In press vision research

How faces (and cars) may become special

Mackenzie A. Sunday[1]*, Michael D. Dodd[2], Andrew J. Tomarken[1] & Isabel Gauthier[1]

[1]Vanderbilt University, Department of Psychology

[2]University of Nebraska-Lincoln, Department of Psychology and Center for Brain, Biology and Behavior

*Correspondence to:
Mackenzie Sunday
Department of Psychology
Vanderbilt University
226 Wilson Hall
Nashville, TN 37204
USA
Email: mackenzie.a.sunday@vanderbilt.edu

Keywords: face recognition, object recognition, experience

Abstract:

Recent reports have shown that individuals from small hometowns show relatively poor face recognition ability as measured by the Cambridge Face Memory Test or CFMT (Balas & Saville, 2015; 2017), suggesting that the number of faces present in an individual's visual environment relates to that individual's face recognition ability. We replicate this finding in a sample from a different region (Nebraska) and with more variable age distribution. We extend the study by using another test of face recognition ability that does not require learning over trials, and with non-face object recognition tests that share the learning format with the CFMT. We find no hometown effect in these other tests, although more power would be required to show the CFMT effect is significantly larger. We use the same dataset to explore whether experience with more faces and cars in larger hometowns leads to specialization of these abilities. We find strong and substantial support for the hypothesis that the recognition abilities for faces and for cars are more independent from general object recognition in people from larger hometowns. This suggests that experience may be critical to the specialization of these abilities.

People differ in how well they can recognize faces and objects (Dennett et al., 2012; Duchaine & Nakayama, 2006; McGugin, Richler, Herzmann, Speegle, & Gauthier, 2012; Richler et al., n.d.; Russell, Duchaine, & Nakayama, 2009). What drives these individual differences remains an unanswered question. There appears to be a strong genetic influence on these abilities (Shakeshaft & Plomin, 2015; Wilmer et al., 2010) and correlational studies also suggest an influence of life experience (Gauthier et al., 2014; Ryan & Gauthier, 2016; Tanaka, Kiefer, & Bukach, 2004). In particular, experience stemming from interest in certain domains relates to recognition abilities (Gauthier et al., 2014; Ryan & Gauthier, 2016) and experience due to categories present in one's environment also relates to recognition abilities (e.g., the other-race effect, De Heering, De Liedekerke, Deboni, & Rossion, 2010; Sangrigoli, Pallier, Argenti, Ventureyra, & De Schonen, 2005).

Aside from the distribution of different kinds of faces one experiences or the interest one may have individuating objects from various categories, recent work suggests that the number of exemplars in a category—in this case, faces—within one's environment might also impact recognition ability (Balas & Saville, 2015). Balas & Saville (2015) reported that people from larger hometowns (those with higher population densities) performed better on a measure of face recognition than those from smaller hometowns, a difference the researchers attributed to the fact that those from less dense hometowns likely encounter fewer faces during their childhood. Small hometown individuals would have grown up basing face recognition judgments on a smaller "face space" (Valentine, 1991) relative to people from larger towns, which the authors suggest could impair recognition. The result

also aligns with exemplar models wherein performance gains due to automaticity arise from accumulations of exposures in a given task and domain (Logan, 1988; Palmeri, 1997). Though an important finding, the original Balas & Saville (2015) demonstration did not test a specific explanation for the phenomenon, and was limited in a few important ways. First, there was no behavioral test with non-face objects to determine whether hometown size influences faces specifically, or extends beyond faces into other object domains. Second, only one type of face learning task was used (the CFMT), leaving open the question of whether performance on other face tasks would be similarly impacted. A more recent study provides some evidence that this face recognition advantage may not extend to all tasks with face stimuli since Balas & Saville replicated the relative deficit on the CFMT but found no difference between groups in a card-sorting task with faces and bodies (Balas & Saville, 2017). Finally, in the 2015 paper, an effect of hometown population density (hereafter, HPD) was observed on the face-selective N170 ERP, but the effect was entirely accounted for by a difference in N170 amplitudes between face and non-face (chair) categories in the large hometown group but no difference between faces and chairs in the small hometown group. In sum, a relative deficit was observed on the CFMT but it is not clear how specific the effect may be in terms of domain and task.

Our first goal was to replicate the effect found in Balas & Saville (2015) in a larger and more heterogeneous sample recruited from the University of Nebraska-Lincoln, increasing statistical power and the ecological validity of the result. Second, we measured recognition abilities for both faces and other object domains to assess

whether the effect would generalize to another face task that is not a learning task like the CFMT, and to other learning tasks that share the format of the CFMT, for non-face domains.

In addition to determining if those from hometowns with lesser population density showed relatively poor face recognition ability, our second goal was to compare the degrees of "specialization" of faces and cars between HPD. It is well established that car recognition correlates below-average with the recognition of other object categories (McGugin et al., 2012; Van Gulick, McGugin, & Gauthier, 2016), suggesting that car recognition is more independent from general object recognition than other object categories (e.g., birds, mushrooms). Indeed, the dissociation between cars and other object domains is often similar in extent to what is found between faces and other object domains (McGugin et al., 2012; Van Gulick et al., 2016). Since the dissociation between faces and other object domains is often used as evidence that faces are "special" (e.g. McKone, Kanwisher, & Duchaine, 2007; Yue, Tjan, & Biederman, 2006), by this standard, cars would also have to be considered "special."

Determining that the recognition of cars is "special" (i.e. independent from that of other object categories) would have important theoretical ramifications, since an evolutionary explanation for why faces are special could not apply to cars (given cars have only existed for the past century or so). Instead, we would have to explore other possible explanations for the independence of car recognition. For instance, people could have more knowledge about cars, though a recent study found little evidence that knowledge mediates the correlation between car

recognition and the recognition of novel objects (Richler, Wilmer, & Gauthier, 2017). Another possibility is that, given the ubiquity of cars in the modern world, people have more experience with cars as compared with other object domains. If experience was responsible (or partially responsible as it is likely more than one explanation could apply) for the "specialness" of car recognition, then independence of car recognition should be modulated by HPD. Thus, we predict that in a lower density hometown sample, both face and car recognition will be more strongly correlated with other domains than in a higher density hometown sample. This prediction assumes people from low-density hometowns encounter fewer cars than those from high-density hometowns given that a less dense population would imply fewer cars in the visual environment. It is of course possible that this may not be the case given that visual experience with cars can occur not only through in-person interactions but also through perception of images on the Internet, television, magazines, etc. This is also true of faces, however, and given the results from Balas & Saville (2015), we assume that people who live in less dense towns encounter fewer people on a daily basis than those from more dense towns, leading to relatively less experience with faces and cars. Balas & Saville (2015) found a significant difference between face- and chair- evoked N170 amplitudes in their large hometown group but no difference in the small hometown group (although the interaction was not significant), suggesting that face recognition is more distinct from object recognition in those from more dense hometowns than those from less dense hometowns (Balas & Saville, 2015). Therefore we have two main hypotheses: (1) we will replicate the previous finding that people from high density hometowns

perform better on a face recognition measures – and (2) face and car recognition will correlate more strongly with the recognition of other categories for our low-density hometown sample than our high-density hometown sample. As an extension, we ask whether this finding of better recognition in people from high density hometowns generalizes to a non-learning face task or to learning non-face tasks.

Methods

Subjects

A total of 172 subjects were recruited using flyers placed around the University of Nebraska – Lincoln campus. Many UNL students are in-state and come from towns just outside of Lincoln or Omaha. Subjects were compensated $37.50 ($15/hr) for completing all tests and all work was conducted under the approval of both Vanderbilt and UNL Institutional Review Boards and was conducted in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). Informed consent was obtained from all subjects. Of these 172 subjects, 111 reported their hometown zipcode in a follow-up email (as we determined in preliminary stages of analyses that hometown population size and self-reported hometown size were not good predictors of population density, and reasoned that population density is likely more relevant than is hometown population to day-to-day experience with faces and cars). From hometown zipcodes, population density could be determined from www.unitedstateszipcodes.org. One subject was excluded

because of below or near chance performance levels (range .21-.38) on all recognition tasks, leaving 110 subjects.

Procedure

Subjects completed all of the following tests through an online website. Total, the tests took approximately 2.5 hours to complete and subjects were given a week to complete the tests, in a single order: SVET- Bird, SVET-Mushroom, SVET-Car, SVET-Plane, VET- Bird, VET-Mushroom, VET-Car, VET-Plane, CFMT, VFMT and CCMT.

Semantic Vanderbilt Expertise Tests (SVETs)

The SVET is designed to measure semantic knowledge about a particular domain. For 48 trials (with 3 catch trials), subjects choose the real subordinate-level label among two plausible distractor labels (Van Gulick et al., 2016). For example, subjects must choose the option displaying the text "Evening Grosbeak" as the correct bird label, instead of "Dakota Raven" or "Antietam." Here we used the SVET for birds, mushrooms, planes and cars to provide measures of semantic knowledge to accompany every VET. This task takes approximately five minutes to complete.

Vanderbilt Expertise Tests (VETs)

The Vanderbilt Expertise Tests were developed to measure object recognition for several domains using a learning-exemplar task similar to that used in the CFMT (McGugin et al., 2012). Thus, subjects study six exemplars at the beginning of each VET for 20 seconds and then complete an initial 12 three-alternative forced-choice trials (See Figure 1). On each trial, subjects have to

determine which of three items is identical to one of the six previously studied objects. Following the first six trials, there is a further 20-second study period, after which subjects complete 36 trials where the correct response is not an identical

Figure 1. Example Stimuli from the VETs and the CFMT. Top row from left to right; mushrooms, planes and birds. Bottom row from left to right; car, CFMT stimuli, CFMT stimuli with noise added.

image to the image of studied exemplar (so no image matching is possible). Feedback is provided on the first 12 trials but not the later 36. Our subjects completed VETs for birds, planes, cars, and mushrooms to provide both living and non-living domains. Responses were un-speeded and each VET for a single domain takes approximately 10 minutes to complete.

Cambridge Face Memory Test (CFMT)

The CFMT was designed as a measure of face recognition ability (Duchaine & Nakayama, 2006). Subjects begin by studying six Caucasian male grayscale faces and then complete three-alternative forced choice trials to distinguish the target faces from two distractors (Figure 1). On the first 18 trials, faces are presented with viewpoints that are identical to the studied viewpoint, followed by 30 trials in which the faces vary in viewpoint and lighting, and then 24 trials in which Gaussian noise is added to the images (bottom right image in Figure 1). Here, we used the longer CFMT (Russell et al., 2009), which includes 30 additional difficult trials at the end of the test in which more Gaussian noise is added to the images. Subjects studied the target faces between each block and responses were un-speeded. The test takes approximately 15 minutes to complete.

Cambridge Car Memory Test (CCMT)

The CCMT was designed to measure car recognition ability using the same task used in the CFMT (Dennett et al., 2012). Because of limitations of the online website we used to record responses, we had to modify the presentation of stimuli from the original presentation format (three cars staggered along a diagonal from upper left to bottom right) to a new format in which we presented the three car options vertically centered and stacked. Unfortunately, we discovered that some subjects misinterpreted the instructions and we could not reliably determine from the responses collected which subjects were properly responding and which were incorrectly responding. Thus, we did not include the CCMT in any of our analyses.

Vanderbilt Face Matching Test (VFMT)

The VFMT was created to measure face recognition ability using a different task from that used in the CFMT, CCMT and VETs (Sunday, Lee, & Gauthier, in press). In contrast to these tests, the VFMT does not require learning about a small set of faces over a series of trials, but instead only requires short-term visual memory to match face identity on a new set of faces on each trial. We included the VFMT as another measure of face recognition ability that uses a different task from that used by the CFMT. This inclusion allows us to determine whether the hometown-related effects found in Balas & Saville (2015) generalize to all tests that tap into face recognition ability or are specific to the learning exemplar CFMT task. Each of the 95 trials uses a new set of 5 face images of either male or female Caucasian faces (same within a trial). Subjects study two faces for four seconds and then in a test display, they must choose which of three faces matches one of the two studied faces (Figure 2). Subjects are instructed to match identity and not image, since the studied and correct responses target faces are different images of the same individual. Feedback is provided only on the practice trials and first three test trials. Face genders were interleaved to reduce proactive interference and responses were unspeeded. The VFMT takes approximately 15 minutes to complete.

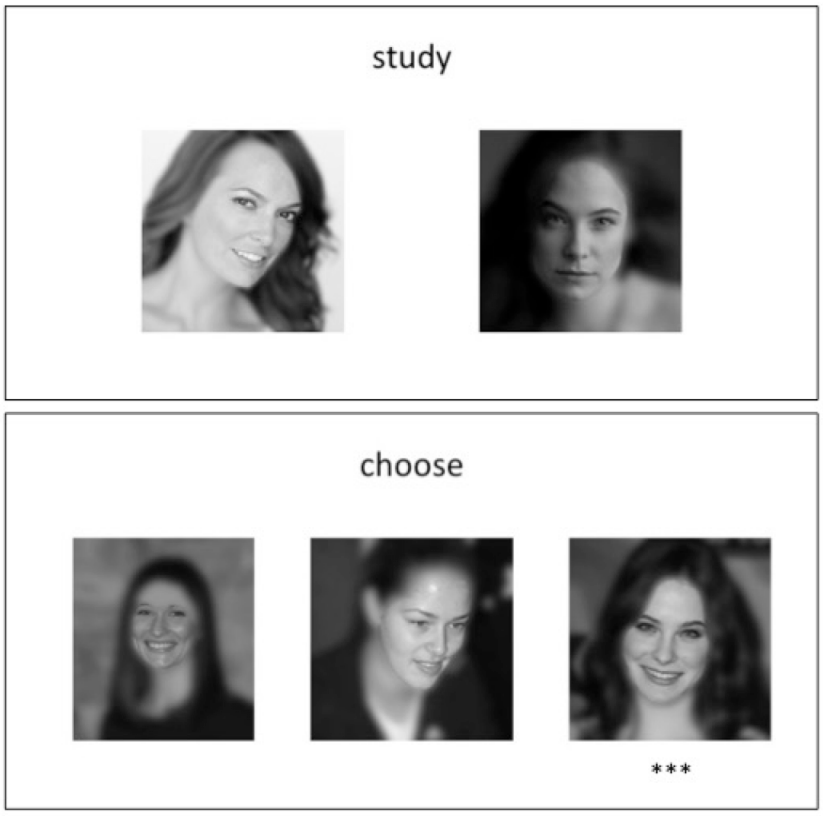

Figure 2. Example VFMT trial. Subjects studied the two top panel faces for four seconds and then choose which of the three following faces were one of the two studied faces. The correct response is indicated by the asterisks.

Self-Reported Expertise

Subjects were also asked to report their experience with each domain (bird, mushrooms, planes, and cars) on a Likert scale from 1-9 using the general statement: Please rate your expertise with {domain}. By expertise we mean your experience with, interest in, and knowledge about items in this category, relative to other people. Subjects also rated their general interest in object recognition through a series of four questions rating from 1-7: (1) their interest in classifying objects in their various sub-categories, (2) how easily they learn to recognize objects visually, (3) how much of their time at work or school involves recognizing things visually,

and (4) how much of their free time involves recognizing things visually (Van Gulick et al., 2016).

Self-Reported Hometown Size and Zipcodes

Subjects answered the question "How would you classify the place you consider your hometown?" on a scale from 1-5 (1: very small town (population less than 1,000), 2: small town (population greater than 1,000 but less than 30,000), 3: small city (population greater than 30,000 but less than 250,000), 4: large city (population greater than 250,000 but less than 1 million), 5: major metropolitan area (population greater than 1 million)). We chose 1,000 and 30,000 as our cutoffs because they are the cutoffs used in Balas and Saville (2015). To get a more continuous estimate of hometown size, we asked subjects to report their hometown zipcode in a follow-up email (111 out of 172 responded). We obtained population and population density (people per square mile) values for each of these zipcodes from www.unitedstateszipcodes.org. These population and population density values are derived from multiple sources, including the U.S. Postal Service, U.S. Census Bureau, Yahoo, Google, FedEx and UPS. We did not ask our subjects to report the exact years during which they lived in the reported hometown, however, meaning that these population values may not exactly correspond to when the subjects lived in their hometowns.

When comparing performance across groups, in addition to NHST results, we provide Bayes Factor (evidence favoring better performance in the high than low population group) and Bayesian estimation of the effect size in the form of the 95% highest density interval (95% HDI) using the BEST program (Kruschke, 2013),

computed using the program's default normal prior, which has been shown to have minimal impact on the posterior distribution. BEST uses an MCMC algorithm to generate the posterior distribution, and we used a chain length of 100,000. When zero does not fall in the 95% HDI, it indicates a credible difference.

Results

Face recognition ability changes over a lifespan, though this ability remains relatively stable across ages 18-50 (Germine, Duchaine, & Nakayama, 2011). Thus, to limit age-related variance, we excluded subjects over the age of 50 (3 subjects), leaving a total of 107 subjects (30 male; mean age = 25.0 years, range = 19–49; 85.0% were Caucasian, 7.5% Asian, 2.8% Hispanic/Latino, 1.9% African-American, 0.9% other and 1.9% Middle Eastern) in the analyses. Within these 107 subjects,

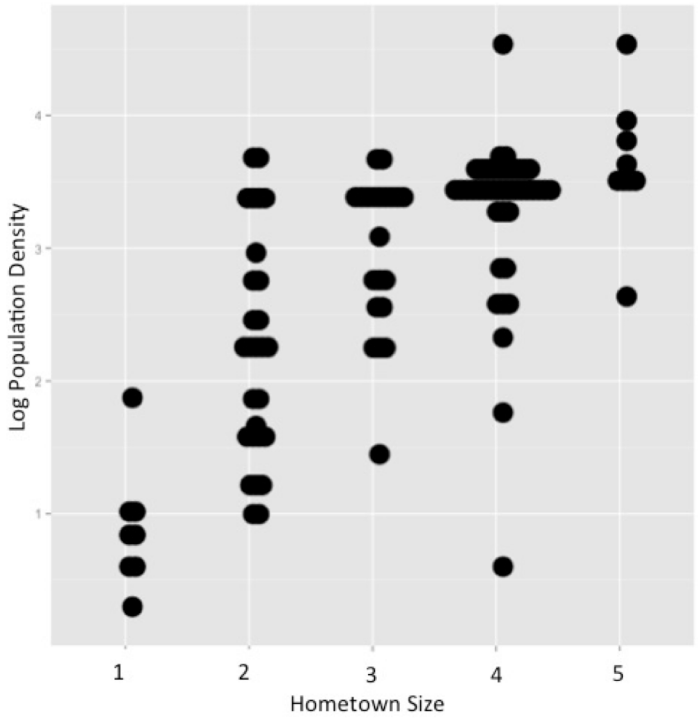

Figure 3. Dot plot showing population density distributions for each self-reported hometown size.

CFMT scores showed no correlation with age (r_{107} = -0.09, p = .38).

While Balas & Saville (2015) categorically compared two groups of subjects for whom they did not obtain the exact hometown size (Small hometown group self-reported hometown populations < 1,000; Large hometown group self-reported hometown populations > 30,000), our more continuous measure affords us the opportunity to examine subjects from a total 72 different hometown zipcodes. Considering the relation between reported hometown size and population density derived from zipcodes, we find that while hometowns of size less than 1,000 (Hometown 1) do show smaller zipcode population density, there is considerable overlap among the other four groups (Figure 3). We reasoned that population density was the variable more relevant to daily experience with faces, thus we grouped our subjects based on population density. Because our population densities ranged from 2 to 34,190, we log-transformed the density measure using base 10, although we report raw un-transformed values in the text and Table 1 for clarity.

Balas & Saville had a small hometown group from towns of less than 10 people per square mile and a large hometown group from a town of around 85 people per square mile (Balas & Saville, 2015). To compare our results to theirs, we created three groups: Small HPD (population density ≤ 10 ppl/mi^2); Medium HPD (10 < population density ≤ 85 ppl/mi^2); Large HPD (population density > 85 ppl/mi^2). We chose 85 ppl/mi^2 as a cutoff point because it approximates the average population density of the entire United States (87.4 ppl/mi^2; Balas & Saville, 2015) and falls within a gap between our largest medium population density (77 ppl/mi^2) and smallest large population density (159 ppl/mi^2). Two of our subjects reported

hometown sizes of less than 1,000 but had zipcode population densities that would place them in the medium group (log population densities of 12 and 75) were included in the small hometown group based on Figure 3. The demographics for each group are reported in Table 1.

Hometown Group	N	Mean Age (SD)	Percent female	Percent Caucasian	Mean Pop. Density (SD)
Small	10	26.7 (3.8)	80%	80%	13.30 (21.9)
Medium	13	28.4 (10.4)	69%	92%	37.9 (21.1)
Large	84	24.3 (5.6)	71%	85%	3142.1 (5166.7)

Table 1. Demographics for small, medium and large hometown groups.

Our small and large density groups correspond well to those groups in the Balas & Saville study, but it was not clear what to predict for our medium density group (i.e. whether we should expect a linear effect, or whether there is a point at which population density reaches a ceiling in its effects). In examining the average accuracies for the CFMT-long with the three groups (small: M = 57.1%, SD = 11.9%; medium: M = 58.1%, SD = 10.7%; large: M = 63.1%, SD = 10.2%), we noted the small vs. large group difference consistent with Balas & Saville (2015), with the medium group very similar in accuracy and variance relative to the small density group. Thus, to increase the power of our analyses, we combined the small and medium groups into one group (now called "low") for comparison with those from places with a large population density ("high"). These two groups had roughly similar demographics (low: N = 23, mean age = 27.7, SD age = 9.6, 74% female, 87%

Caucasian, mean population density = 27.2 ppl/mi^2, SD population density = 24.4 ppl/mi^2; high: N = 84, mean age = 24.3, SD age = 5.6, 71% female, 85% Caucasian, mean population density = 3142.1 ppl/mi^2, SD population density = 5166.7 ppl/mi^2).

Analyses of the patterns of mean performances: Will the small hometown sample show lower performance on the CFMT, as in Balas & Saville (2015)? Will this extend to another test of face recognition or to other learning tests with non-face categories?

For all analyses, no quantitative difference was observed between the short 72-trial CFMT version used in Balas & Saville (2015) and the extended score. Because the longer version is more sensitive to high range performance, from now on we report only this version, which we will call CFMT. Accuracies for each recognition test separated by group are shown in Figure 4.

Because of the difference in sample size, we tested for equality of variance between groups for each test, using Levene's test. The high group had higher variance than the low group on the VET-Bird (F = 5.11, p = .026) so for that test we used a Welch test to adjust degrees of freedom. None of the other tests showed significant evidence of unequal variance (p's > .25).

The only significant difference between low and high group average accuracies was found for CFMT-scores (two-tailed $t(105) = 2.25$, $p = 0.03$, $d = -0.52$, one tail BF: 3.87; 95% HDI: 0.002, 0.110, Figure 4). Indeed, both the long and short CFMT scores show a significant difference between hometown groups, and thus

test-level difficulty does not seem to modulate the observed effect (average performance for the CFMT short form was 71.6 (.12) was 62.0 (.11) for the long form). t-tests and Bayesian estimates qualitatively agreed for all tests: for all the other tests (t-tests and Welch test for birds), *p*-values were large ($p > .4$), BFs favored the null hypothesis and the 95%HDI included 0.

Following Richler et al (2017), we aggregated the four VET scores to produce one total VET score that can be used as an estimate of general object recognition ability. This aggregation is useful since experience and interest in different VET categories do not correlate across categories (Richler et al., 2017). We then ran a 2 (hometown groups) x 3 (VET, VFMT, CFMT) ANOVA in which neither the main effect of group $F(1,105) = 2.57$, $p = .11$, $\eta p^2 = .02$, nor the group x category interaction, $F(2,210) = 1.77$, $p = .17$, $\eta p^2 = .02$, was significant[1].

Thus, we replicate the effect of HPD on CFMT, but we may lack the power to show that this this effect is larger than that for non-face learning tests or a non-learning face recognition task. We performed a power calculation specifying the same group ratios as in the present study and found that to detect the present interaction with 80% power, a sample size 2.6 times as large (278 subjects) would be required[2].

[1] There was a significant effect of Category, $F(2,210) = 4.04$, $p = .02$, which we do not interpret because the different tests were not meant to be equated in difficulty, so only within-tests effects or interactions were of interest.

[2] Note that such power calculations, based on a 95% confidence interval around the noncentrality parameter, are relatively imprecise (Taylor & Muller, 1996).

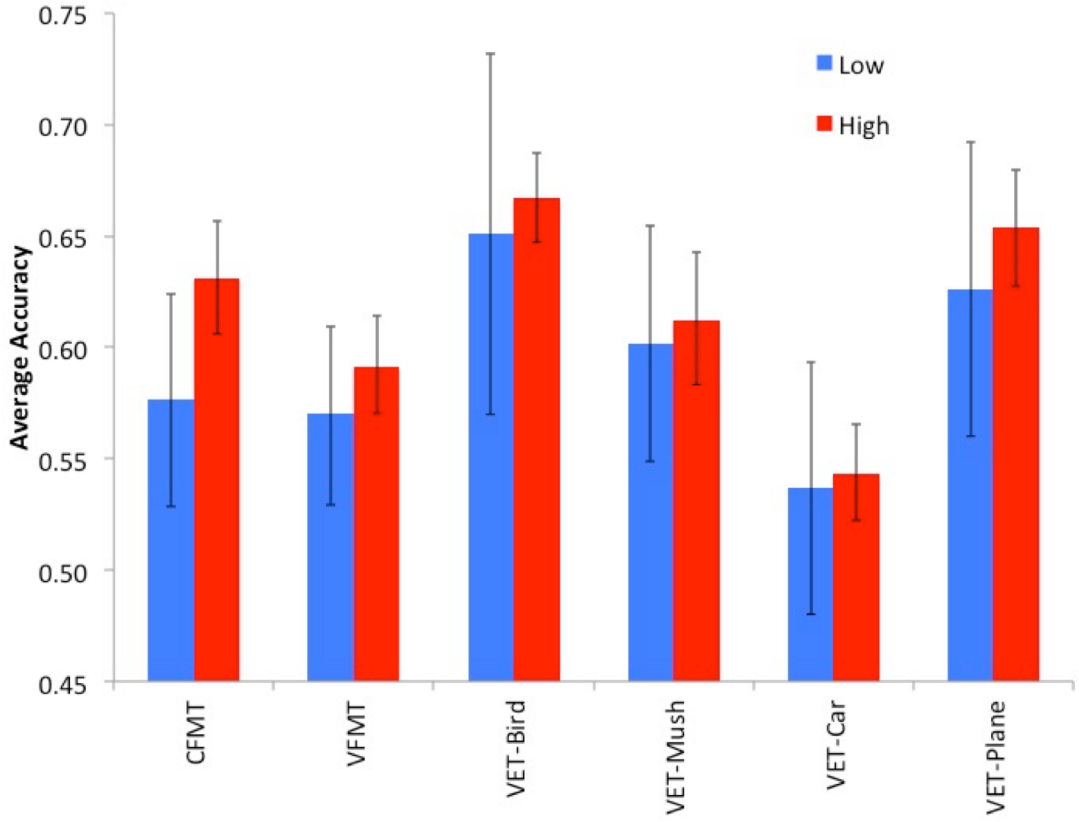

Figure 4. Bar graph of average accuracies for low and high hometown groups. Error bars show 95% confidence intervals.

We provide the full set of first-order correlations for the entire sample here (Table 3), to provide evidence that speaks to the convergent and discriminant validity of the various measures. As expected, because they are the only two tests in the same domain, the two face measures (CFMT and VFMT) showed the strongest correlation ($r_{107} = 0.67$, $p < .001$). Overall, the correlations that involved a face test or the VET-Car were lower (ranging from .28 to .48) than the correlations among the other categories (plane/bird/mushroom, ranging from .56 to .61). This is consistent with face and car recognition being relatively "specialized" abilities and

this issue is addressed specifically under Section 2, "*Analyses of the patterns of correlations as a function of hometown population density".*

	VET-Bird	VET-Mush	VET-Plane	VET-Car	VFMT	CFMT
VET-Bird	α = *.85*	0.61	0.61	0.44	0.48	0.48
VET-Mush	0.45	α = *.64*	0.56	0.28	0.31	0.30
VET-Plane	0.51	0.41	α = *.82*	0.39	0.35	0.35
VET-Car	0.35	0.19	0.30	α = *.74*	0.48	0.44
VFMT	0.38	0.22	0.31	0.35	α = *.74*	0.84
CFMT	0.41	0.22	0.29	0.35	0.67	α = *.85*

Table 3. Correlations between each test are shown in the lower left corner with the Cronbach alpha reliability shown along the diagonal (italicized). Dis-attenuated correlations are reported in the upper right corner. r > .31 are significant at alpha = .001; r > .24 are significant at α = .01; r > .18 are significant at α = .05.

Each VET correlated significantly with the SVET from its respective domain (Table 4, Birds: $r_{107} = 0.38$, $p < .001$; Mushrooms: $r_{107} = 0.23$, $p = .02$; Planes: $r_{107} = 0.24$, $p = .01$; Cars: $r_{107} = 0.40$, $p < .001$). As in prior work (Van Gulick et al., 2016)—and indicative of good validity of the tests as measures of specific experience with various categories—all but the VET-Mush/SVET-Mush ($r_{107} = 0.18$, $p = .13$) within-domain correlations remained significant after regressing out the averaged other domains (e.g. VET-Bird scores after the averaged VET-Mush, VET-Plane and VET-Car score is partialed out; Appendix, Table 2, r's$_{107} > 0.21$, *p*'s < .03)

	VET-Bird	VET-Mush	VET-Car	VET-Plane	SVET-Bird	SVET-Mush	SVET-Car	SVET-Plane
VET-Bird	*a = .85*	0.61	0.44	0.61	**0.58**	0.43	0.23	0.23
VET-Mush	0.45	*a = .64*	0.28	0.56	0.18	**0.40**	0.16	0.08
VET-Car	0.35	0.19	*a = .74*	0.39	0.33	0.21	**0.51**	0.01
VET-Plane	0.51	0.41	0.30	*a = .82*	0.30	0.33	0.24	**0.32**
SVET-Bird	**0.38**	0.10	0.20	0.19	*a = .50*	0.45	0.36	0.03
SVET-Mush	0.29	**0.23**	0.13	0.22	0.23	*a = .52*	0.30	0.43
SVET-Car	0.19	0.12	**0.40**	0.20	0.23	0.20	*a = .82*	0.48
SVET-Plane	0.17	0.05	0.01	**0.24**	0.02	0.26	0.36	*a = .67*

Table 4. Correlations between VETs and SVETs for the 107 subjects. Cronbach alpha values are shown along the diagonal (italicized) and disattenuated correlations are reported in the upper right corner. r > .31 are significant at alpha = .001; r > .24 are significant at α = .01; r > .18 are significant at α = .05. Within-domain correlations are bolded.

Within-domain VETs and relevant self-reported experience showed small correlations, consistent with prior work demonstrating that people are not very good at predicting their recognition performance relative to other people (Van Gulick et al., 2016, see Appendix, Table 1). Correlations between recognition tests and average self-reported general object recognition interest and experience were also small but consistent (Mean r_{107} = .09, range r = -.07-.16).

Population density did not correlate with performance on any of our tests (all r_{107} < .15, *p's* > .13). Adding a quadratic trend for log population density to the linear effect did not substantially improve the fit (R^2 goes from .15 to .18 for the CFMT, the measure that showed the strongest numerical increase).

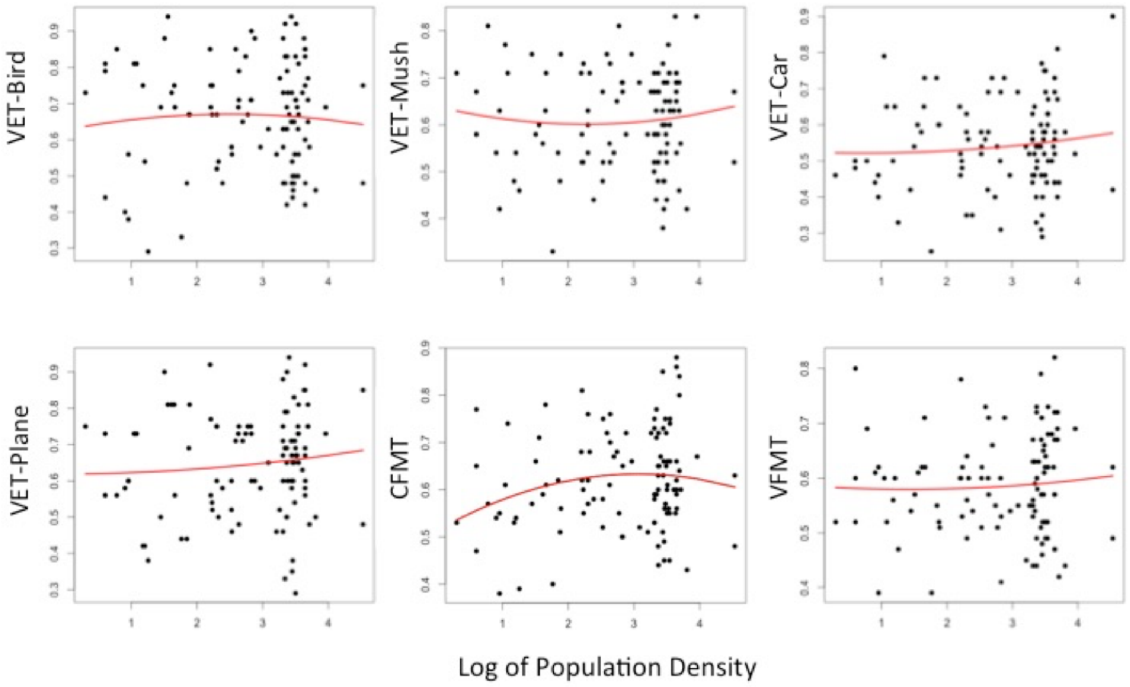

Figure 5. Scatterplots of recognition test scores versus log of the population density with quadratic fits in red.

Analyses of the patterns of correlations as a function of hometown population density: Are face and car recognition abilities more strongly related to other kinds of object recognition in a small hometown sample as compared with a large hometown sample?

This second set of analyses concerns not the mean performance on each test, but whether we can find evidence that in the absence of early experience with a large number of faces and cars, as represented by HPD, face and car recognition is more strongly related to other kinds of object recognition. That is, we already know that in large samples for which HPD is not controlled (but is presumed to be relatively large), performance with faces and with cars shows correlations with object recognition for other categories that are lower than average. Here, we examine whether this effect is stronger in high than low HPD groups, for each individual face and car test.

We performed three separate sets of analyses that focused on the relations among bird, mushroom, and plane recognition and the CFMT, VFMT, and VET-Car recognition measures, respectively. Thus, each analysis involved the correlations among a set of four variables, each assessed within the high and low HPD groups. Below, for the sake of brevity, we state our hypotheses in terms of face recognition (applying to the CFMT and VFMT), but the logic is parallel for cars (VET-Car). Our hypotheses can be framed in terms of the relative magnitude of correlations involving birds, mushrooms, and planes. We predicted that (see Table 5): (1) Of all the correlations involving birds in the high- and low-density groups (e.g., bird-mushroom-high, bird-plane-high, bird-face-high, bird-mushroom-low) the lowest

correlation would be that between birds and faces in the high density group; (2) Of all the correlations involving mushrooms in the high- and low-density groups the lowest correlation would be that between mushrooms and faces in the high-density group; and, (3) Of all the correlations involving planes in the high- and low-density groups the lowest correlation would be that between planes and faces in the high-density group. Thus, within each of the three non-face categories, our hypotheses imposed five inequality constraints. For example, in the case of birds, the correlations for each of the five members of the set bird-mushroom-high, bird-plane-high, bird-mushroom-low, bird-plane-low, and bird-face-low would be greater than the correlation between bird and faces in the high-density group. Thus, across birds, mushrooms, and planes there were 15 inequality constraints in all. We should note that we did not specify any specific pattern of inequalities among pairs of correlations that: (1) Did not involve faces (e.g., there were no inequality constraints on the relation between bird-mushroom-high, bird-plane-high, bird-mushroom-low, and bird-plane-low); (2) Only involved faces within a given density group (e.g., our hypotheses did not constrain the relative magnitude of bird-face-low and mushroom-face-low); and (3) Had no stimuli in common (e.g., the bird-mushroom and plane-face correlations within or across density groups).[3] Based on our prior findings, our predictions here were strongest for the CFMT and VET-Car, as these tests have been used in combination with tests for several other object categories (e.g., VET battery for birds, mushrooms, planes, motorcycles…) in prior

[3] When we imposed additional constraints that also included these correlations (e.g., r between bird and mushroom > r between plane and face) the pattern of results was very similar to those reported below and conclusions about magnitude of effects were identical.

studies with large samples (unscreened for hometown size) and have repeatedly shown lower than average correlations (McGugin et al., 2012; Van Gulick et al., 2016). In contrast, the VFMT has not been used yet in that context. The VFMT has been found to correlate well with the CFMT (Sunday et al., in press) which could lead to the prediction that performance with the VFMT becomes more independent from object recognition with experience. Our results from the above analysis suggest that the VFMT is less sensitive to experience than the CFMT as it relates to mean performance.

r(H bird,face)	**<**	**r(H bird,mush)**
r(H bird,face)	**<**	**r(H bird,plane)**
r(H bird,face)	<	r(L bird,face)
r(H bird,face)	<	r(L bird,mush)
r(H bird,face)	<	r(L bird,plane)
r(H mush,face)	**<**	**r(H mush,bird)**
r(H mush,face)	**<**	**r(H mush,plane)**
r(H mush,face)	<	r(L mush,face)
r(H mush,face)	<	r(L mush,bird)
r(H mush,face)	<	r(L mush,plane)
r(H plane,face)	**<**	**r(H plane,bird)**
r(H plane,face)	**<**	**r(H plane,mush)**
r(H plane,face)	<	r(L plane,face)
r(H plane,face)	<	r(L plane,bird)
r(H plane,face)	<	r(L plane,mush)

Table 5. The 15 inequality constraints included in the combined groups hypothesis. The 6 constraints in bold are those that form the within group (here, High Density) hypothesis. Here, face could denote either CFMT or VFMT scores, or it would be replaced by the VET-Car. H and L denote the high and low population density groups.

In addition to an analysis that combined both groups, we were interested in testing our hypotheses focusing only on the high population density group. In this

group, we predicted that (see Table 5, bold constraints): (1) The bird-face correlation would be lower than both the bird-mushroom and bird-plane correlations; (2) The mushroom-face correlation would be lower than both the bird-mushroom and mushroom-plane correlations; and, (3) The plane-face correlation would be lower than both the bird-plane and mushroom-plane correlations. Thus, six constraints in all were imposed within the high-density group. Although we did not hypothesize the same effect in the low-density group, for comparative purposes we also assessed the strength of the evidence for this group. The logic of our predictions for cars again directly paralleled that just described for both face tests (i.e., the smallest correlations would be the three involving cars within the high-density group).

Note that the hypotheses across both groups and within the high-density group consist of *sets of ordinal (i.e., inequality) constraints* among pairs of correlations. Each constraint specifies that a given correlation is less than another correlation. Although one-tailed tests are commonly used to test a single inequality constraint considered in isolation, it is difficult to test sets of ordinal constraints using traditional statistical methods. Such predictions can, however, be tested using a Bayesian order-constrained hypothesis testing (BOHT) approach (e.g., Hoijtink, Klugkist, & Boelen, 2008; Klugkist, Landy, & Hoijinkk, 2005; Klutymans, van de Schoot, Mulder, & Hoijtink, 2012; Mulder, 2014, 2016). We used the analytic framework and software program BOCORR developed by Mulder (2016) for testing order-constrained hypotheses on correlations. This approach allowed us to test the two sets of composite hypotheses as a whole, rather than relying on tests of

individual pairs of correlations, one by one. In addition to allowing a more direct test of our hypotheses than a piecemeal approach, the BOHT approach had two additional advantages: (1) It does not require the multiplicity corrections necessitated when testing a large number of differences between pairs of correlations (e.g., Boelen & Hoitjtink, 2008; Hoitjtink, Huntjens, Reijntes et al., 2008); and, (2) It yields Bayes factors (BFs) that allowed us to quantify the degree of support for our hypotheses rather than relying on a series of reject/no-reject decisions.

Because of the complexity of the BOHT approach, we emphasize a more intuitive than mathematically rigorous description and refer readers interested in a more technical description to Mulder (2016) and the other sources cited above. Consider our predictions that span both density groups. Consider the 15 inequality constraints shown in Table 5 as the null hypothesis (H_0). Although some applications of BOHT involve multiple competing hypotheses of interest, in our case we simply compared H_0 to its alternative (H_a), that is, any admissible pattern of correlations *other* than that specified by the null hypothesis. When predictions were tested within the high-density group alone, H_0 specified six inequality constraints. Correspondingly H_A was any possible pattern of correlations in the high-density group other than those that would be consistent with H_0.

In both cases, the overriding goal was to compute BFs that quantify the degree of evidence in the data for H_0 relative to H_A. Before these BFs would be computed, it was first necessary to compute the BF for a given H_0 relative to what is known as the unconstrained, encompassing model, denoted as H_u (e.g., Berger &

Mortera, 1999; Klugkist & Hoijtink, 2007; Klugkist et al. 2005). This model imposes no ordinal constraints on the pattern of correlations but does specify an encompassing prior distribution that was designed to be a reasonable model of the multivariate distribution of correlations. We specified a joint prior for the unstructured correlation matrix that resulted in beta $\left(\frac{1}{2},\frac{1}{2}\right)$ distributions on the interval (-1,1) for the marginal priors of the separate correlations. Relative to alternative priors, this specification has been shown to enhance the sensitivity to detect a valid set of order constraints on correlations (Mulder, 2016).

To compute the BF comparing H_0 to H_U it was first necessary to compute both the prior and posterior probabilities that H_0 is correct. The prior probability that H_0 is correct does not incorporate the actual data collected. It is simply the proportion of outcomes under the encompassing prior that is consistent with the restrictions. To choose a simple example, if a single order constraint was being tested specifying that the difference between two correlations was greater than 0, the prior probability of H_0 would be .50 because half of all possible values of the two correlations would be consistent with this constraint. As the number of constraints within a set increase, the proportion of the total correlation space that is consistent with the complete set of restrictions becomes smaller and smaller. After observing the data, the prior distribution is updated using Bayes' theorem and the posterior probability of H_0 is computed. If the data are consistent with the restrictions, the posterior probability of H_0 is larger than the prior probability of H_0; that is, the

average probability density within the restricted space demarcated by H_o has increased.

These computations are generally analytically intractable. For this reason, using BOCORR (Mulder, 2016), both the prior and posterior probabilities were calculated by generating a large number of samples from the prior and posterior distributions and counting the proportion of samples that were consistent with the restrictions. To test the more complex models that included both the high- and low-density groups, we drew 10,000,000 samples and for the within-group analyses we drew 1,000,000 samples. In each case, we then computed the BF for H_o relative to H_u as the ratio of the posterior probability of H_o to the prior probability of H_o. BFs > 1 indicate that, consideration of the actual data increased the probability of H_o while BFs < 1 indicated that the observed data decreased the probability of H_o. Because the set of outcomes represented by H_0 and H_A are mutually exclusive, prior and posterior probabilities for H_A were simply 1 – the corresponding probabilities for H_0. In turn, these probabilities were used to compute the BF for H_A relative to H_U.

Although the BF for H_o relative to H_u was of interest, our primary goal was to compute a BF indicating the relative strength of the evidence for H_o relative to its alternative, H_A. It can be shown that this quantity is the simple ratio of the BF for H_o relative to H_U and the BF for H_A relative to H_U; that is, $\mathrm{BF}_{H_0,H_A} = \frac{\mathrm{BF}_{H_0,H_U}}{\mathrm{BF}_{H_A,H_U}}$. In interpreting BF_{H_0,H_A}, we used Jeffreys' (1961) guidelines according to which BFs between 3 and 10, between 10 and 30, and between 30 and 100 offered, respectively, substantial, strong, and very strong support for the target hypothesis

relative to its alternative. Although Jeffreys considered BFs between 1 and 3 barely worthy of mention, we favor the descriptor 'only marginal' support. When BFs are less than 1 and appear to favor H_a, the reciprocals of the ranges noted above provide descriptors for the strength of evidence. These descriptors facilitate communication but are essentially rough guidelines.

From a conceptual perspective, there are two critical features concerning the interpretation of BF_{H_0,H_A} that should be emphasized. First, like all Bayes factors, it indicates the proportional *change* in the relative probability (i.e., the odds) of the null and alternative hypotheses brought about by consideration of the actual data (e.g., Lavine & Schervish, 1999). Thus, even if a given H_0 has low a priori probability because it limits the set of possible correlations nested under it, it can be associated with a high BF relative to its alternative if the data are highly consistent with it: The data have markedly changed the relative probability of the two hypotheses. The second essential feature is that BFs can be considered the ratio of model fit to model complexity (e.g., Kluytmans, Schoot, Mulder, & Hoijtink, 2012). Prior probabilities are linked to model complexity. In this context, complexity is inversely related to precision and specificity: Less complex models make more precise and restrictive predictions and thus have lower prior probabilities. On the other hand, the better the fit of the data to the model, the higher the posterior probability. Because a BF is the ratio of posterior to prior probabilities, they will especially favor models that fit well despite being highly restrictive. That is, at equivalent levels of complexity, the higher the fit the higher the BF and at equivalent levels of fit, the less complex (i.e., more restrictive) the model the higher the BF. Finally, we note that: (1) BFs take

into account sample size; (2) There is evidence that a Bayesian approach to the analysis of correlations has better properties than frequentist approaches when ns are relatively small (e.g., as was the case in the low-density group; Fosdick & Raftery, 2012); and, (3) The BOCOR program can accommodate both within-group and across-group predictions (Mulder 2016).

Table 6 shows the correlations within the high- and low-density groups for each of three measures of interest (CMFT, VFMT, and VET-Car) and Table 7 shows BFs for both the combined groups and within-group analyses. Recall that our predictions applied to both the combined groups and within-high density analyses, with the low-density analyses included for comparative purposes. In addition, our predictions were strongest for CFMT and cars. We first consider the hypotheses that involved the pattern of correlations across both hometown groups. The CFMT analyses provided strong support for our predictions. An examination of the correlations in Table 6 show that the three lowest correlations among faces (as assessed by the CFMT), birds, mushrooms, and planes were the three correlations involving faces within the high-density group. Consistent with this observation, the BF for the combined group indicated strong support for hypotheses (BF = 22.91). In contrast the BOHT analysis performed on the VFMT showed only marginal support the target hypothesis (BF = 2.36). The BF for VET-Car indicated "substantial" support for the target hypothesis based on Jeffreys' (1961) criteria.

	CFMT	VET-Bird	VET-Mush
VET-Bird	.76/.27		
VET-Mush	.54/.11	.63/.37	
VET-Plane	.51/.20	.64/.46	.36/.42

	VFMT	VET-Bird	VET-Mush
VET-Bird	.50/.34		
VET-Mush	.52/.12	.63/.37	
VET-Plane	.16/.35	.64/.46	.36/.42

	VET-Car	VET-Bird	VET-Mush
VET-Bird	.49/.30		
VET-Mush	.47/.09	.63/.37	
VET-Plane	.35/.28	.64/.46	.36/.42

Table 6. Pearson Correlations for the Low/High population density groups (note the results for birds, mushrooms and planes are repeated in each sub-table, to form 3 sets of correlations used to test 3 sets of order-constrained correlations).

Measure	Model		
	Combined Groups	High Density	Low Density
CFMT	22.91	15.87	0.11
VFMT	2.36	4.67	1.97
VET-Car	7.91	9.07	1.77

Table 7. Bayes Factors for Inequality-Constrained Hypotheses for Correlations

Note: Bayes Factors > 1 indicate greater support for H_0, relative to H_a, and Bayes Factors <1 indicate greater relative support for H_a. According to Jeffreys' (1961) criteria, 3 > BF <1 indicates only marginal support for H_0, 10 >BF >3 indicates substantial support for H_0 , and 30 >BF >10 indicates strong support for H_0, while the reciprocals of these ranges indicate strength of support for H_a.

Examination of the correlation matrices indicated that correlations for the low-density group were generally higher than those for the high-density group, especially on the CFMT. Tests of the equality of correlation matrices (Steiger, 1980)

conducted for the three 4-variable sets (CMFT-bird-mushroom-car, VFMT-bird-mushroom-car, and VET-car-bird-mushroom-car) indicated significant differences on the matrix that included CFMT scores , $\chi^2_6 = 16.45, p < .02,$ but no significant differences on the matrices that included VFMT and VET-Car scores, $\chi^2_6 = 7.34, p > .25,$ and $\chi^2_6 = 6.33, p > .35,$ respectively.

Although there were also no significant differences on tests of the equality of covariance matrices, all Box M (Morrison, 1976) *p*'s > .20, the overall differences in magnitude evident on the CFMT correlation matrices could themselves in part account for the high BF for the cross-groups analysis. Thus, for this variable in particular we deemed the separate within-high and within-low density analyses particularly determinant. On the CFMT measure, the BOHT analysis of the high-density group indicated strong support for the target hypothesis (BF = 15.87). In sharp contrast, if anything, the correlations involving the CFMT in the low-density group indicated that consideration of the actual data yielded increased support for the alternative relative to the null hypothesis (BF = 0.11). Two of the six target correlations were in the hypothesized direction within the low-density group but four of six were in the opposite direction. For both the VFMT and VET-Car analyses, the strength of the evidence for the target hypothesis within the high-density group was in the "substantial" range, with the magnitude for cars midway between that of the CFMT and VFMT (BFs = 4.67 and 9.07, respectively for VFMT and cars). On both measures, BFs within the low-density group were only marginal (BFs = 1.97 and 1.77, respectively for VFMT and VET-Car).

In sum, Bayesian tests of order-constrained hypotheses that specifically addressed the prediction that faces and cars are more independent from object recognition in the high than in the low HPD group found strong and substantial support, respectively, for the CFMT and the VET-Car, but only marginal support for the VFMT. These results offer the first evidence in support of any hypothesis for why face recognition (and car recognition) may be "special". Interestingly, our analyses that addressed patterns of correlations among abilities are independent from those addressing mean patterns of performance, in both cases we find that the VFMT did not show the same sensitivity to HPD as the CFMT. This illustrates how variance on any one test includes both aspects that tap into a construct of interest (here, face recognition ability) as well as more test-specific components. We speculate that the CFMT format measures a face-learning component that is not as important in the VFMT where each trial is independent, but future work could test this hypothesis with a number of different face recognition tests designed to tap or not into such a process. Interestingly, we find no evidence that the CFMT and VFMT are any less related in one group than the other (Low-density group: $r = .68$, High-density group: $r = .66$).

Discussion

First, we compared mean performance on a number of tests of face and object recognition in people who came from hometowns with relatively low vs. high population density. We replicated a relative disadvantage on the CFMT for people

from low population density hometowns. While this effect has been reported twice (Balas & Saville, 2015; 2017), we extend it in a few ways. The original two studies were conducted in undergraduates in North Dakota who came from small (less than 1000) or large (>30000) hometowns. Here, we did not restrict on hometown size, and after collecting information on both hometown size and hometown zipcode, used HPD measured directly to consider hometown experience on a more continuous basis. Accordingly, we were able to determine that those coming from the untested intermediate range of population density in the original study (those from hometowns with population densities between 10 and 85 per square mile) showed results on the CFMT similar to those from the smaller hometowns. In addition, the Balas & Saville studies tested only undergraduates, between 18 and 24. We tested subjects in a much wider age range (19-49, with 37% of our sample older than 24) and nonetheless replicated the population density effect on CFMT performance. While this suggests that later life experiences do not erase the influence of hometown environment, we did not collect data on current location population density or how it may have varied through our subjects' lives, and we would not exclude that such later experience could account for meaningful variance in face recognition if it was measured.

In addition to replicating the effect of HPD on the CFMT, we compared the face learning task to a non-learning face task, and to a battery of tests of learning various non-face objects. On the one hand, the CFMT was the only task that showed a significant effect of HPD. On the other hand, we did not have sufficient power to claim that the effect for CFMT was larger than the non-significant effects, in the

same direction, observed on the other tasks. Our results may suggest that this effect could be relatively specific to a face-learning task. We speculate that this may reflect those from small hometowns needing to learn and repeatedly individuate fewer faces than those from larger hometowns, but the ubiquity of the internet and television make face-processing skills unrelated to learning more equal between different hometown sizes.

Before any strong conclusion is made about whether the effect is specific to faces and the extent to which it depends on the learning format of the CFMT, it is clear that a larger sample size, especially in the low population density group, will be necessary. Future work should include other tests in which subjects learn faces over trials as in the CFMT (e.g., the Vanderbilt Face Expertise Tests, Ryan & Gauthier, 2016) with tests that involve less face learning (e.g., ensemble-perception face tasks like Haberman, Brady, & Alvarez, 2015) in samples of varying hometown populations.

Second, we compared the pattern of correlations among various abilities in the two hometown groups. Car and face recognition have been reported to be special abilities that are surprisingly independent of other object recognition abilities and from each other (McGugin et al., 2012; Van Gulick et al., 2015; Richler et al., 2017). While no study to date has offered an explanation for this, one suggestion is that high levels of experience for both categories—as mediated by population density—could lead to the development of specialized recognition mechanisms (Gauthier, in press). Here, using sensitive Bayesian tests of order-constrained correlations, we found support for the hypothesis that car and face recognition as

measured by learning tests (CFMT and VET-Car) are more independent from general object recognition in people who come from large hometowns than in those who come from much smaller hometowns. The differences in the direction and/or magnitude of BFs were particularly striking when computed separately within the high and low population density groups. We found only meager support for the same pattern when measuring face recognition ability with the VFMT, despite the fact that the tests are strongly related. Just as for the effect of population density on mean performance, it appears that although relatively small, the non-error related, unshared variance between the CFMT and VFMT is critical in revealing the role of experience.

In sum, this work replicates prior findings that people in low population density hometowns are poorer at face learning than those in larger hometowns. It also reveals for the first time that face and car recognition abilities are not particularly "special" for people who grew up in small hometowns, while there is much stronger evidence that they are special for people who grew up in larger hometowns. When only faces are special, a nativist account may be plausible. The finding that in terms of individual differences, cars are equally as special made such account less plausible, and the current results point further in the direction of experience as a driving factor. However, HPD is only an indirect measure of experience with faces or cars and we and others (Balas & Saville, 2015; 2017) did not collect a great deal of information on other ways these individuals may differ. Future studies should consider gathering converging evidence from other correlates of experience, both early and late. Finally, our work shows the importance of using a

multiplicity of measures with different formats to help clarify the nature of these effects. That is, we have previously cautioned against strong conclusions based on only two object categories (Gauthier & Nelson, 2001; Gauthier, in press) and here we add caution about strong conclusions based on abilities measured using a single test format.

Acknowledgements: This work was funded by the Temporal Dynamics of Learning Center (National Science Foundation Grant SBE-0542013 and SMA-1640681). This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. (1445197). We would like to thank Susan Benear for her help collecting data for this study.

Appendix:

Table 1. Correlations between VETs and Self-Reported Expertise (SR). Within-domain correlations are bolded. r > .24 are significant at α = .01; r > .18 are significant at α = .05. Within-domain correlations are bolded.

	VET-Bird	VET-Mush	VET-Car	VET-Plane	SR-Bird	SR-Mush	SR-Car
VET-Mush	0.448	-					
VET-Car	0.351	0.19	-				
VET-Plane	0.507	0.406	0.3	-			
SR-Bird	**0.265**	0.055	0.116	0.27	-		
SR-Mush	0.102	**-0.142**	-0.055	0.131	0.45	-	
SR-Car	0.209	0.069	**0.219**	0.289	0.306	0.326	-
SR-Plane	0.021	-0.115	-0.043	**0.28**	0.402	0.592	0.436

Table 2. Correlations between VETs and SVETs on the averaged score of the other domains was been regressed out. Within-domain correlations are bolded.

	VET-Bird	VET-Mush	VET-Car	VET-Plane	SVET-Bird	SVET-Mush	SVET-Car
VET-Mush	-0.15	-					
VET-Car	-0.21	-0.33	-				
VET-Plane	-0.16	-0.17	-0.24	-			
SVET-Bird	**0.28**	-0.06	0.07	0.00	-		
SVET-Mush	0.12	**0.15**	-0.05	0.03	-0.00	-	
SVET-Car	-0.13	-0.05	**0.37**	-0.04	-0.13	-0.30	-
SVET-Plane	0.06	-0.04	-0.17	**0.21**	-0.36	-0.07	-0.11

References:

Balas, B., & Saville, A. (2015). Neuropsychologia N170 face specificity and face memory depend on hometown size. *Neuropsychologia*, *69*, 211–217. https://doi.org/10.1016/j.neuropsychologia.2015.02.005

Balas, B., & Saville, A. (2017). Hometown size affects the processing of naturalistic face variability. *Vision Research*. https://doi.org/10.1016/j.visres.2016.12.005

De Heering, A., De Liedekerke, C., Deboni, M., & Rossion, B. (2010). The role of experience during childhood in shaping the other-race effect. *Developmental Science*, *13*(1), 181–187. https://doi.org/10.1111/j.1467-7687.2009.00876.x

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. https://doi.org/10.3758/s13428-011-0160-2

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. https://doi.org/10.1016/j.neuropsychologia.2005.07.001

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), 7–7. https://doi.org/10.1167/14.8.7

Germine, L. T., Duchaine, B., & Nakayama, K. (2011). Where cognitive development and aging meet: Face learning ability peaks after age 30. *Cognition*, *118*(2), 201–210. https://doi.org/10.1016/j.cognition.2010.11.002

Haberman, J., Brady, T. F., & Alvarez, G. A. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *Journal of Experimental Psychology: General*, *144*(2), 432–446. https://doi.org/10.1037/xge0000053

Kluytmans, A., Schoot, R. Van De, Mulder, J., & Hoijtink, H. (2012). Illustrating Bayesian evaluation of informative hypotheses for regression models. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00002

Kruschke, J. K. (2013). Bayesian Estimation Supersedes the t Test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. https://doi.org/10.1037/a0029146

Logan, G. D. (1988). Toward an Instance Theory of Automatization. *Psychological Review*, *95*(4), 492–527. https://doi.org/10.1037/0033-295X.95.4.492

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*, *69*, 10–22. https://doi.org/10.1016/j.visres.2012.07.014

McKone, E., Kanwisher, N., & Duchaine, B. C. (2007). Can generic expertise explain special processing for faces? *Trends in Cognitive Sciences*, *11*(1), 8–15. https://doi.org/10.1016/j.tics.2006.11.002

Palmeri, T. J. (1997). Exemplar Similarity and the Development of Automaticity.

Journal of Experimental Psychology: Learning, Memory, and Cognition, *23*(2), 324–354.

Richler, J. J., Tomarken, A., Vickery, T. J., Ryan, K. F., Floyd, R. J., Sheinberg, D., … Gauthier, I. (n.d.). Individual Differences in Object Recognition.

Richler, J. J., Wilmer, J. B., & Gauthier, I. (2017). General object recognition is specific: Evidence from novel and familiar objects. *Cognition*, *166*, 42–55. https://doi.org/10.1016/j.cognition.2017.05.019

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: people with extraordinary face recognition ability. *Psychonomic Bulletin & Review*, *16*(2), 252–257. https://doi.org/10.3758/PBR.16.2.252

Ryan, K. F., & Gauthier, I. (2016). Gender differences in recognition of toy faces suggest a contribution of experience. *Vision Research*, *129*, 69–76. https://doi.org/10.1016/j.visres.2016.10.003

Sangrigoli, S., Pallier, C., Argenti, A. M., Ventureyra, V. A. G., & De Schonen, S. (2005). Reversibility of the Other-Race Effect in Face Recognition During Childhood. *Psychological Science*, *16*(6), 440–444. https://doi.org/10.1111/j.0956-7976.2005.01554.x

Shakeshaft, N. G., & Plomin, R. (2015). Genetic specificity of face recognition. *Proceedings of the National Academy of Sciences*, *112*(41), 201421881. https://doi.org/10.1073/pnas.1421881112

Sunday, M. A., Lee, W., & Gauthier, I. (n.d.). Age-related differential item functioning in tests of face and car recognition ability. *Journal of Vision*.

Tanaka, J. W., Kiefer, M., & Bukach, C. M. (2004). A holistic account of the own-race

effect in face recognition: Evidence from a cross-cultural study. *Cognition*, *93*(1), 1–9. https://doi.org/10.1016/j.cognition.2003.09.011

Taylor, D. J., & Muller, K. E. (1996). Bias in linear model power and sample size calculation due to estimating noncentrality. *Communications in Statistics-Theory and Methods*, *25*(7), 1595–1610. https://doi.org/10.1080/03610929608831787.BIAS

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

Van Gulick, A. E., McGugin, R. W., & Gauthier, I. (2016). Measuring nonvisual knowledge about object categories: The Semantic Vanderbilt Expertise Test. *Behavior Research Methods*, *48*(3), 1178–1196. https://doi.org/10.3758/s13428-015-0637-5

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

Yue, X., Tjan, B. S., & Biederman, I. (2006). What makes faces special? *Vision Research*, *46*(22), 3802–3811. https://doi.org/10.1016/j.visres.2006.06.017