Age-related differential item functioning in tests of face and car recognition ability

Mackenzie A. Sunday Vanderbilt University, Nashville, TN, USA

Woo-Yeol Lee Vanderbilt University, Nashville, TN, USA

Isabel Gauthier Vanderbilt University, Nashville, TN, USA

The presence of differential item functioning (DIF) in a test suggests bias that could disadvantage members of a certain group. Previous work with tests of visual learning abilities found significant DIF related to age groups in a car test (Lee, Cho, McGugin, Van Gulick, & Gauthier, 2015), but not in a face test (Cho et al., 2015). The presence of age DIF is a threat to the validity of the test even for studies where aging is not of interest. Here, we assessed whether this pattern of age DIF for cars and not faces would also apply to new tests targeting the same abilities with a new matching task that uses two studied items per trial. We found evidence for DIF in matching tests for faces and for cars, though with encouragingly small effect sizes. Even though the age DIF was small enough at the test level to be acceptable for most uses, we also asked whether the specific format of our matching tasks may induce some age-related DIF regardless of domain. We decomposed the face matching task into its components, and using new data from subjects performing these simpler tasks, found evidence that the age DIF was driven by the similarity of the two faces presented at study on each trial. Overall, our results suggest that using a matching format, especially for cars, reduces age-related DIF, and that a simpler matching task with only one study item per trial could reduce age DIF further.

Introduction

There is growing interest in measuring individual differences in visual abilities, an approach that can offer evidence as to the number, real-world relevance, and developmental origins of visual mechanisms (Wilmer, 2008). This approach requires the creation of measurement tools that are sufficiently reliable to provide information about individual differences, which is often not the case with standard cognitive tasks (e.g., Hedge, Powell, & Sumner, 2017; Ross & Gauthier, 2015). Another important psychometric property is that of measurement invariance across groups: We want to be confident that we can test the same ability in individuals of different groups (e.g., females vs. males or younger vs. older individuals). This is required to interpret group differences. For instance, many studies report a small advantage for women in tests of face recognition (e.g., Lewin & Herlitz, 2002; Lovén, Herlitz, & Rehnman, 2011), but to interpret this as reflecting a true difference on the same underlying unidimensional construct, it is necessary to ensure that the difference does not stem from bias within the test. Even when group effects are not of interest, it is important to ensure that a test is not biased against individuals from a certain group.

Not only is the issue of test bias important for practical and ethical reasons, but it is also of scientific importance, as the presence of bias in a given measure can indicate that the construct of interest is not fully understood (Millsap & Everson, 1993). In the study of intelligence and personality but also in new areas of individual differences study, like decision-making and neuroimaging, the use of item response theory (IRT) provides powerful means of understanding constructs and bias that go beyond classical item theory (Klein Entink, Kuhn, Hornke, & Fox, 2009; Millsap, 2010; Thomas et al., 2013; van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011). IRT models have several advantages over classical models, as they provide interpretation of individual performance that go beyond simple sum scores on a test. To put it simply, two individuals with the same total score but with a very different pattern of errors across individual items (or trials) may have different true levels of ability, and using IRT, researchers can build a model of the information that each item provides about the target ability and provide better ability estimates for each person. IRT also provides estimates of precision at

Citation: Sunday, M. A., Lee, W.-Y., & Gauthier, I. (2018). Age-related differential item functioning in tests of face and car recognition ability. *Journal of Vision*, *18*(1):2, 1–17, https://doi.org/:10.1167/18.1.2.

https://doi.org/10.1167/18.1.2 Received December 9, 2016; published January 5, 2018 ISSN 1534-7362

both the individual and test-level and can reduce reliance on normative samples (Wilmer et al., 2012). IRT models can provide information not only about the test taker, but also about the test itself. They can also be particularly helpful during test development and validation. Among other types of information, IRT methods provide ways to match individuals on the underlying target ability so that we can evaluate whether certain items on a test are biased against them in some way. These issues are important in high-level vision, as we are just beginning to characterize the manner in which visual abilities vary in the normal population, for instance across development (Germine, Duchaine, & Nakayama, 2011).

The present work investigates age-related differential item functioning (DIF) in two new matching tasks designed to measure face recognition and car recognition abilities. This work is motivated by earlier reports of more age-related DIF in the measurement of car recognition ability (Lee, Cho, McGugin, Van Gulick, & Gauthier, 2015) than in measurement of face recognition ability (Cho et al., 2015). The presence of considerable age bias in the measurement of car recognition ability is a concern because many questions in high-level vision use a comparison between different categories, very often face and nonface objects, with cars often serving as the sole nonface category (Richler, Wilmer, & Gauthier, 2017; Shakeshaft & Plomin, 2015). Here, we postulated that age DIF in man-made categories like cars could be amplified in tests that focus on a small number of target objects, and we created new tests that use a larger number of different objects. Our hope was to create tests of face and car recognition ability with minimal age DIF. Though here we use DIF analyses to investigate possible bias in specific measures of face and car recognition, similar issues having to do with the test format or the nature of the domain tested and how it relates to various groups in the population should be broadly relevant and will need to be considered as authors create new measures of visual abilities.

Studies 1 through 3 present the new tests and offer evidence of small and comparable age DIF for the two categories at the whole test level, thereby providing useful tools for future work. However, the finding of even a small but significant level of age-DIF on the new face test was unexpected based on previous results on another test of face recognition ability (Cho et al., 2015). Study 4 was therefore designed to explore the role of age-related strategies that may explain the small but significant age-DIF in the face matching task.

The first test of visual ability subjected to DIF analysis was the Cambridge Face Memory Test (CFMT; Duchaine & Nakayama, 2006), one of the most commonly used tests of face recognition for studies in the normal population. In this task, subjects study six faces and are tested in a series of trials where they have to find which of three faces is one of the original six faces. Several tests of nonface object recognition have been created using the same format (Dennett et al., 2012; McGugin, Richler, Herzmann, Speegle, & Gauthier, 2012; Richler et al., in press; Van Gulick, McGugin, & Gauthier, 2015). Two recent studies employed IRT methods to test the assumption of invariance across groups in the CFMT (Cho et al., 2015) and in a similar test of car recognition, the Vanderbilt Expertise Test for cars (VET-Car; Lee et al., 2015). To evaluate bias within tests, IRT DIF can identify differences in item parameters or item response functions after controlling for differences in levels of performance on the latent trait. In other words, DIF analysis asks whether people in different groups who have the same level of latent ability tend to make similar errors on the test.

Cho et al. (2015) found that, at the test level, the CFMT showed very little DIF across gender groups, across individuals tested online or in the lab, and across younger versus older individuals. In contrast, while Lee et al. (2015) found no evidence of test-level gender DIF in the VET-Car, they found a significant number of DIF items between age groups (<27 or ≥ 27 years). When the age-related DIF was further investigated, several items on the test (which used modern car makes and models) were easier for the younger subject group than the older group and this resulted in a substantial test-level bias. The authors postulated that age-related DIF could be a challenge when measuring abilities with familiar object categories for which different age groups might have different experiences.

Indeed, it is well known that experience with an object category can improve recognition accuracy (Van Gulick et al., 2015), shorten reaction times for individual identification as compared with categorical classification (Tanaka & Taylor, 1991) and increase holistic processing of objects within the category (Bukach, Philips, & Gauthier, 2010; Gauthier, Williams, Tarr, & Tanaka, 1998; Wong, Palmeri, & Gauthier, 2009). Experience with objects can vary for different reasons, including the presence of certain objects in our environment and our interests, and age is one factor that can influence our interests and environment. In some domains, objects may vary by location but do not systematically change over generations (like birds or mushrooms), and experience may be similar across age cohorts. However, for domains that tend to change over time (like cars or planes), experience is likely to vary with age cohorts. For example, a 60-year-old individual likely bought his or her first car in the 1970s and has seen the almost four decades worth of cars since. Compared with the average 20-year-old, who only rarely even sees a car manufactured before the year 2000, the older individual's experience perceiving cars has been wildly more

variable than the Millennial. There may also be an age at which objects from a certain category are most attended, so that each person may be most familiar with car models that are available during their working years.

Therefore, one interpretation of the difference between no age DIF for the CFMT (found in Cho et al., 2015) and age DIF for the VET-Car (found in Lee et al., 2015) is that any test of car recognition ability would be more susceptible to age DIF than tests of face recognition ability because every car has an associated epoch in history when it was most popular, which can interact with the age of the individuals taking the test. By this account, the CFMT would not show age DIF because faces are not as associated with epochs that vary as rapidly as for cars.[1] An alternative account is that the finding of age DIF for the VET-Car is specific to this particular test in this particular domain (for example, because of idiosyncrasies in the stimuli used on the test), with no implication for other tests of face or car recognition.

Here we ask if the same pattern of age-related DIF (age DIF for car but not face recognition) would be found using other tests targeting the same recognition abilities. To accomplish this, we first created two new measures of face and car recognition ability that do not depend on learning a small set of examples, as do the CFMT and VET-Car, but should otherwise tap into similar abilities: the Vanderbilt Face Matching Test (VFMT) and Vanderbilt Car Matching Test (VCMT). To be clear, our goal was not to investigate aging mechanisms in object recognition nor was it to explore differences between face and car recognition abilities, but rather to investigate the generality of the previously found greater age-DIF for cars than faces. As a byproduct of rejecting this general rule, we offer new tests in a new format that provide means of testing subject samples that vary in age on both face and car recognition abilities without having to worry about age bias.

Aside from these goals, the new field of research on individual differences in high-level vision can benefit from the availability of multiple tests for face and object recognition abilities to facilitate their measurement as latent variables (Borsboom, Mellenbergh, & Van Heerden, 2003). Recent evidence shows that different measures that would otherwise not show a correlation because the underlying constructs are not related can show inflated correlations when they both contain a great deal of stimulus repetition (i.e., small stimulus sets; Richler, Floyd, & Gauthier, 2015). The creation of new matching tasks in which each trial on the test uses a new set of exemplars of faces (Study 1) and of cars (Study 2) will help reduce the possibility of inflated correlations since stimulus repetitions in these matching tasks will be minimized. To determine if the previous age-DIF results (Cho et al., 2015; Lee et al., 2015) generalize to tests using different paradigms, in Study 3 we used IRT methods to test for DIF between age groups in these new matching tests. As mentioned before, Study 4 was motivated by the surprising finding of significant (although small) age DIF on the face matching test, and was thus designed to investigate what aspects of the matching task may be responsible for these surprising effects.

Study 1

The VFMT was designed to measure face recognition ability using a series of independent matching items. We chose to use a matching task in which subjects were only required to match faces across one item, not the entire test, so that the ability to learn about individual faces over the course of the test was not relevant to performance.

Using a matching task to measure face recognition is not new (e.g., Burton, White, & McNeill, 2010; Konar, Bennett, & Sekuler, 2010), but we made a few critical modifications. On each item, subjects study two faces for 4 s and are then presented with a three-alternative forced choice, only one of which matches one of the study faces in identity. We chose to use two different study faces on each item to preserve one property of the CFMT, uncertainty about which target is relevant on any given item. The correct response was a different image of the same face identity to ensure subjects matched identity, not images. To better discriminate subjects across all face recognition ability levels and achieve good reliability, we varied item difficulty by including foils of varying similarity to targets. We varied face position and perspective, and similarity between the faces presented as study face pairs. The test was refined over several iterations.

While the CFMT uses only Caucasian male faces, previous work found that face learning tasks with male and female faces yielded scores that are highly correlated (Ryan & Gauthier, in press). Therefore, to increase construct coverage, we used Caucasian faces from both genders. This allowed us to alternate gender on each item, to reduce the potential for proactive interference (Underwood, 1957) because each item was at least clearly different from the immediately preceding item. Because experience with different races can affect face recognition (e.g., Tanaka, Kiefer, & Bukach, 2004), we only included Caucasian faces. However, just like the original CFMT, this choice may reduce the VFMT's validity for people who do not have extensive experience with Caucasian faces. In this study, we describe the VFMT and assess its validity by relating it to the CFMT.

Methods

Subjects

VFMT piloting

One thousand and twenty-five subjects were recruited from Amazon Mechanical Turk for the VFMT and compensated $0.65. Several studies have provided evidence that online crowdsourcing tools like Amazon Mechanical Turk can provide high-quality data when used correctly (Buhrmester, Kwang, & Gosling, 2011; Crump, McDonnell, & Gureckis, 2013; Hauser & Schwarz, 2015; Mason & Suri, 2012; Paolacci, Chandler, & Ipeirotis, 2010). For all experiments, only those who had at least 95% of their previous Amazon Mechanical Turk tasks accepted were eligible to participate. We restricted subjects to those with U.S. IP addresses. Thirty-two subjects were excluded for failure to follow instructions, leaving 993 subjects (352 male; mean age = 38.4 years, range = 20–77 years) in the analyses. Of the 993 subjects who completed the VFMT, 80.3% were Caucasian, 7.0% African-American, 5.6% Hispanic/Latino, 4.2% Asian, 2.0% other, 0.5% Pacific Islander, and 0.3% Native American/Alaskan.

Comparison of VFMT and CFMT

Two hundred fifty subjects were recruited from Amazon Mechanical Turk to complete the CFMT. We contacted subjects 1 day after they completed the CFMT and offered them the opportunity to complete the VFMT and other tests not reported here. Subjects were compensated $0.70 for completing the CFMT, and a total of $5.00 for the other tasks. One hundred and fifteen subjects chose to complete the four additional tasks. Eleven subjects were excluded for failure to follow instructions or failure to respond correctly to both catch items in the other tasks. Of the 104 subjects who completed the VFMT (35 male; mean age = 37.60 years, age range = 20–70 years) and satisfactorily completed all tasks, 85.6% were Caucasian, 4.8% were Asian, 3.9% were Hispanic/Latino, 2.9% were African-American, and 2.8% identified as other.

Cambridge Face Memory Test

We used the long version of the CFMT (Russell et al., 2009). Subjects studied six Caucasian grayscale male target faces, then on each item had to correctly identify the target face presented with two foil faces. The first block of 18 items showed target faces in the studied viewpoint. The second block of 30 items required subjects to identify the target across variations in lighting and viewpoint, and Gaussian noise was added to novel target images in the third block of 24 items to make these items more difficult. The last block of 30 items was the most difficult, with uncropped targets and target faces in profile, both with additional noise added. Subjects were allowed to study the target images between each block and responses were unspeeded.

VFMT stimuli

Through several pilot tests, we modified problematic items (items with below chance accuracy or low correlations between the items responses and total test scores) with the intent of improving the reliability and range of scores produced by the test.

The VFMT faces were 485 images of 388 unfamiliar identities that varied in lighting, size, and perspective. Using Adobe Photoshop, the faces were converted to grayscale and a Gaussian blur was applied to all nonface areas (hair, clothing, background) using Photoshop's Gaussian blur filter with variable pixel radii in an attempt to reduce the high-spatial frequency nonface image features in the background, while preserving a more natural look to the images than would have been achieved by cropping.

VFMT procedure

On each item, two novel faces were presented for a 4-s study period, a duration sufficient for encoding two faces (Curby & Gauthier, 2007). Then, subjects were shown one face that matched the identity of one of the two study faces, along with two distractor faces (see Figure 1). The target face images at study and test were different images of the same person. Subjects were instructed to select the face that matched the identity of one of the study faces. There was no response time limit, and male and female items were interleaved. There were two catch items to screen for subjects who might not have understood the instructions in which the target face was presented with two faces of the opposite gender at test. There were three practice items with cartoon faces, and feedback was provided only on practice items and the first three test items. The VFMT had 95 items, plus two catch items.

Results

Mean accuracy on the VFMT was 58.3% (SD = 9.4%, skewness = −0.209, kurtosis = −0.098; Figure 2), and the test produced good internal consistency (α = 0.745). Mean accuracy was not significantly different when only data from Caucasian subjects were analyzed,

Figure 1. Example VFMT item. The study faces (top) were presented for 4 s, followed by the test faces (bottom). Study target face and correct response are shown with asterisks for illustrative purposes only. All images used with permission from depicted individual.

$t(1592) = -0.982$, $p = 0.326$, $d = -0.053$. Mean accuracy on the CFMT was 61.4% ($SD = 8.4\%$), with good internal consistency ($\alpha = 0.868$). The VFMT and CFMT were correlated, ($r_{104} = 0.575$, $p < 0.001$; Figure 2), resulting in a strong relation when the correlation is disattenuated for measurement error (Nunnally, 1970; $r_{corr} = 0.730$). The VFMT is available online at http://gauthier.psy.vanderbilt.edu/resources/ and at https://doi.org/10.6084/m9.figshare.5709682.v1.

Discussion

The VFMT was developed to capture variation between individuals in face recognition ability. Unlike the CFMT, the VFMT introduces new study faces on every item, making learning of study faces irrelevant to subsequent items. The test was honed to have good reliability, passing the first hurdle to make it useful for evaluating individual differences. Moreover, VFMT scores showed good convergent validity with an existing face recognition measure, the CFMT. The VFMT could therefore replace the CFMT in situations where a test without stimulus repetition is preferable, and it could be used alongside the CFMT and other face recognition tasks in research aiming to measure face recognition as a latent variable.

Study 2

We created the VCMT to measure car recognition ability using the same task format as the VFMT. As with the VFMT, we developed the VCMT through several pilot iterations. We then assessed the test's convergent validity by relating it to an existing measure of car recognition, the Vanderbilt Expertise Test-Car (VET-Car).

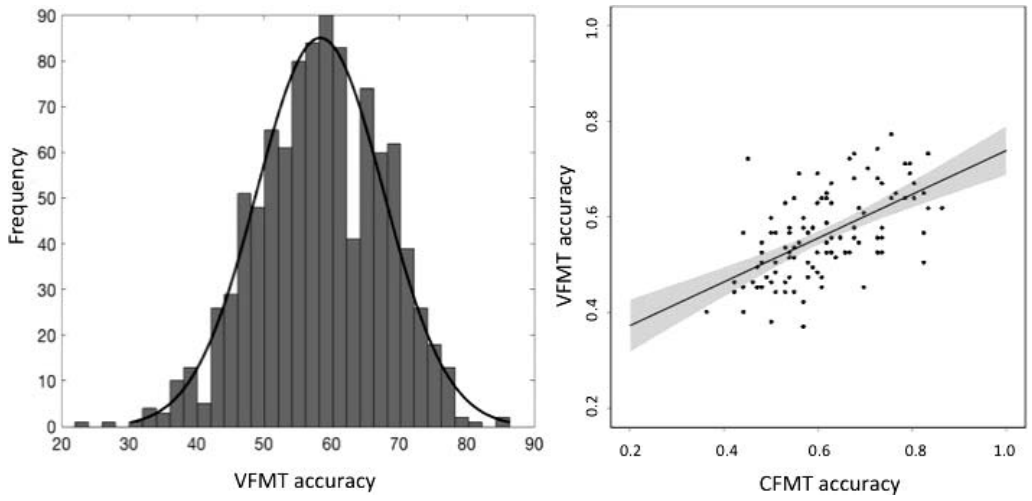

Figure 2. Left: Frequency distribution of VFMT accuracies (as percentages, $N = 993$). Chance is 33%. Right: Scatterplot of CFMT and VFMT accuracies ($N = 104$) with 95% confidence intervals represented by the shaded region.

Methods

Subjects

VCMT piloting

One thousand subjects recruited from Amazon Mechanical Turk were compensated $0.30 to complete the VCMT (395 male; mean age = 37.5 years, range = 19–75 years). Of these subjects, 79.4% were Caucasian, 7.6% African-American, 4.5% Asian, 4.3% Hispanic/Latino, 2.8% other, and 0.7% Native American/Alaskan.

Comparison of VCMT and VET-Car

One hundred and four subjects were recruited from Amazon Mechanical Turk to complete the VCMT and VET-Car. Subjects were compensated $2.00 for completing both tests. Five subjects were excluded for failure to follow instructions. Of the 99 remaining subjects (38 male; mean age = 37.01 years, age range = 21–66 years), 74.8% were Caucasian, 11.1% were African-American, 8.1% were Asian, 4.0% were Hispanic/Latino, and 2.0% identified as other.

Vanderbilt Expertise Test

We used the Vanderbilt Expertise Test (VET) as another car recognition measure. The test was developed as an object recognition measure with a similar format to the CFMT (see McGugin et al., 2012, for more details on VET development). Subjects studied six exemplars from an object category for 20 s, and were then tested with identical exemplars for six items with feedback. This was followed by another 20-s study period, then six more items with feedback. Finally, subjects completed 36 items where the target exemplar was not an identical image to the study exemplar and no feedback was provided. Different versions of the VET with different categories have been used (e.g., McGugin et al., 2012; Van Gulick et al., 2015). Here, we used the car version to compare with our new car recognition measure.

VCMT stimuli

The VCMT used 485 images of 212 car models that varied in lighting, size, and perspective. Using Adobe Photoshop, the car images were converted to grayscale and a Gaussian blur was applied to any identifying features (hood ornaments, license plates, etc). The background of the image was left unaltered, but the study-target and correct response car image did not share similar backgrounds, such that the background was not informative. All cars were modern cars models produced after the year 2000. Car images were of sedans, vans, SUVs, and convertibles.

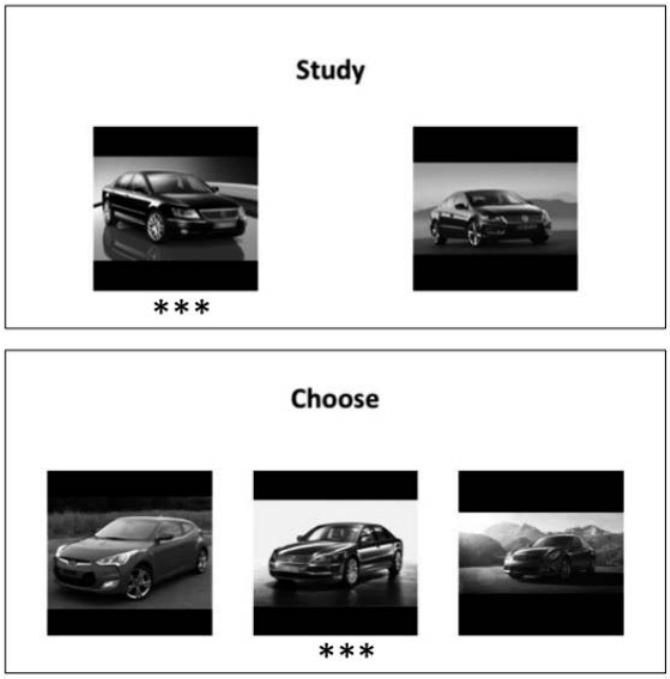

Figure 3. Example VCMT item. Study target car and correct response are shown with asterisks for illustrative purposes only.

VCMT procedure

The VCMT followed the same procedure as the VFMT. On each item, two cars were presented for 4 s to study, and then subjects were asked to choose the car image that matched one of the cars they studied from a triplet of three new car images (see Figure 3). Subjects were instructed to match the car make and model, but not year (since car models do not differ greatly from year to year). Thus, for example, an image of a 2012 Honda Civic would match another image of a 2014 Honda Civic. The study target and correct response images were different images of the same car make and model. As with the VFMT, there was no response time limit, and sedan, van, and SUV items were randomized. There were two catch items in which the target car was presented with two images of planes. Two practice items with famous cars preceded the experimental items and feedback was provided only on practice items and the first three test items. Because there is a finite number of cars models produced after the year 2000, some car models were repeated once (but at most once and using a different image of the same model) in the test. No target car model was ever repeated, only the

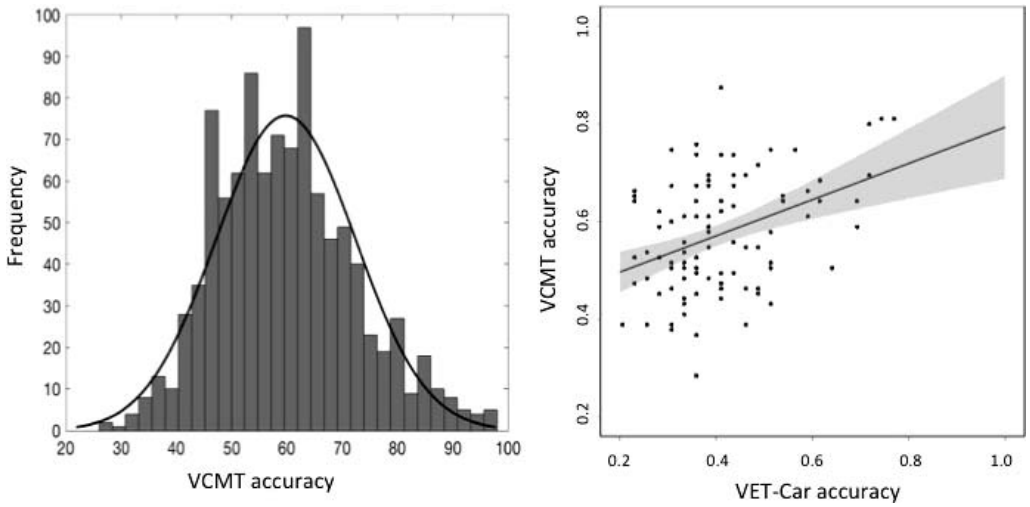

Figure 4. Left: Frequency distribution of VCMT accuracies (as percentages, $N = 1{,}000$). Chance is 33%. Right: Scatterplot of VET-Car and VCMT accuracies ($N = 99$) with 95% confidence intervals represented by the shaded region.

distractor cars. The VCMT has 95 items, plus the two catch items (totaling 97 items).

Results

The final version of the VCMT had an internal consistency of $\alpha = 0.868$ and a mean accuracy of 59.8% ($SD = 12.6\%$, skewness = 0.402, kurtosis = −0.020; Figure 4). Mean accuracy on the VET-Car was 40.2% ($SD = 12.8\%$), with moderate internal consistency ($\alpha = 0.700$). The VCMT and VET-Car were significantly correlated ($r_{99} = 0.389$, $p < 0.001$, $r_{corr} = 0.512$; Figure 4). The VCMT is also available online at http://gauthier.psy.vanderbilt.edu/resources/ and at https://doi.org/10.6084/m9.figshare.5709682.v1.

Discussion

We created the VCMT to measure car recognition ability using the same task used in the VFMT. The VCMT produced good internal consistency and score distributions. Moreover, the VCMT showed convergent validity by correlating with the VET-Car, another measure of car recognition ability. That the correlation between the VCMT and VET-Car is somewhat lower than that between the VFMT and CFMT could reflect the fact that the car domain more variable than the face domain (with cars varying more than faces both in their shape at any one time, and also across time), such that a test like the VET-Car (or the CCMT, Dennett et al., 2012) that focuses on learning only six specific cars may not provide sufficient domain coverage. That logic would suggest that the VCMT format may provide a more representative measure of "car expertise" than learning tests, an idea that could be explored in future work.

For our present purposes, armed with the VCMT and VFMT, we can now investigate whether the prior finding of more age DIF for cars than faces in learning tests generalizes to matching tasks with larger stimulus sets.

Study 3

Previous work has found more age-related DIF on the VET-Car than on the CFMT, with larger effect sizes at the test level (see Cho et al., 2015; Lee et al., 2015; VET-Car, ETSSD = −0.476; CFMT, expected test score standardized difference [ETSSD] = 0.023). While these tests used similar tasks, the two studies differed both in the stimuli categories (cars and faces for the VET-Car and CFMT, respectively) and in the use of different subject populations. Since the CFMT subjects were younger on average (Figure 5), the cutoff used to divide subjects into the younger or older group was less than that used during the VET-Car DIF analysis (CFMT –20 years Cho et al., 2015; VET-Car 27 years, Lee et al., 2015). To address this issue, we redid the CFMT and VET-Car DIF calculations using the same age cutoff.

Methods

DIF analysis of VFMT and VCMT

Data from the final VFMT and VCMT versions (reported in Studies 1 and 2, respectively) were

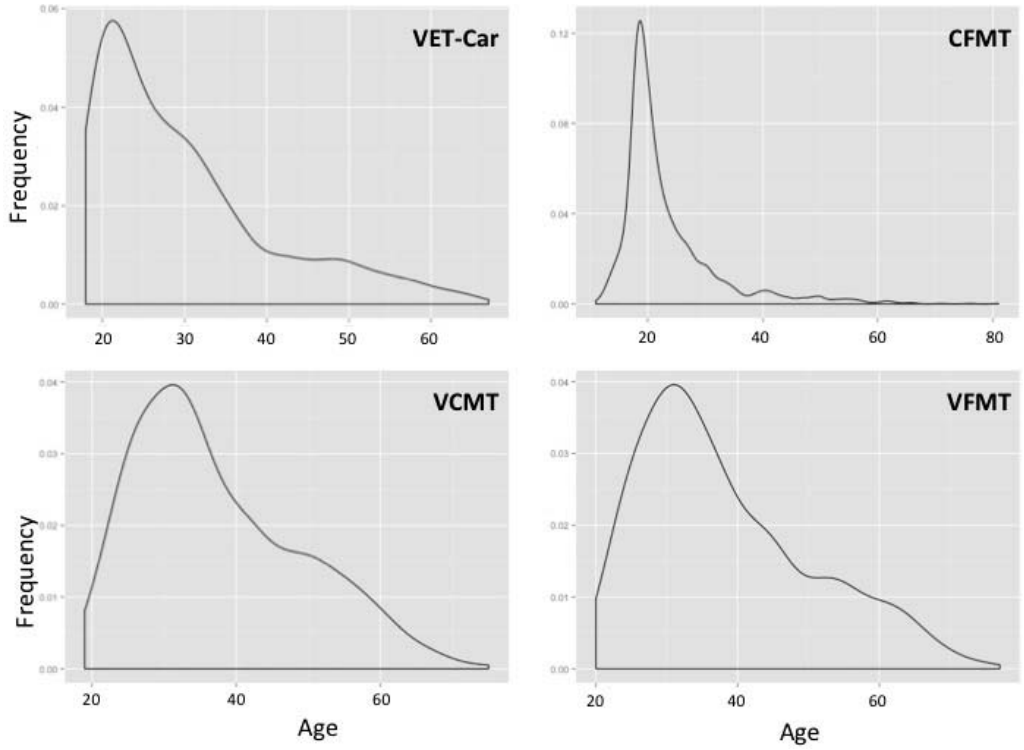

Figure 5. Density plots of subject age distributions for each of the tests discussed here: VET-car from Lee et al. (2015) and CFMT from Cho et al. (2015).

analyzed. To begin the IRT analyses, we fit a three-parameter logistic item response model to each dataset. The three-parameter model includes item discrimination, location, and guessing parameters and was chosen because the three-parameter model was used in previous work with the VET-Car and CFMT and because both the VCMT and VFMT involve three-alternative forced choices that make a guessing parameter a reasonable addition to the IRT model. The unidimensional three-parameter logistic model fit better than the unidimensional two-parameter logistic model according to the likelihood ratio test (LRT) for both the VFMT, $\chi^2(94) = 437.28$, $p < 0.001$, and VCMT, $\chi^2(95) = 788.94$, $p < 0.001$. Items with negative item discrimination parameters were excluded from further analysis (VCMT – Item 95, VFMT – Items 6, 37, 67, and 70). The average accuracy between age groups on the VFMT did not differ significantly, $t(991) = 1.543$, $p = 0.123$, $d = 0.105$. However, the average accuracy of the younger group on the VCMT was significantly higher than the older group, $t(998) = 4.686$, $p < 0.0001$, $d = 0.308$. Because DIF is found after controlling for the main effects (group difference and the item parameter of the reference group), the detection of DIF is independent of this difference in accuracies between groups.

DIF Analysis

To perform the DIF analysis, subjects were split into two groups, a younger group ($\leq$30 years) and an older group (>30 years). We chose this cutoff as a compromise to provide groups large enough to compare in the new tasks, but also allow a reanalysis of the data from Cho et al. (2015) and Lee et al. (2015) using the same cutoff (to ensure that any difference in DIF results for the new matching tasks relative to the learning tasks cannot be attributed to cutoff difference).

Similarly, analyses for the CFMT and VET-Car from previous studies (Cho et al., 2015, using publicly available online dataset; Lee et al., 2015) were repeated, but with the same 30-year cutoff as was used with the VCMT and VFMT. The number of subjects in each group for each test is reported in Table 1.

For our VCMT and VFMT analyses, we used three different methods of DIF detection: Lord's chi-square test (Lord, 1980), Raju's z-statistic (Raju, 1990), and the LRT method (Thissen, Steinberg, & Wainer, 1988). The difR package in R (Magis, Beland, & Raiche, 2013) was used to perform the Lord's and Raju's test, and the LRT test was performed using IRTLRDIF Version 2.0 (Thissen, 2001). An item was considered to show DIF if the critical value from the respective test was above the 5% significance level (7.815 for Lord's,

Test	Number of subjects		Median age of sample	Percent of subjects tested online
	Younger group (≤30)	Older group (>30)		
VCMT	317	683	35	100%
VFMT	316	692	35	100%
VET-Car	900	526	20	67.6%
CFMT	2030	308	20	57.8%

Table 1. Subject age summaries. *Notes*: VCMT = Vanderbilt Car Matching Test; VFMT = Vanderbilt Face Matching Test; VET-Car = Vanderbilt Expertise Test-Car; CFMT = Cambridge Face Memory Test.

with $df = 3$, 1.96 for Raju's, and 3.85 for LRT). These values are reported in Supplementary Table S1. The younger group was used as the reference group and the older group as the focal group. Though the designation of focal and references is theoretically arbitrary in our case, we chose to be consistent with what was previously chosen for the DIF analyses of the VET-Car and CFMT.

Age DIF analysis for VET-Car and CFMT

Because we are only interested in comparing the DIF effect sizes for VET-Car and CFMT to determine if different age cutoffs impact these effect sizes, we are not interested in identifying specific DIF items. Thus, for the sake of brevity, here we do not report DIF items and performance of each DIF detection method, but instead report the total number of DIF items identified by two out of three methods (Table 2).

Comparison between age cutoffs

We report three different measures of DIF effect size. First, we report the signed and unsigned test differences between the groups (STDS and UTDS; Meade, 2010). The STDS and UTDS are both the expected differences between percent correct of all subjects in the focal group that is solely due to DIF. While the STDS is signed, the UTDS is not, meaning that UTDS does not allow for cancellation across items or subjects. Because DIF magnitude can be either positive or negative, cancellation can occur when some DIF items have a positive DIF and others have a negative DIF, causing the DIF effect to cancel out on the test level. Thus, by comparing STDS and UTDS values, we can determine the type of DIF for that item. A large difference between STDS and UTDS values could be interpreted as DIF on that item that works in opposing directions, such that it cancels out when the STDS is calculated but produces a large UTDS value. The ETSSD (Meade, 2010) is also reported. Because this statistic expresses DIF magnitude on the standardized scale at the test level, we can compare the DIF effect sizes between tests having a different metric. To calculate DIF effect sizes the item parameters were first estimated for each group using IRTLTDIF software. Next, items were deleted if the item discrimination parameters were unusually high (>100), then item parameters were re-estimated using the remaining items. This step was repeated until all item discrimination parameter estimates were less than 100. Last, the DIF effect sizes were calculated using VisualDF program (Meade, 2010) based on the final item estimates.

Results

DIF analysis of VFMT and VCMT

The DIF results for VCMT and VFMT are reported in Supplementary Table S1. Items were considered as showing DIF if they showed significant DIF in two out

Test	Number of total items	Percent of items showing DIF	STDS	UTDS	ETSSD
VET-Car	44	81.25%	−0.323	2.089	−0.077
CFMT	72	18.06%	0.023	1.979	0.003
VCMT	93	65.29%	−0.165	3.405	−0.019
VFMT	89	51.58%	−0.070	3.152	−0.013

Table 2. Summary of items with differential item functioning (DIF) for the four tests as well as effect sizes when the age cutoff was at 30 years. *Notes*: In the Table, items were considered as showing DIF if they showed significant DIF in two out of the three DIF detection methods (Lord, Raju, and LRT). The percentage of items showing DIF is reported, as well as the signed test differences (STDS), unsigned test differences (UTDS), and expected test score standardized difference (ETSSD) values. VCMT = Vanderbilt Car Matching Test; VFMT = Vanderbilt Face Matching Test; VET-Car = Vanderbilt Expertise Test-Car; CFMT = Cambridge Face Memory Test.

of the three DIF detection methods, so that any item considered showing DIF had a degree of convergence between DIF detection methods (though this criterion is admittedly arbitrary, Supplementary Table S1 reports the results from each method—Lord, Raju, and LRT). As a result, 62 items (65.29%) in VCMT and 49 items (51.58%) in VFMT were found as DIF items.

Age DIF analysis for VET-Car and CFMT

Results from the DIF analysis of the VET-Car and CFMT using the 30-year age cutoff are reported in Table 2. For the VET-Car, 39 out of 48 items (81.25%) were identified as DIF items. For the CFMT, 13 out of 72 items (13.06%) were identified as DIF items.

Comparison of DIF test-level effect sizes

Because the sample size varied between measures both in total number of subjects and the number of subjects in the focal groups, we examined the effect sizes of the DIF found in all four tests. Test-level bias is arguably more important than item-level bias, since these tests are typically used to make inferences at the level of the ability the entire test is meant to measure. Forty-four items of VET-Car were included for DIF effect sizes analysis. Items 12, 40, 44, and 46 were deleted due to the extreme discrimination parameter estimate. All 72 items of CFMT were included for DIF effect sizes analysis. Ninety-three items of VCMT were included. Items 40 and 95 were excluded. Six items (Items 6, 37, 62, 67, 70, and 76) of VFMT were excluded due to high discrimination estimates.

Results are summarized in Table 2. The VET-Car, VCMT, and VFMT all have negative average STDS values, meaning that if two groups had the same latent recognition ability, the older group would nonetheless be expected to have a percent correct that is 0.734, 0.177, or 0.079 percentage points lower than the younger group on the VET-Car, VCMT, and VFMT respectively. Conversely, the older group would be expected to score 0.032 percentage points higher on the CFMT.

For all tests, we found small effect sizes at the test level, expressed as ETSSD values, with absolute values all less than 0.08 (according to the guideline of Cohen's *d*; Cohen, 1988), these effect sizes are interpreted as small; i.e., <0.2). Since the ETSSD values were small for our age cutoff of 30, especially given the larger ETSSD value initially reported with 27-year cutoff for the VET-Car (–0.476; Lee et al., 2015), we postulated that perhaps something about the age cutoff of 30 was decreasing the ETSSD values. To test this, we also calculated ETSSD values for the VCMT and VFMT at 27- and 35-year cutoffs. With the 27-year cutoff, the VFMT and VCMT still produced small ETSSD values (VFMT: ETSSD = –0.033; VCMT: ETSDD = –0.006). The same was found at the 35 year cutoff (VFMT: ETSSD = –0.061; VCMT: ETSDD = –0.034). Thus it seems as though these small ETSSD values (that is, the much smaller age DIF in the matching tasks compared to that for the VET-Car in Lee et al., 2015) cannot be explained by the age cutoff used to divide the sampling distributions.

Discussion

Using IRT, we examined the VCMT and VFMT for DIF between age groups and revisited previous DIF analyses of the VET-Car and CFMT. Our analyses addressed two questions: (a) Was the smaller DIF for faces in the CFMT in Cho et al. (2015) compared to cars in the VET-Car in Lee et al. (2015) due to the different age cutoffs? and (b) Is the pattern of car recognition showing more DIF than face recognition specific to the tasks used on the VET-Car and CFMT or will it extend to the new matching tasks?

To answer the first question we redid the VET-Car and CFMT effect size calculations using the same age cutoff for both (30 years). With this new cutoff age, the VET-Car showed a considerably smaller test-level effect size than in the analyses by Lee et al. (2015), but the CFMT still did not show any evidence of DIF. Even with the diminished effect size, the VET-Car still had a larger DIF effect size than the CFMT, so at the very least the qualitative pattern of more DIF for cars than faces was found regardless of cutoff. Nonetheless, DIF in the VET-Car was sensitive to the cutoff, and because the age distribution of the samples we gathered for the matching tasks is relatively similar to that of the Lee et al. study, it highlights the importance of comparing different cutoffs. It is possible that the higher age DIF in the VET-Car than the CFMT was not entirely due to the difference in cutoff as it was to the different age distributions of the samples in those prior studies. However, if that were the case, we would expect that the new matching tasks would also show a large amount of DIF, similar to the VET-Car.

To address this and our second question, we performed a DIF analysis on the VCMT and VFMT. We found that the VCMT had a slightly higher DIF effect size on average compared with the VFMT when the age cutoff was 30 years. Conversely, the VFMT has a slightly higher DIF effect size than the VCMT when the age cutoff was at 27 and 35 years. Importantly, though approximately half of VCMT and VFMT items showed evidence for DIF, the test-level DIF effect sizes were quite small, and no cutoff we used resulted in DIF

effects as large as those observed for the VET-Car by Lee et al. (2015).

Thus, to answer our initial question, these results provide little support for the idea that any testing with cars would produce large age DIF while any testing with faces would not. It appears that either the format of the VET-Car, perhaps the repeated use of a few specific cars, made the test particularly sensitive to age cohort effects. It would be interesting to assess whether other learning tasks with a few car models (for instance the Cambridge Car Memory Test, Dennett et al., 2012) would also show large age DIF effects.

From a practical perspective, the finding that matching tasks led to only minimal test-level age DIF effects is reassuring. Our critical finding of small DIF effect sizes on both the VFMT and VCMT means that unless even a small amount of age bias is a concern, these tests can be used for most research purposes. While the average test-level DIF effect sizes were small for both tests, we did find that both tests had many items with DIF. From a theoretical perspective, finding many items with DIF in the VFMT is especially interesting, as age DIF cannot be easily attributed to cohort-specific face stimuli. It is true that there is an own-age effect in face recognition (Anastasi & Rhodes, 2005; He, Ebner, & Johnson, 2011; Picci & Scherf, in press), but because the VFMT used Caucasian adult faces just as in the CFMT, this does not seem a plausible explanation for the age DIF in the VFMT. Instead, and unexpectedly, it seems there might be a domain-general type of age DIF in the matching tests that was not present in the learning tasks (because it was not observed in the CFMT). We sought to explore this surprising finding further in Study 4, to see if something about the task structure of the VFMT interacted with age. While we specifically focused on the age DIF we observed in the VFMT because it presented such a contrast with the absence of age DIF in the CFMT, our goal in Study 4 was to investigate whether the task structure used in both the VFMT and VCMT may have associated age DIF.

Study 4

Here, we investigate if some of the DIF found in the VFMT may be due to different age groups employing different strategies on the specific matching task we used. We use the VFMT rather than the VCMT because the difference in DIF between the VFMT and CFMT (with the CFMT showing no DIF and the VFMT showing some), is more striking than that between the VCMT and VET-Car. In addition, because the age DIF found in tests using face stimuli is much more difficult to attribute to cohort-specific effects, it begs for an explanation. As a reminder, motivated by a desire to preserve some of the features of the learning tasks such as having to hold several targets in working memory, the VFMT requires subjects to encode two faces and then compare the representations of these faces in their memory to the triplet of possible matching images. Accordingly, item difficultly in this task could be modulated by the similarity of encoded images, the similarity of these images to the options, and the similarity of the options to each other. For the sake of simplicity, we grouped these similarities into four comparisons (Figure 6) and collected data in four simpler matching tasks that together assess most of the similarities that may be relevant to the more complex VFMT task. We then quantified the extent to which variability across items on these four tasks predict the magnitude of age DIF of the corresponding VFMT items.

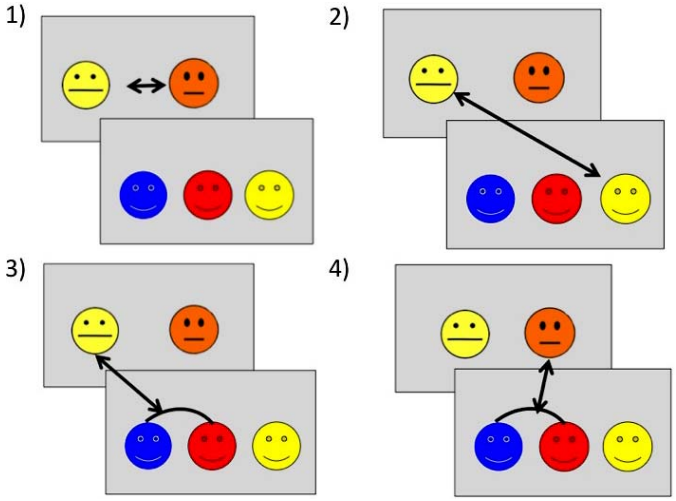

Figure 6. Cartoon graphic of each of the four similarity comparisons. Yellow faces represent study and test target faces, the orange face represents the study foil, and red and blue faces the test distractors. (1) Comparison 1 of study target and study-foil; (2) Comparison 2 of study target and test target; (3) Comparison 3 of study target and both test distractors; (4) Comparison 4 of study foil and test distractors.

The simpler matching tasks are different from the parent task used in the VFMT and so we cannot assume that subjects employ the same strategy in these different paradigms. Our specific goal in this study is to ask whether the similarity of different subsets of faces relevant to each trial can help explain the source of age DIF in the VFMT. Our broader goal is to determine if the task structure used on both the VFMT and VCMT may account some of the age DIF we found in both tests in Study 3. Note that we did not analyze the results of the simpler tasks based on age, as we are here primarily interested in variability across items in the various tasks, rather than variability across subjects. While it is possible that there may be age DIF in the

simpler tasks, we cannot determine this at the sample size we collected here.

Methods

To quantify how each type of similarity influenced the VFMT items, we had an independent sample of subjects perform a classic same/different matching task for each comparison for each VFMT item. The same/different matching task differs from the task used in the VFMT and so we cannot expect subjects to necessarily employ the same strategy on these two differing paradigms. By using the same/different tasks, we hope to deconstruct the VFMT paradigm to better understand how the relative similarities of different pairs of faces in the items may relate to differences in strategies that different age groups may have used on the original task. Because the VFMT used accuracy and not reaction time as a dependent measure, we also focused here on accuracy: The more people were able to say that two images depicted the same person (or a different person), the more similar (or dissimilar) they were inferred to be.

Stimuli

For Comparisons 1 and 2 (Figure 6), the study-target face was presented for 2 s followed by either the foil (Comparison 1) or the test-target face (Comparison 2). Subjects were instructed to click "same" if the faces were of the same person and "different" if the faces were of two different people. This was done for each of the 95 VFMT items, totaling 190 items. All of the items were presented in random order.

For Comparison 3 (Figure 6), the study-target face was again presented for 2 s, followed by the two distractor faces shown together. Subjects were tasked with clicking "yes" if either of the two faces was the same person as the study face and "no" if neither of the faces was the same person as the study face. Because the answer for Comparison 3 is "no" for every comparison, we included 45 items in which one of the two test faces was the study target so that one-third of the items would have a correct response of "yes." Items were randomized and totaled 140. Comparison 4 (Figure 6) was prepared in a similar way to Comparison 3, except because all the correct responses for Comparison 4 would be "yes," we included 45 "no" items. The tests of Comparisons 1 and 2 (combined) took approximately 25 min to complete and Comparisons 3 and 4 took approximately 20 min to complete.

Model and predictor	*B*	*SE*	*t*	*p*
VFMT item accuracy ($R^2_{adj} = 48.3\%$)				
Intercept	−0.383	0.121	−3.180	0.0021
Comp 1	−0.074	0.083	−0.892	0.3750
Comp 2	0.623	0.077	8.090	≤0.0001
Comp 3	0.353	0.104	3.390	0.0011
Comp 4	0.397	0.093	4.270	≤0.0001

Table 3. Results of a multiple regression predicting Vanderbilt Face Matching Test (VFMT) item accuracy.

Subjects

Subjects were recruited from Amazon Mechanical Turk and were compensated $1.00 to complete the Comparison 3 and 4 tests and $1.30 to complete the Comparison 1 and 2 test (so that each task used to get comparison data took approximately the same amount of time). Because stimulus repetition can inflate correlations between tests when it is a shared feature (Richler et al., 2015), we used separate samples from the VFMT sample in Study 3. One hundred ten subjects were recruited for Comparisons 1 and 2. Ten were excluded for failure to follow instructions, leaving 100 subjects for analysis (60 male; mean age = 34.45 years, range = 19–71 years). One hundred eleven subjects were recruited for Comparison 3. Eleven were excluded for failure to follow instructions, leaving 100 subjects for analysis (43 male; mean age = 36.84 years, range = 18–74 years). One hundred sixteen subjects were recruited for Comparison 4, but nine were excluded for failure to follow instruction, leaving 107 subjects (56 male; mean age = 36.48 years, range = 20–74 years).

Results

The four comparisons had similar mean accuracies (1: 79.2%, *SD* = 15.7%; 2: 73.7%, *SD* = 17.2%; 3: 74.3%, *SD* = 13.6; 4: 76.2%, *SD* = 14.5%). To assess how each comparison contributed to the difficulty of each VFMT item, a multiple regression was performed with the four comparisons as predictors of the overall VFMT item accuracy. Comparisons 2, 3 and 4 were all significant predictors of VFMT item accuracy, and as a whole, even though they were measured in a separate group of subjects, accounted for 48% of the variability in performance in the task (Table 3).

To examine how the similarity comparisons differed for VFMT items that showed age-related DIF, we looked at how the average similarity comparison accuracies differed between items with uniform (as opposed to nonuniform) DIF and compared this with items that showed no DIF. These items were categorized as either items with higher difficulty parameters

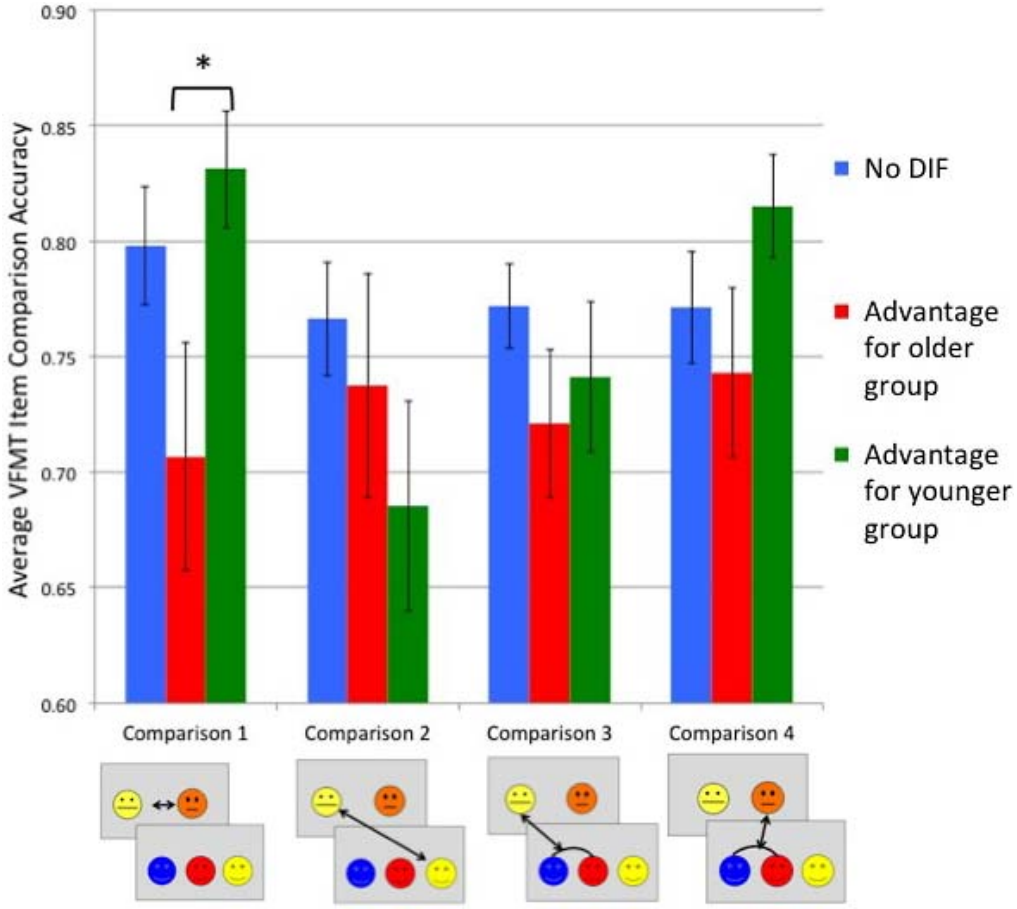

Figure 7. Average accuracies (in decimal form) of each similarity comparison, graphed for items with higher difficulty parameters for younger subjects (red), older subjects (green), and items with no DIF (blue). Chance was at 0.50 for all comparisons.

for the older group (Items 9, 10, 19, 22, 23, 29, 38, 50, 51, 58, 59, 60, 65, 77, and 86), items with higher difficulty parameters for the younger group (Items 12, 20, 25, 26, 28, 32, 48, 52, 64, 71, 73, 78, 80, 89, 92, and 93), and items which showed no DIF (Items 1, 2, 3, 4, 5, 7, 13, 14, 15, 16, 17, 18, 21, 24, 27, 30, 31, 33, 35, 36, 39, 40, 41, 43, 44, 45, 46, 47, 55, 61, 63, 66, 68, 72, 74, 79, 81, 82, 83, 91, and 94).

Within each comparison, there was only one statistically significant difference between groups of items: VFMT items that advantaged younger subjects had study targets that were less similar to the study foil (Comparison 1) than those items that advantaged older subjects, $t(29) = 2.192$, $p = 0.037$, $d = 0.796$ (Figure 7). In addition, when comparing the relative difficulty of the different comparisons (correct rejection for Comparisons 1, 3, and 4, hit rate for Comparison 2), items that showed no DIF or an advantage for older subjects showed no differences, whereas items that advantaged younger subjects showed poorer performance for Comparison 2 relative to both Comparisons 1 and 4 (Comparisons 1 and 2: $t[14] = -2.982$, $p = 0.001$, $d = -1.594$; Comparisons 4 and 2: $t[14] = -2.384$, $p = 0.032$, $d = -1.274$), and the accuracy for Comparison 3 was also lower than for Comparison 1 ($t[14] = 2.619$, $p = 0.020$, $d = 1.400$). In other words, items that advantaged younger subjects were those that had relatively distinct study faces, test foils that were not similar to the study foil, but in which the test target was not highly similar to the study target. It is unclear whether this reflects a different strategy (e.g., emphasis on the distinguishing features of the two study faces) and/or a difference in ability to generalize across different images of the same person (for Comparison 2).

Discussion

Our goal in Study 4 was to explore what aspects of the matching task used in the VFMT may have given rise to the age DIF found in Study 3. While the DIF on the VFMT was small in terms of effect size (which is good news for practical purposes), the fact that many VFMT items showed age DIF is puzzling theoretically. It is not easily explained by the account given for the larger DIF effects obtained in the VET-Car (Lee et al., 2015). The stimulus set used on the VFMT does not

easily lend itself to the sorts of cohort explanations we can invoke for cars (that is, people of a certain age are more familiar with certain car models). Because the VFMT shows more age DIF than the CFMT, and because the VFMT and VCMT both show small but similar levels of DIF, we postulated that the task format may lead to these effects. Specifically, we hypothesized that different age groups may employ different strategies to accomplish the same matching task and investigated this by comparing how the similarity between different faces in each VFMT item accounted for whether these items favored the younger age group versus the older age group.

We found that the similarity between the study-target and test-target distractors, study-target and test-distractors, and study-foil and test-distractors (Comparisons 2, 3, and 4) accounted together for almost half of the variance in the total score item accuracy. This indicates, perhaps not surprisingly, that the image similarities on a given VFMT item modulate this item's difficulty. Interestingly, the similarity between the study-target and study-foil (Comparison 1) did not significantly contribute to items' overall accuracy (when age was not taken into account). Nonetheless, this comparison differed between items with DIF that favored younger versus older subjects. This may be because the similarity between the two study items was related to test performance in different ways for younger and older subjects. Though the work here was not motivated by work on eyewitness identification, this work does show some evidence of an age-related difference in eyewitness identification, with older witnesses making more false alarms than younger witnesses (Lamont, Stewert-Williams, & Podd, 2005; Memon, Bartlett, Rose, & Gray, 2003; Searcy, Bartlett, & Memon, 1999). These age differences may arise due to differences in the strategy used by older adults during suspect identification compared to that used by younger adults (specifically by relying more on familiarity; Edmonds, Glisky, Bartlett, & Rapcsak, 2012).

On items with study-targets and study-foils that were more similar, an older subject with the same ability level as a younger subject would be more likely to choose the correct response (i.e. the item favors older subjects). A similar trend was observed for similarity Comparison 4: When the study-foil and test-distractors were more similar, the item favored older subjects. Both similarity Comparisons 1 and 4 involve the study-foil, suggesting that any age-related differences may arise during encoding of the study faces (keeping in mind that on this task, subjects are not aware of which study stimulus is relevant). It is possible that when the faces are similar, older subjects sometimes give the correct response for the wrong reason (matching the test-target to the study-foil). Further work is needed to better pinpoint how strategies differ between age groups, but follow-up work should consider testing whether eliminating the study-foil decreases the amount of age-related DIF observed (which we note was, from a practical standpoint, small). It will also be important for future work to determine if these results found with faces in the VFMT extend to cars in the VCMT, and by extension to any category that would be tested using a similar format.

General discussion

We investigated the generalizability of age-related DIF found in car and face recognition tests. Our goal was to test if previous findings of age-related DIF in car recognition but not face recognition extended to measures of these abilities that used a different task. To do so, we first created the VFMT, which was designed to measure face recognition by presenting two faces to study followed by a three-alternative forced choice matching task. The final VFMT version produced acceptable reliability and a roughly normal score distribution. The test also showed convergent validity by correlating with the CFMT. We then created the VCMT in a similar way, using the same task that was used on the VFMT. This test also produced reliable scores and showed convergent validity with the VET-Car.

There are limitations to our study and interpretation, which include the fact that it is difficult to compare age DIF effects in the VFMT to their absence in the CFMT in previous work given that the sample for CFMT in Cho et al. (2015) had relatively fewer older adults. The CFMT should be examined again in a sample more variable in age before it is concluded that there is no age DIF on this test. In addition, while we provided convergent validity for the VFMT and CFMT, we did not here address divergent validity with other object categories (e.g., Van Gulick et al., 2015).

In two new tests of face and car recognition, we looked for age DIF using DIF detection methods based on IRT. We found that both the VFMT and VCMT had many items showing age-related DIF, but that both tests has small average DIF effect sizes. Moreover, we found that the moderate DIF effect sizes previously found in the VET-Car decreased when the age cutoff was increased, consistent with our conjecture that the test was sensitive to age cohort effects.

In contrast to what was found with the VET-Car, the age DIF observed on the matching tasks was small regardless of age cutoff, supporting their usefulness in future research when it is important to avoid age bias. We speculated that the small amount of age DIF observed on these matching tasks likely stemmed from

a different source than the more substantial DIF observed on the VET-Car, since the matching tasks' DIF was not sensitive to age cutoff and was comparable for faces and cars. To explore the sources of this age DIF, we compared how different aspects of the matching tasks differed between DIF items favoring the older age group and those favoring the younger age group.

We discovered that VFMT items showing DIF favoring old versus young subjects tend to have study-targets and study-foils that were more similar. While the VFMT and CFMT likely draw on the same construct, it may be that the VFMT format allows for more variability in strategy. More studies are needed to better understand what these strategies are and why these differences arise. Here, we presented two new tests of face and car recognition, which can be used with other tests, and combined using aggregation or IRT methods to reduce the influence of biases tied to strategies relevant to each specific test. More generally, for those moving into the study of individual differences from fields, such as high-level vision, that traditionally do not emphasize psychometrics, we have provided a detailed and practical example of how IRT DIF can be used to characterize bias and compare bias across tasks.

Keywords: face and object recognition, differential item functioning

Acknowledgments

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).

Commercial relationships: none.
Corresponding author: Mackenzie Sunday.
Email: mackenzie.a.sunday@vanderbilt.edu.
Address: Department of Psychology, Vanderbilt University, Nashville, TN, USA.

Footnote

[1] Of course it is possible to imagine examples using portraits from a very different era, but here we assume that the faces used on the CFMT are similar to adult faces experienced by all age groups in a contemporary sample.

References

Anastasi, J. S., & Rhodes, M. G. (2005). An own-age bias in face recognition for children and older adults. *Psychonomic Bulletin & Review*, *12*(6), 1043–1047.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203–219.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5.

Bukach, C. M., Phillips, W. S., & Gauthier, I. (2010). Limits of generalization between categories and implications for theories of category specificity. *Attention, Perception, & Psychophysics*, *72*(7), 1865–1874.

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior Research Methods*, *42*(1), 286–291.

Cho, S. J., Wilmer, J., Herzmann, G., McGugin, R. W., Fiset, D., Van Gulick, A. E., . . . Gauthier, I. (2015). Item response theory analyses of the Cambridge Face Memory Test (CFMT). *Psychological Assessment*, *27*(2), 552–566.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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

Curby, K., & Gauthier, I. (2007). A visual short-term memory advantage for faces. *Psychonomic Bulletin Review*, *14*(4), 620–628.

Dennett, H. W., McKone, E., Tavashmi, R., Hall, A., Pidcock, M., Edwards, M., & Duchaine, B. (2012). The Cambridge Car Memory Test: A task matched in format to the Cambridge Face Memory Test, with norms, reliability, sex differences, dissociations from face memory, and expertise effects. *Behavior Research Methods*, *44*(2), 587–605.

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, *44*(4), 576–585.

Edmonds, E. C., Glisky, E. L., Bartlett, J. C., & Rapcsak, S. Z. (2012). Cognitive mechanisms of false facial recognition in older adults. *Psychology and Aging*, *27*(1), 54–60.

Gauthier, I., Williams, P., Tarr, M. J., & Tanaka, J. W. (1998). Training "Greeble" experts: A framework for studying expert object recognition processes. *Vision Research*, *38*, 2401–2428, doi:10.1016/S0042-6989(97)00442-2.

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.

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*, 400–407.

He, Y., Ebner, N. C., & Johnson, M. K. (2011). What predicts the own-age bias in face recognition memory? *Social Cognition*, *29*(1), 97–109.

Hedge, C., Powell, G., & Sumner, P. (2017). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, Advance online publication, doi:10.3758/s13428-017-0935-1.

Klein Entink, R. H., Kuhn, J. T., Hornke, L. F., & Fox, J. P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods*, *14*(1), 54–75.

Konar, Y., Bennett, P. J., & Sekuler, A. B. (2010). Holistic processing is not correlated with face-identification accuracy. *Psychological Science*, *21*(1), 38–43, doi:10.1177/0956797609356508.

Lamont, A. C., Stewert-Williams, S., & Podd, J. (2005). Face recognition and aging: Effects of target age and memory load. *Memory & Cognition*, *33*, 1017–1024, doi:10.3758/BF03193209.

Lee, W. Y., Cho, S. J., McGugin, R. W., Van Gulick, A. B., & Gauthier, I. (2015). Differential item functioning analysis of the Vanderbilt Expertise Test for cars. *Journal of Vision*, *15*(13):23, doi:10.1167/15.13.23. [PubMed] [Article]

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

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New York and London: Routledge.

Lovén, J., Herlitz, A., & Rehnman, J. (2011). Women's own-gender bias in face recognition memory. *Experimental Psychology*, *58*, 333–340.

Magis, D., Beland, S., & Raiche, G. (2013). difR: Collection of methods to detect dichotomous differential item functioning (DIF) in psychometrics (R Package version 4.5) [Computer software]. Retrieved from https://cran.r-project.org/web/packages/difR

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23.

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.

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology*, *95*(4), 728–743.

Memon, A., Bartlett, J., Rose, R., & Gray, C. (2003). The aging eyewitness: Effects of age on face, delay, and source-memory ability. *Journal of Gerontology, Series B*, *58*(6), 338–345.

Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives*, *4*(1), 5–9.

Millsap, R. E., & Everson, H. T. (1993). Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*(4), 297–334.

Nunnally, J. C. (1970). *Introduction to psychological measurement*. New York: McGraw-Hill.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5), 411–419.

Picci, G., & Scherf, K. S. (2016). From caregivers to peers: Puberty shapes human face perception. *Psychological Science*, *27*, 1461–1473.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*(2), 197–207.

Richler, J. J., Floyd, R. J., & Gauthier, I. (2015). About-face on face recognition ability and holistic processing. *Journal of Vision*, *15*(9), 1–12, doi:10.1167/15.9.15. [PubMed] [Article]

Richler, J. J., Wilmer, J. B., & Gauthier, I. (2017). General object recognition is specific: evidence from novel and familiar objects. *Cognition*, *116*, 42–55.

Ross, D. A., & Gauthier, I. (2015). Holistic processing in the composite task depends on face size. *Visual Cognition*, *23*(5), 533–545.

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face

recognition ability. *Psychonomic Bulletin & Review*, *19*(2), 252–257.

Ryan, K., & Gauthier, I. (2016). Gender differences in recognition of toy faces suggest a contribution of experience. *Vision Research*, *129*, 69–76.

Searcy, J. H., Bartlett, J. C., & Memon, A. (1999). Age differences in accuracy and choosing in eyewitness identification and face recognition. *Memory & Cognition*, *27*(3), 538–552.

Shakshift, N. G., & Plomin, R. (2015). Genetic specificity of face recognition. *Proceedings of the National Academy of Sciences*, *112*(41), 12,887–12,892.

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), B1–B9, doi:10.1016/j.cognition.2003.09.011.

Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, *23*(3), 457–482.

Thissen, D. (2001). IRTLRDIF (Version 2.0b) [Computer software]. Chapel Hill, NC: L. L. Thurstone Psychometric Laboratory.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainerand H. I. Braun (Eds.), *Test Validity* (pp. 147–169). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thomas, M. L., Brown, G. G., Thompson, W. K., Voyvodic, J., Greve, D. N., Turner, J. A., & Potkin, S. G. (2013). An application of item response theory to fMRI data: Prospects and pitfalls. *Psychiatry Research: Neuroimaging*, *212*(3), 167–174.

Underwood, B. J. (1957). Interference and forgetting. *Psychological Review*, *64*(1), 49–60.

van der Maas, H. L., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: on the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, *118*(2), 339–356.

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*, 1178–1196.

Wilmer, J. B. (2008). How to use individual differences to isolate functional organization, biology, and utility of visual functions; with illustrative proposals for stereopsis. *Spatial Vision*, *21*(6), 561–579.

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Gerbasi, M., & Nakayama, K. (2012). Capturing specific abilities as a window into human individuality: The example of face recognition. *Cognitive Neuropsychology*, *29*(5–6), 360–392.

Wong, A. C.-N., Palmeri, T. J., & Gauthier, I. (2009). Conditions for face-like expertise with objects: Becoming a Ziggerin expert—but which type? *Psychological Science*, *20*, 1108–1117, doi:10.1111/j.1467- 9280.2009.02430.x.