.037
VET Plane	0.081331	0.08	1.08	.285
VET (all except planes)	0.176116	0.07	2.47	.016
Sex × VET plane	−0.14737	0.08	−1.95	.055
Sex × VET (all except cars)	0.057999	0.07	0.81	.419

Table 8
Results of separate multiple regression analyses for male and female participants.

Model and predictor	*B*	*SE*	*t*	*p*
Car d′, males (R^2 adjusted = 66.1%)				
Intercept	1.42632	0.07	21.80	<.001
VET car	0.48204	0.08	6.26	<.001
VET (all except cars)	−0.01617	0.08	−0.21	.835
Car d′, females (R^2 adjusted = 49.5%)				
Intercept	1.40945	0.06	22.50	<.001
VET car	0.481039	0.08	6.39	<.001
VET (all except cars)	−0.06395	0.07	−0.87	.391
Plane d′, males (R^2 adjusted = 40.3%)				
Intercept	1.49518	0.08	19.00	<.001
VET plane	0.2287	0.11	1.99	.055
VET (all except planes)	0.118117	0.10	1.15	.259
Plane d′, females (R^2 adjusted = 10.9%)				
Intercept	1.27679	0.07	19.10	<.001
VET plane	−0.06604	0.10	−0.67	.509
VET (all except planes)	0.234115	0.10	2.35	.024

because we were *a priori* interested in sex differences, we followed-up on this analysis with the multiple regressions for each sex separately (Table 8). For men, the results predicting Plane matching were very similar to those predicting Car matching. VETplane was the strongest predictor, and although its contribution independent of VETall_except_plane was not quite significant (again, $p = .055$), both predictors together (VETplane and VETall_except_plane) accounted for 40% of the variance in Plane Matching. Thus, the pattern is qualitatively the same as for cars, with evidence that Plane matching depends on domain-specific variance. For women however, the pattern of results is very different. First, VET performance (Plane and All_except _plane) together account for only 11% of the Variance in Plane matching, and in this case the domain-general VETall_except_plane is the strongest predictor, with little evidence of a domain-specific contribution. Note that these differences between men and women for Plane matching and not Car matching do not appear to be due to a restriction of range specifically for women with planes. While plane d' is poorer for women than men (Supplemental Table 5), range is actually larger for women than men. For VETplane the range is less for women than for men (.60 vs. .71, note that the sample sizes are different) but the range is very similar for women with VETplane and VETcar (.62 vs. .60). Instead, we provide an admittedly speculative, but perhaps more interesting, interpretation of this difference below. Before doing so we note that Experiment 2 provides convergent validity for the VET, at least for a sex-congruent category of expertise (cars or planes in males), by relating it to a sequential measure of expertise that has proven useful in several prior studies.

Coming back to the interpretation of the entire pattern of results, including the sex difference, we believe the results provide evidence for the ability of the VET to separate domain-specific and domain-general contributions to perceptual expertise. In studies of perceptual expertise with real world-experts, perceptual expertise is generally defined as performance on a perceptual task (e.g., Gauthier et al., 2000; Harel et al., 2010; Xu, 2005). Because in such cases we do not measure experience *per se*, we do not know to what extent performance is due to experience with this domain vs. other factors. Cars is a category for which both men and women are likely to have significant exposure, even though it is clearly possible for someone with an interest in cars to seek more exposure. The results of our analyses suggest that both men and women who perform particularly well with cars do so because of domain-specific variance, which can reasonably be attributed to experience. A car expert is likely to have more car-responsive neurons in their visual system (Gauthier et al., 2000; McGugin et al., submitted for publication), which may be associated with domain-specific but task-general advantages for cars. Likewise, because the results for plane matching are almost identical to those for car matching in men, we can surmise that men who perform well with planes do so because of domain-specific practice. In contrast, it is possible that the women who do best with planes might be better at applying domain-general strategies, which is why a VET aggregate score for several categories was a better predictor of plane matching performance than the score for planes alone. Because the matching task and the VET are somewhat different, this domain-general strategic contribution accounts for only 11% of the variance in Plane Matching.

Admittedly this is a post hoc explanation, but the results at least reveal the benefit of using several categories as a measure of perceptual expertise. Regressing out performance for all categories apart from the category of interest is an improvement over only measuring performance in the domain of interest or comparing it to a single control category. Importantly, as shown for cars, this approach appears to work well to capture domain-specific variance in a behavioral task, regardless of whether this task correlated with performance for several other categories (as it did for men) or not (as it did not for women). Furthermore, the results for plane matching converge with Experiment 1 to demonstrate that sex needs to be taken into account when interpreting individual differences across object categories.

The main advantage of the VET over the standard matching task used as a measure of expertise is the number of control categories. Future work could compare the VET and perceptual matching tasks for the same number of categories on how well they predict a target criterion, such as FFA activity for various objects. Until then, it should not be assumed that the memory-related learning component of the VET presents an advantage over a more perceptual matching task. However, our interpretation of the sex differences builds on the assumption that measuring individual differences across a variety of tasks may be another way to better isolate domain-specific effects from domain-general, task-specific strategies.

5. Experiment 3

In Experiment 3, we had a subset of the participants from Experiment 1 (non-overlapping with those in Experiment 2) also perform a sequential matching composite task with faces to measure holistic face processing. Holistic processing is operationalized in this task as a failure to selectively attend to one half (e.g., top) of a face when the face halves are aligned relative to misaligned (e.g., Richler, Cheung, & Gauthier, 2011b; Richler et al., 2008). Theoretically, given the central role that holistic processing has played in the literature on face recognition, one would expect that people who process faces holistically experience an advantage in recognizing faces. Recent work (Konar, Bennett, & Sekuler, 2010; Richler, Cheung, & Gauthier, 2011b; Wang et al., in press[3]) has failed to support this relationship between holistic processing and face recognition skill when holistic processing is measured with one version of the composite task called the *partial design*, in which the parts of face composites that are to be ignored are always different, leading to complicating confounds from response bias (Cheung et al., 2008; Richler, Cheung, & Gauthier, 2011a; Richler et al., 2011). Richler, Cheung, and Gauthier (2011b) also tested a different version of the composite task, called the *complete design*, which provides a measure of holistic processing that is robust to response biases (Cheung et al., 2008; Richler, Cheung, & Gauthier, 2011a) and found that it correlated with performance on the CFMT. The correlation was of moderate strength, however ($r = .4$), and given that performance on the CFMT shows significant relationships with all VET categories in Experiment 1, it is possible that holistic processing of faces is mainly accounted for by domain-general object-recognition variance. Thus, one goal of Experiment 3 is to test whether holistic face processing (measured with the complete design of the composite task) predicts face recognition ability independent of object recognition ability (captured by the VET). In other words, does holistic processing of faces account for face-specific recognition abilities?

We entered all predictors simultaneously in the multiple regression to assess their independent contributions, and examined the contribution to face recognition of both VET factors, and their interaction with sex. Thus, a second goal of Experiment 3 was to examine whether face recognition and object recognition are independent, as was either suggested or assumed by previous work where only one non-face object category was used (e.g., Wilhelm et al., 2010; Wilmer et al., 2010; Zhu & et al., 2010). In other words, is some portion of face recognition ability accounted for by more general object recognition abilities, and more importantly does this relationship depend on sex?

5.1. Methods

5.1.1. Participants

One hundred and nine Caucasian individuals (61 female; mean age 22.05 years) received a small honorarium for participation. All participants had normal or corrected-to-normal visual acuity. The experiment was approved by the Institutional Review Board at Vanderbilt University, and all participants provided written informed consent. Participants completed three tasks in the following order: composite task, Cambridge Face Memory Test (CFMT), Vanderbilt Expertise Test (VET) (the latter two as part of Experiment 1).

5.1.2. Composite task

Stimuli in the composite task were images of twenty female faces from the Max Planck Institute Database (Troje & Buthoff, 1996) converted to gray-scale and cut in half to produce 20 face top halves and 20 face bottom halves, each 256×128 pixels in size. Face halves were randomly combined to create composite faces. A white line, 3 pixels thick, separated face halves resulting in faces that were 256×259 pixels. The white line was added to make it unambiguous where the top half ends and the bottom half begins, which, if anything, should facilitate selective attention to one half. Misaligned faces were created by moving the top half of the face to the right by 35 pixels, and the bottom half of the face

[3] While the authors of this study argued a correlation was found with performance on a face task when a baseline with objects was subtracted, the individual correlations with the partial composite design measure of holistic processing revealed no correlation with face recognition but a small but negative correlation with the object task.

to the left by 35 pixels, such that the edge of one face half fell in the center of the other face half.

On each of 160 trials, a fixation cross was presented (200 ms), followed by the study face (200 ms). The test face was then presented following a 500 ms ISI for 200 ms. Participants were instructed to judge whether the top half of the test face was the same as or different from the top half of the study face while ignoring the irrelevant bottom half. Participants had a maximum of 2500 ms to respond. Time-outs were rare (<1% of trials), and excluded from our analyses. The study face was always aligned. The test face could be either aligned or misaligned.

A trial sequence that contained 20 trials for each combination of congruency (congruent/incongruent), alignment (aligned/misaligned) and correct response (same/different) was randomly generated, and the same trial sequence was used for all participants. The experimental block was preceded by a 16-trial practice block.

5.1.3. Results

Data from one participant were discarded according to the criterion in Experiment 1 (reaction times in the VET below 200 ms on more than 40% of the trials in at least 3 out of eight object categories). Avoiding ceiling effects is particularly important in individual differences analyses, and this particular sample had more participants at ceiling than a previous study (Richler, Cheung, & Gauthier, 2011b). In the composite task this can be especially problematic as the measure of holistic processing is a difference of differences: if participants are at ceiling in any of the cells of the design, the magnitude of the difference of differences is artificially limited. Therefore, we discarded data from 42 participants who had average accuracy on the composite task greater than 90%, resulting in a total of 66 participants (34 female; mean age 22.03 years) in the analyses.[4] In the CFMT, we avoided ceiling effects without further rejecting participants by considering performance on noise trials only, which are more difficult than the no-noise trials. Importantly, performance on the noise trials is highly correlated with performance on the no-noise trials (r_{66} = .768, p < .001), and the results are qualitatively the same when we use all CFMT trials vs. noise trials only. Average overall accuracy on the CFMT was 77.53% (SD = 12.57), and average performance on the CFMT noise trials was 62.25% (SD = 18.59). Average overall accuracy on the VET was 60.97% (SD = 6.62).

A 2 × 2 repeated-measures ANOVA on d' in the composite task with factors congruency (congruent, incongruent) and alignment (aligned, misaligned) revealed a main effect of alignment ($F_{1,65}$ = 7.23, MSE = 2.37, p < .01, η^2 = .03), a main effect of congruency ($F_{1,65}$ = 64.01, MSE = 2.09, p < .001, η^2 = .19) and a significant congruency × alignment interaction ($F_{1,65}$ = 32.20, MSE = 2.11, p < .001, η^2 = .10). At the group-level, participants in our sample showed evidence of holistic processing: performance was better on congruent vs. incongruent trials, and this difference was reduced when face halves were misaligned.

5.1.4. Zero-order correlations

For each participant, holistic face processing (HP) is indexed by the magnitude of the congruency × alignment interaction in d' [(aligned congruent − aligned incongruent) − (misaligned congruent − misaligned incongruent)] in the composite task. Average performance on the CFMT noise trials provides a measure of face recognition, and Factors 1 and 2 loadings from the VET provide measures of object recognition.

[4] Confirming our intuition that ceiling effects in the composite task can be problematic because holistic face processing is operationally defined as a difference of differences, the critical correlation between the CFMT (noise trials) and holistic face processing is not significant when these participants are included in the analyses (r_{108} = .126, p = .197).

Table 9
Zero-order correlations between measures. 95% confidence intervals are shown in brackets.

	HP	VET-F1	VET-F2	Age
CFMT	.256* (.015, .468)	0.15 (−.095, .378)	0.091 (−.154, .325)	.287* (.049, .494)
HP		−0.075 (−.311, .170)	−0.173 (−.398, .072)	−0.019 (−.259, .224)
VET-F1			−0.063 (−.300, .181)	<.000 (−.242, .242)
VET-F2				0.151 (−.094, .379)

* p < .05.

The zero-order correlations between all our variables of interest are shown in Table 9.

5.1.5. Multiple regression

We conducted a multiple regression analysis (Table 10) with performance on the CFMT (noise trials) as the dependent variable, entering all predictors simultaneously, including z-transformed measures of holistic face processing, object recognition (VET Factors 1 and 2), and age. Because Experiment 1 revealed an interaction between Sex and VET Factor loadings, we also included sex and the interaction between sex and each predictor. Holistic face processing, age, and VET Factor 2 × Sex were found to be significant predictors of face recognition. There was a trend for VET Factor 1 × Sex to also be a significant predictor.

5.1.6. Partial correlations

Partial correlations between each significant predictor and performance on the CFMT noise trials are shown in Fig. 5. In summary, the results of Experiment 3 replicate Richler, Cheung, and Gauthier (2011b) in finding that holistic face processing predicts face recognition performance, and demonstrates that this relationship survives even after age, sex, and most importantly performance on the VET have been factored out. This validates the central importance of holistic processing in face recognition, in contrast to earlier claims (Konar, Bennett, & Sekuler, 2010). Age was also a significant predictor of face recognition abilities, consistent with the finding that face recognition abilities improve until 30 years of age (Germine, Duchaine, & Nakayama, 2010).

Interestingly, we found that general object recognition ability was correlated with face recognition performance but only when sex-congruent categories are used. One possibility is that sex-congruent categories best correlate with face recognition because of an underlying potential that is domain-general but which becomes expressed in domain-specific skills through experience (see Section 6).

Table 10
Results of multiple regression analysis.

Model and predictor	*B*	*SE*	*t*	*p*
CFMT (R^2 adjusted = 17.9%)				
Intercept	59.0070	2.38	24.8	<.001
Holistic face processing	5.76156	2.17	2.65	.010
VET Factor 1	2.55556	2.22	1.15	.255
VET Factor 2	3.90197	2.48	1.57	.121
Age	5.17035	2.34	2.21	.031
Sex	1.55148	2.38	0.65	.518
Holistic Face Processing × Sex	−.892232	2.17	−0.41	.683
VET Factor 1 × Sex	3.74840	2.22	1.69	.097
VET Factor 2 × Sex	−5.61263	2.48	2.26	.028
Age × Sex	0.943705	2.34	0.40	.683

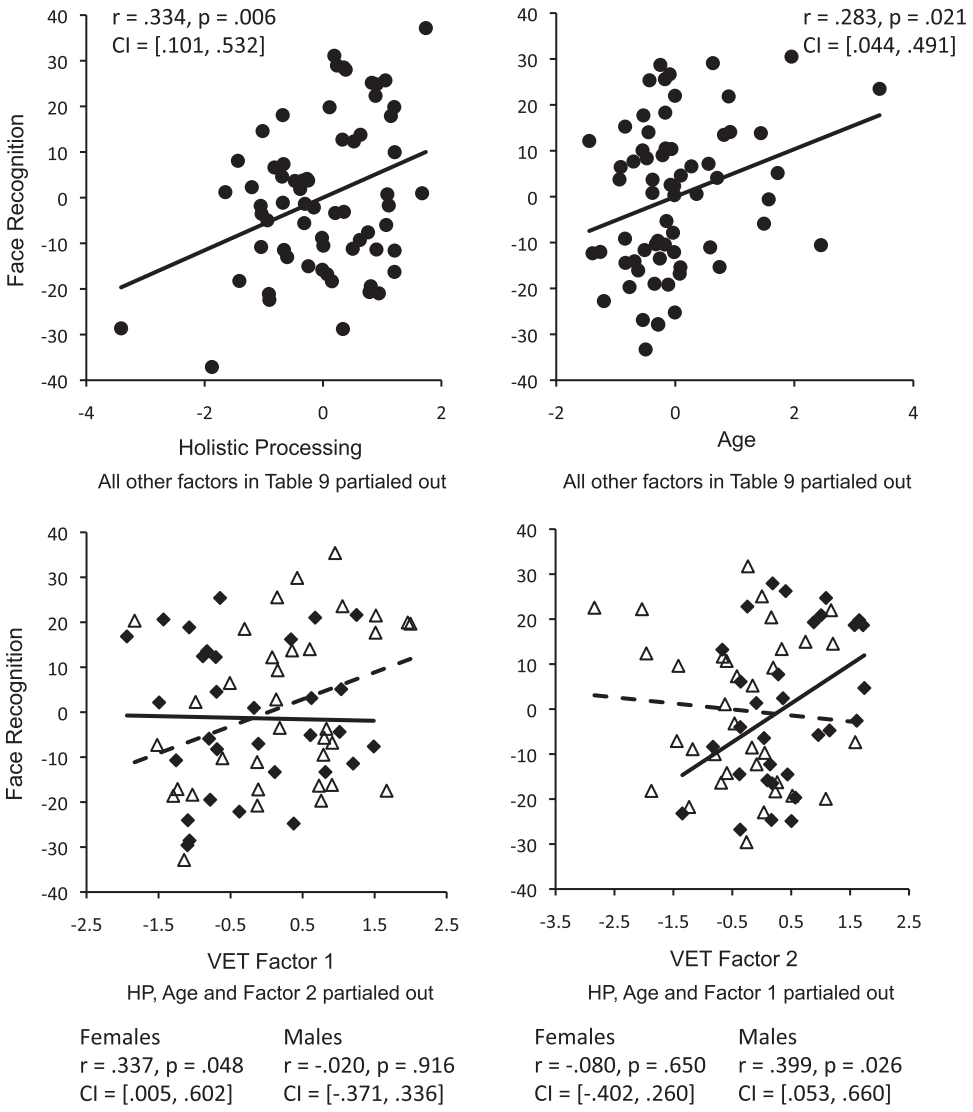

Fig. 5. Partial correlations and 95% confidence intervals between face recognition (accuracy on CFMT noise trials) and holistic face processing (congruency × alignment interaction in d'), age, and object recognition (VET Factor 1 and VET Factor 2). In the bottom two plots, triangles and dashed lines show datapoints and the trendline for female participants, and diamonds and solid lines show datapoints and the trendline for male participants.

6. General discussion

We offer a new test of object recognition, the VET, which provides reliable measures of object recognition similar to the CFMT for eight different object categories. Our results demonstrate important advantages of using several object categories. The VET allows one to quantify perceptual expertise with specific object categories while the remaining categories can be used to assess more general object recognition skills. And while with only a single category one group could appear superior at object recognition (for instance men in Dennett et al. (2011)), with many categories different domain-specific advantages can emerge for each group, leading to a very different interpretation.

Our results clearly question common wisdom whereby any effect that is similar for faces and another non-face object category is presumed to capture domain-general variance that would apply to any other object category. For one thing, as shown in Experiment 2, within the realm of non-face objects it is possible to dissociate domain-specific from domain-general influences. In addition, In Experiment 1 the VET captured two factors associated with opposite sex advantages. Perhaps more critically, not only were men better than women on average on Factor 2, but in Experiment 3 Factor 2 was more strongly related to face recognition ability for men than women. While not significant, the same trend was observed for Factor 1 and women. This result has important implications, suggesting that using a single object category could be even more problematic when a sample contains individuals from both sexes, because while object categories that are not sex-congruent (below, we address how this is interpreted) may not correlate with face recognition, those that are sex-congruent can reveal this relationship. Any analysis that fails to take this interaction into account may overstate the independence of face and object recognition abilities.

It is not clear what drives the observed sex effects for Factors 1 and 2. In a meta-analysis of patients with category-specific semantic disorders, Gainotti (2005) found that men are usually more impaired with plants whereas women are more impaired with animals. Based on the fact that this distinction was not associated with different foci of lesion, whereas the distinction between deficits for living vs. non-living categories does map onto relative

ventral vs. dorsal lesions, this author argued that the sex effect for plants vs. animals was due to familiarity. One study that used semantic fluency as opposed to naming tasks, which can suffer from ceiling effects, found that age and sex interacted to predict performance (Moreno-Martinez, Laws, & Schulz, 2008). Younger adults showed no significant sex differences but elderly females had better fluency for flowers, vegetables and kitchen utensils and elderly males showed better fluency for musical instruments. These results suggest familiarity effects in line with gender roles. However, other authors have proposed a more evolutionary account for most of these effects (Barbarotto et al., 2002; Laiacona, Barbarotto, & Capitani, 2006). Importantly, although there is a literature on sex by category interactions in naming and fluency tasks, our results extend this interaction to visual recognition skills.

Interestingly, like Dennett et al. (2011), we find that self-reported expertise or familiarity does not account for these effects, but our interpretation of this result differs. That is, the fact that women show an advantage for living categories while men show an advantage for cars appears to suggest that stereotypical interests may play a role in these effects. Dennett et al. ruled out this explanation in favor of a general male advantage in mental rotation, but this can be excluded here by the use of other categories for which women show an advantage.[5] We suggest that these self-reported measures of expertise or interest may often provide relatively poor predictors of perceptual performance, as measured by the VET or a matching task (see also Barton, Hanif, & Ashraf, 2009). It is clear from Experiment 1 that people are often poor judges of how they rank relative to the general population on recognition ability in specific domains. There is a parallel for this discrepancy in the literature on sex-by-category interactions in the naming performance of normal subjects for living and non-living things. Females are slower than males to name nonliving things and males slower to name living things (McKenna & Parry, 1994), but this has not been found to be accounted for by either conceptual or visual familiarity ratings (Laws, 1999).

One interpretation of such findings is that we need to look outside of experience to understand these sex differences, for instance to evolutionary influences, but another account is that what self-ratings of familiarity fail to capture is the quality of one's perceptual experience with object categories. Several studies indicate that it is not exposure to a category, but the kind of experience with it, that determines perceptual expertise (McGugin et al., 2011; Tanaka, Curran, & Sheinberg, 2005; Wong, Palmeri, & Gauthier, 2009). We propose that performance for sex-congruent categories best correlates with face recognition because of an underlying potential that is domain-general but which becomes expressed in domain-specific skills through experience. Assuming that both men and women experience considerable pressure to develop face recognition skills and experience constant opportunities to practice this skill, face recognition performance is likely to express each individual's potential at object recognition. Likewise, men could have more motivation and opportunity to individuate cars, and the same may be true for women and natural categories.

It cannot be assumed that the only influences on object recognition are associated with sex. Separate factors may influence interest and experience with other categories, such as age, culture, occupation etc. It may also not be prudent to extrapolate predictions to other categories such as other living things or other vehicles, since there were non-trivial differences within the categories for each factor, such as between cars and planes or leaves and mushrooms. At the minimum, it is clear that no single category can stand for a general construct of "object recognition"; it is always possible that a new object category not tested in the VET (e.g., shoes, vegetables, trains, etc.) would not fit well within the two factors we have uncovered here. In other words, when it comes to individual differences at least, the comparison of face recognition to object recognition may not be valid unless a sufficient number of categories are tested such that a true domain-general latent factor can be extracted. The very use of the term "non-face recognition" in the literature illustrates a bias to assume homogeneity among non-face categories; it may be no more useful than the construct of "non-mushroom recognition". Thus, any conclusion that face recognition is independent from object recognition based on a single control domain is fundamentally limited (e.g., Wilhelm et al., 2010; Wilmer et al., 2010; Zhu et al., 2010).

[5] In addition, there are other reasons to doubt that men are better at recognizing cars due to an advantage for mental rotation, given the dissociation of the systems involved in mental rotation and object recognition (e.g., Gauthier et al., 2003; Hayward, Zhou, Gauthier, & Harris, 2006).

Acknowledgments

This work was supported by NIH Grants 2 RO1 EY013441-06A2 and P30-EY008126 and by the Temporal Dynamics of Learning Center, NSF Grant SBE-0542013. We would like to thank Lisa Weinberg for assistance in data collection and Brad Mahon for helpful guidance.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.visres.2012.07.014.

References

Barbarotto, R., Laiacona, M., Macchi, V., & Capitani, E. (2002). Picture reality decision, semantic categories and gender: A new set of pictures with norms and an experimental study. *Neuropsychologia, 40*, 1637–1653.

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

Bowles, D. C., McKone, E., Dawel, A., Duchaine, B., Palermo, R., Schmalzl, L., et al. (2009). Diagnosing prosopagnosia: Effects of ageing, sex, and participant stimulus ethnic match on the Cambridge Face Memory Test and Cambridge Face Perception Test. *Cognitive Neuropsychology, 25*, 423–455.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433–436.

Bukach, C. M., Phillips, S. W., & Gauthier, I. (2010). Limits of generalization between categories and implications for theories of category specificity. *Attention, Performance and Psychophysics*, 1865–1874.

Carey, S. (1992). Becoming a face expert. *Philosophical Transcripts of the Royal Society London, 335*, 95–102.

Cheung, O. S., Richler, J. J., Palmeri, T. J., & Gauthier, I. (2008). Revisiting the role of spatial frequencies in the holistic processing of faces. *Journal of Experimental Psychology: Human Perception and Performance, 34*, 1327–1336.

Curby, K. M., Glazek, K., & Gauthier, I. (2009). A visual short-term memory advantage for objects of expertise. *Journal of Experimental Psychology: Human Perception and Performance, 35*, 94–107.

Dennett, H. W., McKone, E., Tavashmi, R., Hall, A., Pidcock, M., Edwards, M., et al. (2011). The Cambridge Car Memory Test: A task matched in format to the Cambridge Face Memory Test, with norms, reliability, sex differences, dissociations from face memory, and expertise effects. *Behavioral Research*. http://dx.doi.org/10.3758/s13428-011-0160-2.

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

Duchaine, B., Germine, L., & Nakayama, K. (2007). Family resemblance. ten familymembers with prosopagnosia and within-class object agnosia. *Cognitive Neuroscience, 24*, 419–430.

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*, 576–585.

Duchaine, B., Wendt, T. N., New, J., & Kulomaki, T. (2003). Dissociations of visual recognition in a developmental agnostic: Evidence for separate developmental processes. *Neurocase, 9*, 380–389.

Duchaine, B., Yovel, G., Butterworth, E. J., & Nakayama, K. (2006). Prosopagnosia as an impairment to face-specific mechanisms: Elimination of the alternative hypotheses in a developmental case. *Cognitive Neuroscience, 23*, 714–747.

Gainotti, G. (2005). The influence of gender and lesion location on naming disorders for animals, plants and artefacts. *Neuropsychologia, 43*, 1633–1644.

Gauthier, I., Curby, K. M., Skudlarski, P., & Epstein, R. A. (2005). Individual differences in FFA activity suggest independent processing at different spatial scales. *Cognitive, Affective and Behavioral Neuroscience, 5*, 222–234.

Gauthier, I., Curran, T., Curby, K. M., & Collins, D. (2003). Perceptual interference supports a non-modular account of face processing. *Nature Neuroscience, 6*, 428–432.

Gauthier, I., & Nelson, C. (2001). The development of face expertise. *Current Opinion in Neurobiology, 11*, 219–224.

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience, 3*, 191–197.

Gauthier, I., & Tarr, M. J. (1997). Becoming a ''greeble'' expert: Exploring mechanisms for face recognition. *Vision Research, 37*, 1673–1682.

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (1999). Activation of middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nature, 2*, 568–573.

Germine, L., Cashdollar, N., Duzel, E., & Duchaine, B. (2011). A new selective developmental deficit: Impaired object recognition with normal face recognition. *Cortex, 47*, 598–607.

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

Harel, A., Gilaie-Dotan, S., Malach, R., & Bentin, S. (2010). Top-down engagement modulates the neural expressions of visual expertise. *Cerebral Cortex, 20*, 2304–2318.

Hayward, W. G., Zhou, G., Gauthier, I., & Harris, I. (2006). Dissociating viewpoint costs in mental rotation and object recognition. *Psychonomic Bulletin & Review, 13*, 820–825.

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

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17*, 4302–4311.

Konar, Y., Bennett, P. J., & Sekuler, A. B. (2010). Holistic processing is not correlated with face-identification accuracy. *Psychological Science, 21*, 38–43.

Laiacona, M., Barbarotto, R., & Capitani, E. (2006). Human evolution and the brain representation of semantic knowledge: Is there a role for sex differences? *Evolution and Human Behavior, 27*, 158–168.

Laws, K. R. (1999). Gender affects latencies for naming living and nonliving things. *Cortex, 35*, 729–733.

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Labels facilitate learning of novel categories. *Pyschological Science, 18*, 1077–1083.

McGugin, R. W., Gatenby, J. C., Gore, J. C. & Gauthier, I. (submitted for publication). High-resolution imaging of expertise reveals reliable object selectivity in the FFA related to perceptual performance.

McGugin, R. W., & Gauthier, I. (2010). Perceptual expertise with objects predicts another hallmark of face perception. *Journal of Vision, 10*, 1–12.

McGugin, R. W., McKeeff, T. J., Tong, F., & Gauthier, I. (2010). Irrelevant objects of expertise compete with faces during visual search. *Attention, Perception & Psychophysics, 73*, 309–317.

McGugin, R. W., Tanaka, J. W., Lebrecht, S., Tarr, M. J., & Gauthier, I. (2011). Race specific perceptual discrimination improvement following short individuation training with faces. *Cognitive Science, 35*, 330–347.

McKeeff, T. J., McGugin, R. W., Tong, F., & Gauthier, I. (2010). Expertise increases the functional overlap between face and object perception. *Cognition, 117*, 355–360.

McKenna, P., & Parry, R. (1994). Category-specificity in the naming of natural and man-made objects: Normative data from adults and children. *Neuropsychological Rehabilitation, 4*, 255–281.

Moreno-Martinez, F. J., Laws, K. R., & Schulz, J. (2008). The impact of dementia, age and sex on category fluency: greater deficits in women with Alzheimer's disease. *Cortex, 44*, 1256–1264.

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 Psychology, 9*, 555–604.

Pelli, D. G. (1997). The VideoToolbox software of visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437–442.

Rhodes, G., Byatt, G., Michie, P. T., & Puce, A. (2004). Is the fusiform face area specialized for faces, individuation, or expert individuation? *Journal of Cognitive Neuroscience, 16*, 189–203.

Richler, J. J., Cheung, O. S., & Gauthier, I. (2011a). Beliefs alter holistic face processing...if Response Bias is Not Taken into Account. *Journal of Vision, 11*, 1–13.

Richler, J. J., Cheung, O. S., & Gauthier, I. (2011b). Holistic processing predicts face recognition. *Psychological Science, 22*, 464–471.

Richler, J. J., Mack, M. L., Palmeri, T. J., & Gauthier, I. (2011). Inverted faces are (eventually) processed holistically. *Vision Research, 51*, 333–342.

Richler, J. J., Tanaka, J. W., Brown, D. D., & Gauthier, I. (2008). Why does selective attention to parts fail in face processing? *Journal of Experimental Psychology: Learning, Memory and Cognition, 34*, 1356–1368.

Rossion, B., & Curran, T. (2010). Visual expertise with pictures of cars correlates with RT magnitude of the car inversion effect. *Perception, 39*, 173–183.

Rossion, B., & Gauthier, I. (2002). How does the brain process upright and inverted faces? Behavioral and cognitive neuroscience reviews. *Developmental Psychobiology, 1*, 63–75.

Rossion, B., Kung, C.-C., & Tarr, M. J. (2004). Visual expertise with nonface objects leads to competition with the early perceptual processing of faces in the human occipitotemporal cortex. *Proceedings of the National Academy of Science USA, 101*, 14521–14526.

Scott, L. S., Tanaka, J. W., Sheinberg, D. L., & Curran, T. (2006). A re-evaluation of the electrophysiological correlates of expert object processing. *Journal of Cognitive Neuroscience, 18*, 1453–1465.

Tanaka, J. W., Curran, T., & Sheinberg, D. L. (2005). The training and transfer of real world perceptual expertise. *Psychological Science, 16*, 145–151.

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology, 46*, 225–245.

Thompson, P. (1980). Margaret Thatcher: A new illusion. *Perception, 9*, 483–484.

Troje, N. F., & Buthoff, H. H. (1996). How is bilateral symmetry of human faces used for recognition of novel views? *Vision Research, 38*, 79–89.

Wang, R., Li, J., Fang, H., Tian, M., & Liu, J. (in press). Holistic processing and face recognition. *Psychological Science*. http://dx.doi.org/10.1177/0956797611420575.

Wilhelm, O., Herzmann, G., Kunina, O., Danthiir, V., Schacht, A., & Sommer, W. (2010). Individual differences in perceiving and recognizing faces—One element of social cognition. *Journal of Personality and Social Psychology, 99*, 530–548.

Williams, N. R., Willenbockel, V., & Gauthier, I. (2009). Sensitivty to spatial frequency and orientation content is not specific to face perception. *Vision Research, 49*, 2353–2362.

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

Wong, A. C., Palmeri, T. J., & Gauthier, I. (2009). Conditions for facelike expertise with objects: Becoming a Ziggerin expert – but which type? *Psychological Science, 20*, 1108–1117.

Wong, Y. K., Twedt, E., Sheinberg, D., & Gauthier, I. (2010). Does Thompson's Thatcher effect reflect a face-specific mechanism? *Perception*, 1125–1141.

Woolley, A. W., Gerbasi, M. E., Chabris, C. F., Kosslyn, S. M., & Hackman, J. R. (2008). Bringing in the experts: How team ability composition and collaborative planning jointly shape analytic effectiveness. *Small Group Research, 39*, 352–371.

Xu, Y. (2005). Revisiting the role of the fusiform face area in visual expertise. *Cerebral Cortex, 15*, 1234–1242.

Young, A. W., Hellawell, D., & Hay, D. (1987). Configural information in face perception. *Perception, 10*, 747–759.

Zhu, Q. et al. (2010). Heritability of the specific cognitive ability of face perception. *Current Biology, 20*, 1–6.