Measuring non-visual knowledge about object categories:

The Semantic Vanderbilt Expertise Test

Ana E. Van Gulick[1,2], Rankin W. McGugin[3], and Isabel Gauthier[3]

[1] University Libraries, Carnegie Mellon University, Pittsburgh, PA, USA

[2] Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA, USA

[3] Department of Psychology, Vanderbilt University, Nashville, TN, USA

In Press at

Behavior Research Methods

Correspondence to:
Ana E. Van Gulick
Hunt Library, Carnegie Mellon University
4909 Frew Street
Pittsburgh, PA 15213
Email: anavangulick@cmu.edu
Phone: 315-569-4539

Abstract

How much do people differ in their ability to recognize objects, and what is the source of these differences? To address the first question, psychologists created visual learning tests like the Cambridge Face Memory Test (Duchaine & Nakayama, 2006) and the Vanderbilt Expertise Test (VET; McGugin et al., 2012). The second question requires consideration of the influence of both innate potential and experience, but experience is difficult to measure. One solution is to measure the products of experience beyond perceptual knowledge, specifically non-visual semantic knowledge. For instance, the relation between semantic and perceptual knowledge can help clarify the nature of object recognition deficits in brain damaged patients (Barton et al., 2009). We present a reliable measure of non-perceptual knowledge in a format applicable across categories. The Semantic Vanderbilt Expertise Test (SVET) measures knowledge of relevant category-specific nomenclature. We present SVETs for eight categories: cars, planes, Transformers, dinosaurs, shoes, birds, leaves, and mushrooms. The SVET demonstrates good reliability and domain-specific validity. We find partial support for the idea that the only source of domain-specific shared variance between the VET and SVET is experience with a category. We also demonstrate the utility of the SVET-Bird in experts. The SVET can facilitate the study of individual differences in visual recognition.

Keywords: Semantic knowledge, visual object recognition, measurement, experience, individual differences

Introduction

Individual differences have been almost completely overlooked in the study of general object recognition in neuro-typical populations. One exception is faces, due in part to the development of a standardized test of face memory with established reliability and validity: the Cambridge Face Memory Test (CFMT; Brad Duchaine & Nakayama, 2006). The CFMT reveals variability in face recognition performance across the spectrum of performance (Bowles et al., 2009; Germine, Duchaine, & Nakayama, 2011; Hedley, Brewer, & Young, 2011; Russell, Duchaine, & Nakayama, 2009). Extending this line of research to objects, McGugin et al. (2012) developed a battery of tests exploring individual differences in object recognition for a variety of non-face object categories: the Vanderbilt Expertise Test (VET). Most people have a lot of experience recognizing faces, but they vary much more in their experience individuating objects in other categories. Thus, individual differences in performance with objects are likely to reflect both domain-general ability and domain-specific experience. Like the CFMT, the VET is a visual learning test. It includes a variety of non-face object categories (in VET 1.0: birds, cars, planes, owls, wading birds, motorcycles, mushrooms, and leaves), and each subtest has good internal consistency (McGugin, Richler, Herzmann, Speegle, & Gauthier, 2012).

While other object recognition tests focused on a single category (e.g., Dennett et al., 2012), the inclusion of several categories in the VET battery makes it possible to isolate domain-specific skills, allowing researchers to partial out domain-general abilities (McGugin, Newton, Gore, & Gauthier, 2014a; McGugin, Van Gulick, Tamber-Rosenau, Ross, & Gauthier, 2014b). However, a complication is that observers may possess more non-perceptual knowledge for some categories (e.g., names for some of the objects), and this knowledge may contribute to performance. This problem is hard to avoid with familiar objects in a domain of expertise, so

here we seek to *measure* this knowledge rather than control it. We designed a battery of tests for measuring this sort of information: the Semantic Vanderbilt Expertise Test (SVET). The development of parallel VET and SVET batteries for the same domains will allow us to address several important issues.

One of these issues has to do with the validity of domain-specific object recognition measures. Prior work with the Vanderbilt Expertise Test (McGugin et al., 2012) argued that measuring performance with one object category was not sufficient to get a representative measure of object recognition ability. Because experience can influence performance, a domain-general ability is best understood as a latent construct that contributes to performance for many categories. But the VET can also be used to measure skill in a single domain, as in fMRI studies of expertise for cars (e.g. McGugin et al., 2014a; 2014b). How do we validate the domain-specific component of a test for cars, birds, or mushrooms? What sort of evidence can suggest that are we tapping into a coherent domain of object knowledge, distinct from other categories? Here, using variance on the VET and SVET for the same categories, we can estimate whether there is category-specific shared variance between these two different types of measures, over and beyond what is shared across categories. This can provide convergent (within category, across task type) and divergent (within task type, across category) validity for both the VET and the SVET.

A related issue is what this shared domain-specific variance across categories may represent. We postulate that the reason for shared variance between visual and semantic performance for a category (say cars), over and beyond differences shared with measurements for other categories (say birds or planes), is experience with that domain. A measure derived from performance to estimate experience eliminates problems associated with self-report

measures (Zell & Krizan, 2014). One drawback of self-reports is that the same question about experience can be answered on the basis of different dimensions depending on the specific category (Gauthier et al., 2014). Here we will explore the link between these self-reports of experience and our estimates of experience derived from performance.

Finally, we wished to provide a non-visual semantic test in a common format for many object categories – a tool for research that does not currently exist. Prior tests of semantic category knowledge present with some limitations. Some tests require visual knowledge as well as semantic knowledge (e.g., naming images; Barton, Hanif, & Ashraf, 2009). Others (e.g., Tanaka & Taylor, 1991) use subordinate-level object naming or feature listing, which is difficult for novices who simply may not have enough knowledge to list distinguishing features for many objects. Lastly, the structure of these tests is often specific to a single domain (e.g., name the manufacturer of a car model; Barton et al., 2009), making them hard to use across categories.

The SVET battery seeks to provide a more versatile estimate of non-perceptual knowledge that can be used across a variety of object categories. The test was designed with the following attributes in mind: 1) it would measure semantic knowledge in a manner that is theoretically independent of visual knowledge; 2) it could be easily implemented for a wide range of object categories; 3) it should discriminate across levels of performance over the entire continuum of experience with a category. While for some applications one may desire a test to discriminate among experts in a particular area, here we also sought good discrimination in a larger range, from those with little experience up to individuals with a moderate to high amount of knowledge. Unlike most knowledge tests that are designed with experts in mind, we started with the goal of obtaining information about people at the low end of the continuum. It should be easier to add obscure nomenclature to increase the difficulty of tests for experts, but there is

always a concern that below a certain minimal level of expertise, poor performers simply get lumped into an undifferentiated "novice" group. Research in some domains of expertise (e.g., chess) simply do not have reliable measures below a certain skill level (Bilalic, Langner, Erb, & Grodd, 2010; Bilalic, Langner, Ulrich, & Grodd, 2011). In keeping with these goals, we mainly focused on testing subjects from the "normal" population (an undergraduate subject pool and a sample of workers from Amazon Mechanical Turk with US IP addresses), but in Experiment 2 we also test an expert population for one category. An inherent limitation of measuring domain-specific knowledge (which applies to the VET and SVET equally) is that the specific categories and items need to be relevant to the population studied (e.g., bird species and car models vary in different parts of the world). As is the case for face recognition tests, researchers working in other populations may need to create new measures but should be able to use the same general formats.

To achieve the above-mentioned goals, the SVET subtests measure knowledge of subordinate-level object labels and names for a given category. The decision to use knowledge of domain-relevant nomenclature as a proxy for semantic knowledge was motivated by the desire to find a format that would apply to a wide variety of categories for which other aspects of semantic knowledge may have very different structures (e.g., for living and non-living objects). While most semantic networks will include more than names, we assume that for most categories, the size of the semantic network is directly related to the size of the relevant vocabulary. Some models of object naming postulate that naming is a stage subsequent to semantic access (e.g., Bruce & Young, 1986). Likewise, the interactive activation and competition model of object naming (Humphreys, Lamote, & Lloyd-Jones, 1995) links name representations directly to semantic representations. These models, and common sense, suggest that a reasonable

assumption is that in most domains, people who have the richest semantic knowledge also know more names for objects in that domain. We do not have people name images of objects, as this would tap into both perceptual and non-perceptual representations. Instead, the SVET structure is similar to knowledge tests created to measure individual differences in print exposure: the Author Recognition Test (ART) and the Magazine Recognition Test (MRT) (Stanovich & Cunningham, 1992). In these tests, subjects must select the names of real authors (mostly novelists) and magazine titles among foils. The SVET uses a similar format of recognizing target names among foils, but employs a triplet (one target, two foils) trial structure to reduce guessing and allow for more precise design of easy and difficult trials based on foil selection. Stanovich and Cunningham (1992) also provide an example of using a theoretical framework to relate domain-specific and domain-general performance with domain experience to make inferences about why people differ, which is a framework we ultimately hope to use with the VET and SVET.

In Study 1, we describe how we created and refined the SVET for eight common object categories by testing iterations of each test in a large online sample from Amazon Mechanical Turk (AMT). We used classical item analyses, factor analysis, and measures from Item Response Theory (IRT) modeling. We measured the reliability (internal consistency) and dimensionality (does the test measure a single body of knowledge?) of each category-specific SVET. We also validated the SVET with self-reports of domain experience to determine if the relation between performance and experience is domain-specific.

In Study 2 we then tested the SVET together with the VET for each category in a large laboratory sample. In addition, we measured face recognition ability (CFMT), which has been related to the VET in prior work (Gauthier et al., 2014; McGugin et al., 2012), and fluid

intelligence (*Gf*), which we postulated could predict the ability to acquire nomenclature. We also used a set of standardized questions (from Gauthier et al., 2014) relevant to experience with each category, to estimate domain-specific experience from self-report. In this dataset, we again looked at the reliability of each VET and SVET. We examined test validity in two ways: 1) is performance domain-specific (greater relation between VET and SVET or self-report experience for the same domain relative to other domains), and 2) does SVET performance separate novices from experts. Lastly, we tested whether the shared variance between VET and SVET can be understood as domain-specific experience.

In Study 3, we demonstrated the utility of the SVET in an expert population. We measured performance on the VET-Bird and SVET-Bird in a group of birders, asking if these measures demonstrate validity relative to self-reports of birding experience, and if the tests are sensitive enough to capture individual differences in visual and semantic knowledge among experts.

Study 1

The first stage of SVET development was creating an initial set of items and testing each set on Mechanical Turk (MTurk) (https://www.mturk.com/, Amazon Web Services, Amazon.com Inc., Seattle, WA; see Crump, McDonnell, & Gureckis, 2013; Paolacci & Chandler, 2014). We went through multiple iterations of data collection, item analysis and revision for each test (between 1 and 4 revisions per category). We also included three catch trials that were extremely easy in each test, to ensure that subjects were following instructions. Subjects who missed more than one catch trial were excluded from all analyses, which was rare (0-4% of subjects per test).

Below we focus the results on the final version of each SVET, for 116 MTurk subjects who completed the tests for all eight categories. In Study 2, we will present results for these same versions of the SVET used in testing 213 subjects in the laboratory, in conjunction with a modified version of the VET battery of visual learning tests (McGugin et al., 2012), adapted to test the same categories as the SVET.

SVET Design

Categories. We selected a set of object categories with available subordinate-level names, including domains of typically greater interest to men or women, based on prior VET results (McGugin et al., 2012) and intuition for new categories. We also wanted a mix of living things and artifacts. We created SVETs for nine categories and, based on pilot testing, abandoned one of them (butterflies) as we could not produce enough items that were known to novices. Our final set consisted of: cars, planes, Transformers, dinosaurs, women's shoes, birds, leaves, and mushrooms.

Trials. The SVET for each category consisted of 48 test trials and three catch trials. Each SVET test trial presented three names: one target name, a real subordinate-level name of an object in that category (e.g. Honda Civic, 737, blue jay, birch), and two foil names, which were either names from a different category not tested in any of the SVETs (e.g., types of rocks, grass, or viruses) or were entirely made-up but plausible words for the given category. Compared to object naming, this trial format allows people with limited category knowledge to provide a response on each trial. Catch trials were very easy trials that followed the same format as the test trials with real target names but much more obvious foil names (e.g., blue jay, JCPenney, lipstick). Catch trials were of particular interest for online testing, to exclude any subjects who did not read or understand the instructions or who lacked motivation. The order of trials and

tasks were the same for all subjects to prevent variance from order effects to confound individual differences. In the SVET 1.0 trials were ordered from easiest to hardest based on trial accuracy from earlier data. The SVET (and VET) battery is meant to be modular so that selection of sub-tests for specific domains can be adjusted based on a research question. Table 1 gives examples of an easy, medium, and difficult SVET trial for each category (see Supplemental Information for complete tests and information about the target items in each category).

Table 1. Example trials from each SVET. Selected trials illustrate an easy, medium, and difficult trial for each category (lower trial numbers are easier). Names in bold are the real name and correct response. See full SVETs in Supplemental Materials.

	Name 1	Name 2	Name 3
CAR			
Trial 6	Volvo Focus	Mercedes-Benz C300	**Mercury Alero**
Trial 26	Suzuki Prestige	**Infiniti G37**	Pontiac S550
Trial 49	Saturn Fusion	**Acura TSX**	Saab S80
PLANE			
Trial 2	**737**	Serpens	Sheffield
Trial 24	8900	A2 Lobo	**Spitfire**
Trial 45	**Mosquito**	Western Lair	A480
TRANSFORMER			
Trial 4	Lavaman	Chromoburn	**Quickstrike**
Trial 28	**Sunstreaker**	Septawave	Proton
Trial 46	Waveracer	**Hound**	Sotter
DINOSAUR			
Trial 7	**Pentaceratops**	Eudontidectes	Microtarius
Trial 22	Stuthioceratops	Centaurisaurus	**Iguanodon**
Trial 48	Corposaurus	Monocyclosaurus	**Mussaurus**
SHOE			
Trial 4	**Nine West**	Rebecca Fox	Aloft
Trial 25	Zetta	Kalden White	**Franco Sarto**
Trial 47	Graham Wood	Gravelle	**Chinese Laundry**
BIRD			
Trial 5	**Savannah Sparrow**	Tufted Gemthroat	Green Huckaloo
Trial 23	**Scarlet Tanager**	Blue-stripe Binbeak	Tri-colored Wheatear
Trial 37	Spot-breasted Pixie	**McCown's Longspur**	Pale-eyed Baylin
LEAF			
Trial 3	Red Mountainwood	Venuswood	**American Sycamore**
Trial 19	**Yellow Poplar**	California Bargo	Feather Willow
Trial 47	Silver Aster	Valley Walnut	**Tulip Poplar**
MUSHROOM			
Trial 4	**White Truffle**	Milky Scaber	Sugar Siullus
Trial 25	Amber Stalk	Tavel	**Enoki**
Trial 46	**Crab Brittlegill**	Elephant Trunk	Glass Cap

Trial presentation. On each trial, three object names (one real name and two foils) were presented in black type on a white background in the center of the screen, in a single row so that the middle name was centered with the other two names to the left and right. The location of the correct response occurred with equal frequency in the left, right, and center locations. Subjects responded by clicking on a name with no time limit. No feedback was provided and subjects could not return to previous trials.

Individual Test Piloting and Revisions

After designing an initial SVET for each category, data were collected independently for each SVET with subjects recruited from MTurk with restrictions that they were English speakers residing in the United States (N=35-101 per test iteration; N=1,383 total).

For each version, we conducted informal item analyses on the data. We computed Cronbach's alpha, a measure of internal consistency, to assess the reliability of the measurements. Our goal was a minimum reliability of $\alpha > 0.7$, and we especially monitored for large changes in reliability across versions, although reliability typically remained stable or improved with versions. As a minimal check for validity, we considered correlations between SVET performance and subjects' self-report of experience with that category (which were always positive and generally increasing with iterations). Using exploratory factor analyses based on polychoric correlations, we considered the dimensionality of each test – our goal was to make each test as unidimensional as possible so that test scores would reflect differences in primarily one type of knowledge across individuals. Therefore, if a test had two large factors, items for one of them were dropped and more items similar in nature to the items that loaded on the other factor were created. In other cases, one main factor was dominant and we replaced items that did not load well on that factor. It was not always clear what explained the various factors on a test.

We changed or modified items with the goal to obtain a relatively even distribution of difficulty with as wide a range as possible, and to have foils chosen with similar probability. A foil that is rarely chosen can increase the chance level, if even those subjects with little knowledge never choose it. We also conducted exploratory analyses using 2-parameter Item Response Theory (IRT) models for each test (see Embretson & Reise (2000), for an introduction to IRT). Such models produce item characteristic curves for each item, based on a difficulty parameter and a discrimination parameter. We used these results to try to spread item difficulty through the range of subject ability, seeking items that would cover the entire range of subject ability. We selected items to maximize discriminability, to provide as much information as possible to separate observers who fall above and below the item's difficulty level. Any item with negative discriminability (people with a higher latent trait perform worse than those with a lower latent trait) was replaced. In a typical test revision, between 5 and 20 target or foil names were changed, but an entire trial rarely had to be dropped.

Testing All SVETs in a Single MTurk Sample

After revising each SVET separately, we conducted an online study in which all subjects completed the final SVET composed of all eight categories. This is the first test for the entire battery in the same subjects.

Subjects. Subjects were recruited online on MTurk. The SVET-Car was posted first and subjects who completed the test without missing more than one catch trial and who obtained above chance accuracy (all but two) were invited to complete the other seven SVETs. The study was approved by the Vanderbilt IRB and subjects gave informed consent before the start of each test. Subjects were paid $0.75 for the SVET-Car and $0.10 for each of the other tests, but were awarded a $1.00 bonus for each of the seven additional tests if they completed the full set within

24 hours ($8.45 total). All eight SVETs were completed by 116 subjects (48 male) aged 18-67 (mean=35.82, SD=12.45). All but one of these subjects reported English as their native language, but all reported being English-speakers residing in the United States. Subjects who completed only a subset of the SVETs (N=9) or completed only the SVET-Car but did not accept the invitation to complete more tests (N=22) were compensated without bonuses and are excluded from the analyses.

Procedure. At the beginning of each SVET, we asked subjects to provide a rating of their experience with the object category (as in Gauthier et al., 2014; McGugin et al., 2012). For example, the self-report for cars asked, "Rate your expertise with: *Cars*. By expertise we mean your experience with, interest in, and knowledge about items in this category, relative to other people." Ratings were on a whole-number scale from 1 (very much below average) to 9 (very much above average). Task instructions to select the *real* name were adjusted for each category to be as specific as possible (e.g. for birds, "the real, common name of a bird species found in North America"). Each SVET trial was then presented as three names printed across the screen until subjects responded by clicking on one of the three names.

Results and discussion. No subjects were excluded due to poor catch trial performance (restriction criterion for catch trial performance across all tests was accuracy <0.66), which was high across tests (0.93-0.99 for each test). Mean SVET performance and variability differed between categories, with cars and mushrooms showing the highest and lowest accuracy, respectively (Table 2, column 1). The greatest variability in accuracy was observed for cars and shoes, for which self-report of experience was the highest (Table 2, column 2), suggesting that greater amounts of experience with cars and shoes is reflected in the SVET score.

For all categories, we observed a significant correlation between SVET performance and self-report of experience with that category (Table 3, column 1). SVET performance and experience within the same domain were always more highly correlated within category (mean r=0.37) than across categories (mean r=0.05). In a further step, we asked how each SVET predicts the average of experience ratings for the other seven categories, referred to as experience-Other (mean r=0.08; Table 3, column 2), and also how category experience predicts the average of SVET performance for the other seven categories, referred to as SVET-Other (mean r=0.07; Table 3, column 3). As noted in Table 3, the within category correlation was nearly always significantly stronger than any between category correlation for SVET and experience (with the single exception of leaves for Experience and SVET-Other).

Table 2. Results of the SVET 1.0 for each category from a single group of subjects on MTurk (N=116) in Study 1.

	Mean Acc (Std Dev)	**Experience (Std Dev)**	**Cronbach's Alpha**
Car	0.69 (0.15)	4.78(1.37)	0.94
Plane	0.53 (0.12)	2.78(1.38)	0.84
Transformer	0.45 (0.11)	2.84(1.65)	0.75
Dinosaur	0.47 (0.08)	3.90(1.48)	0.63
Shoe	0.53 (0.17)	4.07(1.66)	0.92
Bird	0.52 (0.10)	3.41(1.40)	0.75
Leaf	0.57 (0.13)	3.57(1.53)	0.87
Mushroom	0.44 (0.10)	2.63(1.40)	0.71

Table 3. Correlations (*r*) of SVET 1.0 and experience self-reports within and between categories for a single group of subjects on MTurk (N=116) in Study 1. Column 1 shows the correlation between SVET and experience for the same category. Column 2 shows the correlation between SVET for the category and experience-Other (average of self-reports on other 7 categories). Column 3 shows the correlation between experience for the category and SVET-Other (average of SVET performance on other 7 categories). Values in bold are statistically significant ($r_{Crit}(114)=.18$, $p<.05$). The contrast of interest tests if within category relations are stronger than across category relations. In Column 1, this is denoted with a ^ if Column 1 (within) is greater than Column 2 (across category experience) and with a * if Column 1 (within) is greater than Column 3 (across category SVET) as tested using a one-tail Steiger's Z ($p<0.05$).

	SVET and Experience (within category)	**SVET and Experience for other categories**	**Experience and SVET for other categories**
Car	**0.37^***	0.06	0.13
Plane	**0.36^***	0.15	0.06
Transformer	**0.50^***	0.06	0.09
Dinosaur	**0.40^***	**0.21**	**0.19**
Shoe	**0.44^***	-0.09	**0.24**
Bird	**0.33^***	0.14	0.04
Leaf	**0.27^**	0.01	-0.15
Mushroom	**0.27^***	0.08	-0.03

Summary of online testing. The SVET 1.0 battery produced acceptable reliability for most subtests (dinosaurs was lowest at 0.63- see Table 2, column 3). The tests were generally difficult (with performance typically around or below 50%, chance being 33%). This is consistent with the fact that we did not recruit subjects with special knowledge in any of these categories and the mean report of experience for all categories was below 5 (average). Speaking to the validity of the subtests as domain-specific, the correlation between a given SVET and self-report within the same domain was higher than the correlation between that SVET and experience for all other categories, or the correlation between experience for that category and all other SVETs. Study 2 was performed with a larger sample and with the equivalent visual tests (VETs), and so we proceed to that dataset for more extensive analyses (including gender effects). The online dataset is available from the first author.

Study 2

Study 2 Design

Five measures were used in Study 2, three domain-specific measures including an experience questionnaire (Gauthier et al., 2014), the SVET 1.0 as described in Study 1, and subtests from the VET (McGugin et al., 2012) matching categories used in the SVET. We also included two domain-general tests of visual learning and 3 tasks tapping into fluid intelligence (*Gf* - adapted from Redick et al., 2012; see also Hambrick, Meinz, & Oswald, 2007; Hambrick, Pink, Meinz, Pettibone, & Oswald, 2008). Each domain-specific measure tested eight object categories in the following order: cars, birds, dinosaurs, shoes, planes, mushrooms, Transformers, and leaves. This order was selected in an effort to alternate typically male- and female-interest categories.

Face recognition performance was measured using the Cambridge Face Memory Test (CFMT; Duchaine & Nakayama, 2006).

All subjects received the same order of tasks and trials (one exception was the experience questionnaire that had two orders). The experience questionnaire was given first, followed by the CFMT and the VETs (which do not include any object names), then the SVETs, and lastly the fluid intelligence tests[1].

Because this work concerns a novel test, our sample size was determined to be sufficiently large to detect effect sizes of r=0.20 at more then 80% power (Study 1). In the tradition of individual differences work, we do not correct for multiple comparisons but invite readers to focus on effect sizes (r-values) rather then significance per se, especially in using the results to predict further work with the SVET.

[1] After the VETs but before the SVETs subjects also completed a short bird and shoe image naming test that will be described and reported in Study 3.

Methods

Subjects. 217 subjects were recruited from the Vanderbilt University and Nashville communities; they gave informed consent and received course credit or monetary compensation for their participation. The study was approved by the Vanderbilt IRB. All subjects reported normal or corrected to normal visual acuity, were native English-speakers, and had lived in the United States at least 10 years. Subjects were not specifically recruited for their interest in any of the tested categories; most subjects were students at the University. One subject was excluded for not completing all of the tasks and three subjects were excluded for below chance (0.33) performance on two or more SVETs. Data are reported for 213 subjects (86 male) aged 18-55 (mean=22.49, SD=6.31).

Equipment. The experiment was conducted in the laboratory on Apple Mac Minis with LCD monitors. The experience questionnaire was completed using REDCap electronic data capture survey tools (http://redcap.vanderbilt.edu; supported by UL1 TR000445 from NCATS/NIH; Harris et al., 2009) hosted by Vanderbilt University. Subjects were allowed to sit a comfortable distance from the monitor (approximately 40cm).

Tasks.

Experience Questionnaire (modified from Gauthier et al., 2014). The questionnaire measures domain-general and domain-specific self-reported experience along several dimensions. First, subjects answer four domain-general questions, followed by seven domain-specific questions (Table 4). They answered the same question for each of the eight categories before moving to the next question. Even-numbered subjects completed the domain-specific questions in reverse order from odd-numbered subjects to minimize potential order effects. All questions

were answered on a scale from 1 to 9 described for each questions (1 = very little, 9 = a lot) except for duration experience, which was answered from 1 (no interest) to 5 (6 or more years).

CFMT (Duchaine & Nakayama, 2006). Stimuli and procedure were the same as used by Duchaine and Nakayama (2006) in the Cambridge Face Memory Test (CFMT). Stimuli were grey-scale images of faces from varying viewpoints with and without added noise. Subjects study six target Caucasian male faces and then have to select those target faces among two distractors on each trial, despite variations in the target at test relative to study (new orientation, different lighting, or added noise). Following introductory trials, in all subsequent trials, subjects try to select the target face between two foils. There are three blocks of trials shown sequentially, with an opportunity to study the six target faces together on the screen again before each block. The first eighteen test items (6 target faces × 3 presentations) show the faces in a view that is identical to that which was studied in the introduction (Block 1); the next 30 items (6 target faces × 5 presentations) use novel views (Block 2); and the last 24 items (6 target faces × 4 presentations) use novel views with the addition of Gaussian noise to keep performance off ceiling (Block 3). Subjects answered by pressing 1, 2, or 3 on the number pad. Images remained on the screen until subjects made a response.

Table 4. Questionnaire of object experience, generally and with specific domains. Subjects provided self-reports of experience in response to each question on a scale from 1-9.

General Experience
1. Generally speaking, how strong is your interest in classifying objects in their various sub-categories (such as learning about different kinds of insects, plants, vehicles, tools...)?
2. Generally speaking, how easily do you learn to recognize objects visually?
3. Generally speaking, relative to the average person, how much of your time at WORK or SCHOOL involves recognizing things visually?
4. Generally speaking, relative to the average person, how much of your FREE TIME involves recognizing things visually?

Domain-specific Experience
Note the following order is for odd subject numbers, even subject numbers received the reverse order; XXX = category (e.g. birds, cars)
1. Please rate yourself on your expertise with XXX considering your interest in, years of exposure to, knowledge of, and familiarity with XXX.
2. If you are interested in XXX, when did this interest begin?
3. How often do you look at IMAGES of XXX, in movies, television, or other kinds of documents (books, magazines, or online)?
4. How often do you read TEXT (in books, magazines, online) that contains information about XXX?
5. How important is the domain of XXX to you, relative to all the other things you are interested in?
6. If you saw a specific XXX in a TV show, how sure are you that you could recognize that item among similar images if you were tested the next day?
7. If you were asked to write an essay about different kinds of XXX, how extensive and detailed do you think your essay would be?

VET 2.0 (modified from McGugin et al., 2012). Stimuli were all grey-scale images of objects with varied backgrounds (see Figure 1). Some of the subtests used here (dinosaur, Transformer, shoe, and passerine bird) were not included in the original set (McGugin et al., 2012) but were created in the same way to be paired with the SVET. Other subtests (car, plane, leaf and mushroom) were revised (images and trials altered) from the original version to improve coverage of the range of performance and reduce dimensionality (according to the same informal process as described previously for the SVET; see supplemental materials for additional information on the creation of the VET 2.0). Target objects did not overlap between VET and SVET. That is, if an object (e.g. Audi A4) was one of the 6 target objects in the VET, the name of that object was not used as a target in the SVET. Objects that occurred in the VET only as a foil could be used as a SVET target.

The VET for each of the eight categories began with the presentation of a study screen showing an example of each of the six target objects (see Figure 1). Subjects studied these

objects for as long as they chose. No subordinate level labels were ever provided (a departure from the original VET in which labels were provided for the six targets. The images were vertically centered on the screen, to the left of center, at center, and to the right of center. Subjects selected the target image by pressing 1, 2, and 3 on the keyboard, corresponding to the leftmost, center, and rightmost images, respectively. In the first 12 trials, the studied image of one target appeared with two foils, and feedback was provided (correct/incorrect). The study screen with the six target objects was shown again for unlimited study time. Subjects then completed 36 trials with targets that were different examples of the same identities. Subjects were instructed that the target could now be a different image and to generalize across viewpoint, background, color, or size, depending on the category. Subjects did not receive feedback on these trials.

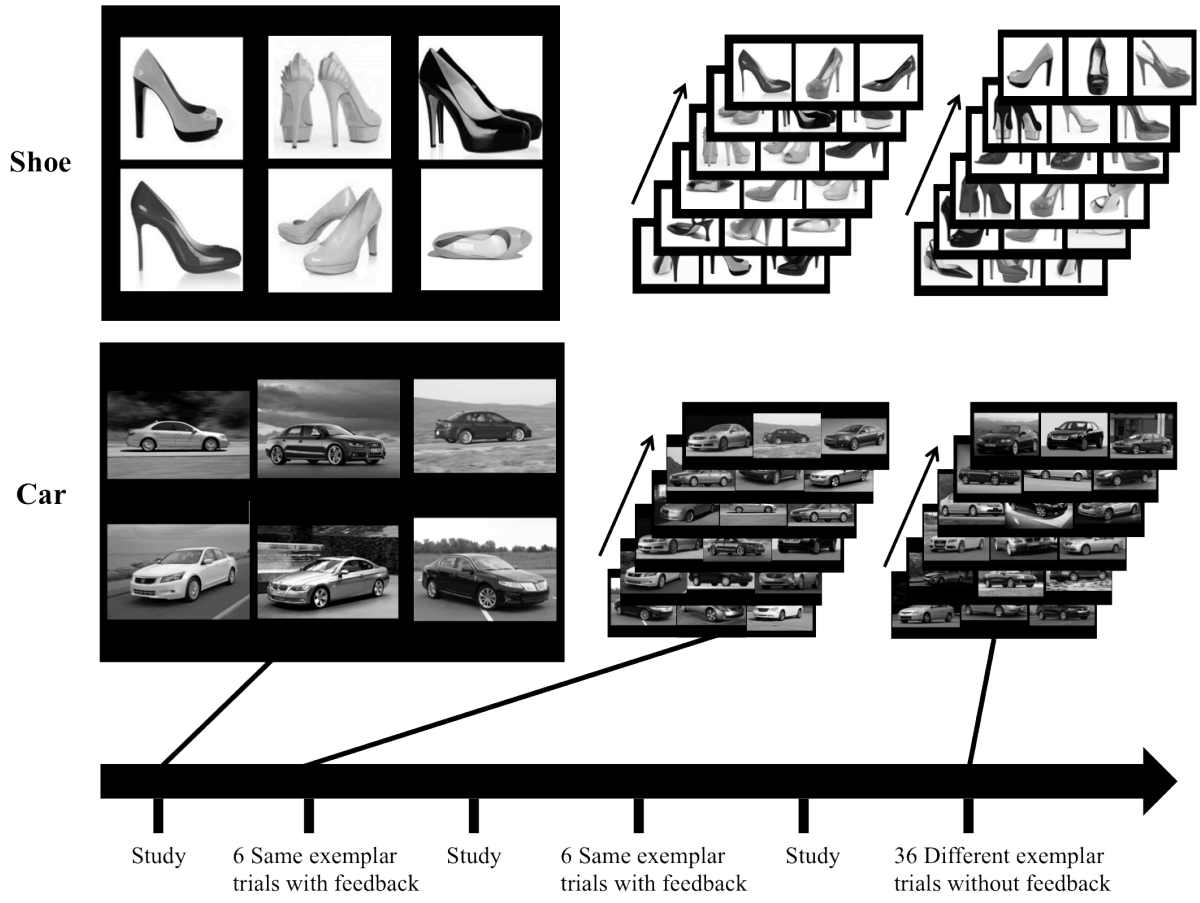

Figure 1. Stimuli and test design for VET 2.0 with women's shoes and cars.

SVET. The stimuli and procedure of the SVET was the same as previously described in Study 1 except that subjects indicated their response by pressing 1, 2 or 3 on the keyboard number pad.

Fluid intelligence. The stimuli and tasks (Figure 2) used in these tests were adapted from Redick et al. (2012) (see also Hambrick et al., 2007; 2008). For each test, subjects had a time limit in which to complete as many trials as possible and were told to focus on accuracy rather than speed in completing the tests. There was no time limit on any specific trial, but a response was necessary to advance to the next trial and subjects could not go back to prior trials. In each task, trials got progressively more difficult. Subjects responded using the keyboard number pad.

Practice trials (2-5 per task) were included at the start of each task with a short explanation of the correct response for each problem. No feedback was provided on any of the test trials.

Raven's advanced progressive matrices (RAPM) (Raven, Raven, & Court, 1998). Stimuli were matrices with one missing piece. Each matrix was a 3 x 3 array of objects with the lower right-hand object missing. The features of the objects in the matrix varied systematically (e.g., object shape, number of lines, direction of lines) according to a pattern. Subjects had to select which of eight objects would appropriately complete the matrix. Eight options were presented below the matrix, labeled 1 to 8. A subset of 18 trials selected from the full advanced matrices test (Raven et al., 1998) was used here. Subjects had 10 minutes to complete as many trials as possible up to 18.

Letter sets (Ekstrom, French, Harman, & Dermen, 1976). Stimuli in each trial were five sets of four letters (e.g. BCCB, GFFG, LMML, QRRQ, WXXW). Four of the five letter sets followed a specific rule and subjects had to select which of the five sets did not follow the same composition rule as the others. The sets were displayed in a row in the center of the screen and labeled with the numbers 1–5. These rules were based on features such as alphabetical order, repetition, or the presence/absence of a specific letter. Subjects had 7 minutes to complete as many trials as possible, up to 30.

Number series (Thurstone, 1938). The stimulus in each trial was an array of 5–12 one or two-digit numbers, selected and arranged so that when read from left to right they followed a particular rule. Subjects had to select which of five (labeled 1-5) number answers would continue the pattern. The pattern governing the number array could apply to single numbers alone or groupings of numbers in the series and could follow either a numerical order (e.g. 1 2 4 1 2 5 1 2

6 interpreted as 124, 125, 126) or a mathematic function (e.g. 2 5 8 11 14 17 interpreted as +3 to each number). Subjects were given 5 minutes to complete as many trials as possible, up to 15.

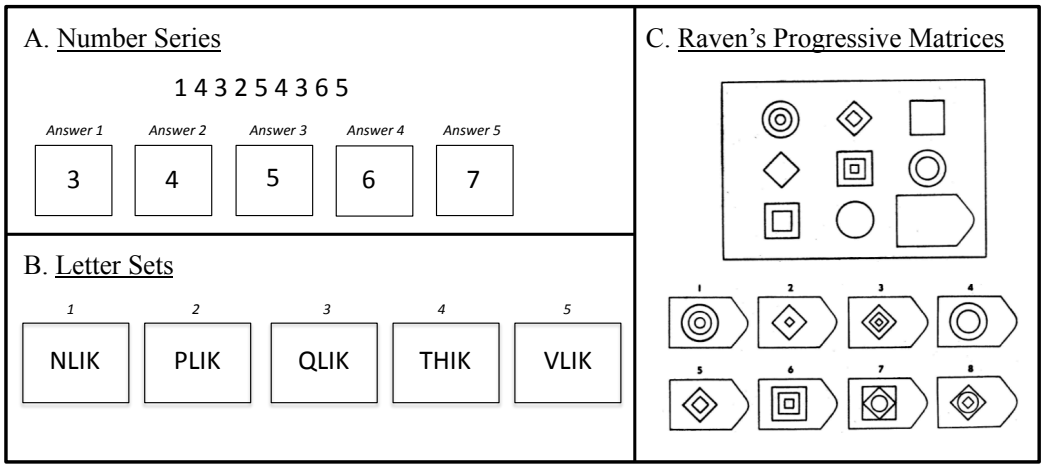

Figure 2. Stimuli and trials for fluid intelligence tests. Examples of a single trial for each task. Stimuli adapted from Redick et al. (2012).

Results and Discussion

Accuracy, variability and reliability for each measure.

Visual and semantic tests. Accuracy on all of the VETs and SVETs was above chance (.33) and not at ceiling (Figure 3 and Table 5). Accuracy distributions of VET and CFMT scores were generally negatively skewed whereas SVET scores were positively skewed, suggesting that the SVETs were more difficult.

Table 5. Accuracy on CFMT, VETs and SVETs in Study 2. Columns show the mean sum score, 95% confidence interval (CI), median, interquartile range (IQR), and skewness.

		Mean	95% CI	Median	IQR	Skewness
	CFMT	0.80	(.78, .82)	0.81	0.17	-0.42
VET	Car	0.59	(.57, .61)	0.60	0.21	0.06
	Plane	0.68	(.66, .70)	0.69	0.21	-0.01
	Transformer	0.72	(.70, .74)	0.73	0.19	-0.19
	Dinosaur	0.70	(.69, .71)	0.69	0.17	-0.28
	Shoe	0.73	(.71, .75)	0.75	0.17	-0.42
	Bird	0.67	(.65, .69)	0.69	0.19	-0.41
	Leaf	0.57	(.55, .59)	0.58	0.15	-0.27
	Mushroom	0.60	(.59, .61)	0.60	0.13	-0.22
SVET	Car	0.62	(.60, .64)	0.60	0.19	0.21
	Plane	0.45	(.44, .46)	0.44	0.13	0.32
	Transformer	0.42	(.41, .43)	0.42	0.13	0.54
	Dinosaur	0.47	(.46, .48)	0.46	0.08	0.80
	Shoe	0.54	(.52, .56)	0.50	0.23	0.43
	Bird	0.47	(.46, .48)	0.46	0.10	0.48
	Leaf	0.53	(.51, .55)	0.52	0.17	0.24
	Mushroom	0.40	(.39, .41)	0.40	0.13	0.30

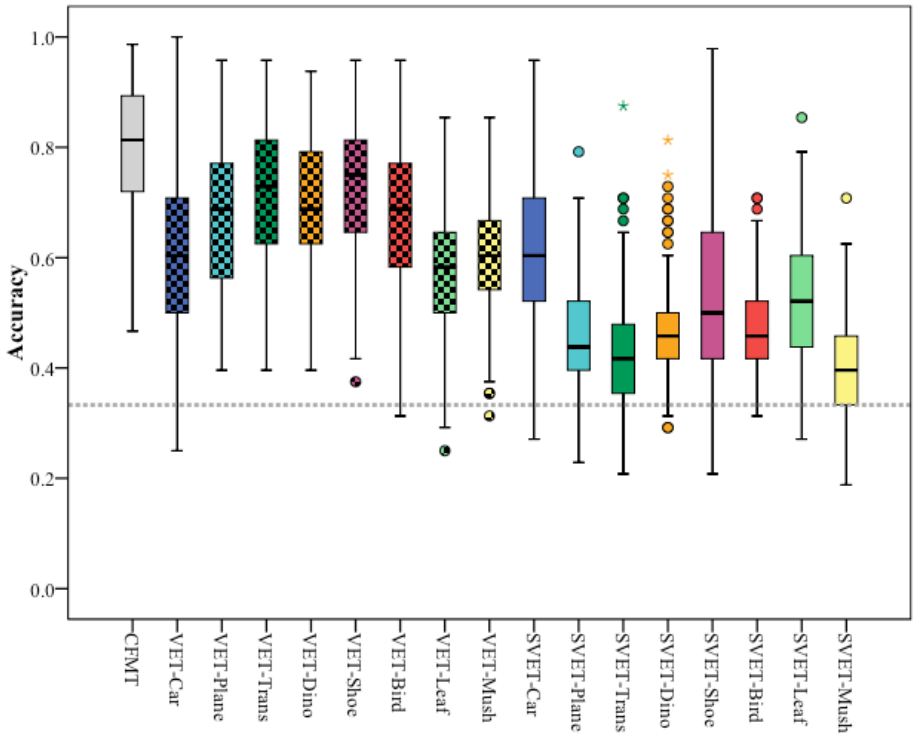

Figure 3. Boxplots of accuracy performance in Study 2 on CFMT (grey), VETs (checkered), and SVETs (solid). The bottom and the top of each box represent the first and third quartile, respectively and the middle line of each box shows the second quartile, or the median. Whiskers show the highest and lowest scores between 1.5 and 3 times the interquartile range while outliers beyond this range are represented by dots (or an asterisk in the case of very extreme outliers). The grey, dotted line shows chance (.33).

The visual measures, CFMT and VETs, showed good reliability measured as internal consistency (Cronbach's alpha in Table 9). In general, the reliability of the SVETs was also good, with the exception of the SVET-Bird (α = .52) and SVET-Mushroom (α = .66), which were less reliable. Exploratory factor analyses with the maximum number of factors and polychoric correlation for each of the SVETs in Study 2 show that SVET-Bird, -Mushroom, and –Dinosaur exhibited the most multidimensionality, which may also translate to lower test reliability. Multidimensionality suggests that a test measures more then a single source of individual differences – in revisions of the SVET, this would represent a target for improvement.

Experience self-report questionnaire. Cronbach's alpha for the four domain-general questions regarding experience individuating objects was 0.65, suggesting a common construct. Self-reports of category experience for each category also demonstrated high reliability across questions (Cronbach's alpha for each category ranged from 0.83-0.93, mean α =0.88).

Table 6. Correlations (*r*) between the experience aggregate scores for general object experience and each domain with age and sex in Study 2. Values shown in bolded red are statistically significant ($r_{Crit}(211)=.132$, $p<.05$).

Experience Aggregate	Correlation with:	
	Age	Sex
General	0.02	0.02
Car	0.03	**0.27**
Plane	-0.04	**0.33**
Transformer	-0.09	**0.30**
Dinosaur	0.09	**0.28**
Shoe	-0.10	**-0.72**
Bird	**0.16**	0.05
Leaf	**0.14**	0.07
Mushroom	0.08	0.11

We computed nine experience aggregate scores for use in later analyses: the average of the four general experience ratings, and category-specific experience aggregates for each of the eight categories (average of the seven category-specific experience questions). Table 6 shows the

correlations of reported experience in each domain with age and sex; we observed the largest sex effects in experience for all categories for which we expected a male interest, but women only reported more interest then men with women's shoes.

Fluid intelligence. Performance on the three measures of fluid intelligence was calculated as the number of trials correctly answered within the time limit (Number series: mean=10.27, SD=2.75; Letter sets: mean=18.08, SD=4.12; RAPM: mean=10.85, SD=2.91). Each of the tests demonstrated high reliability calculated with Cronbach's alpha (Number series: α=0.85; Letter sets: α=0.87; RAPM: α=0.92; all three tasks, α=0.92). As in prior work (e.g., Redick et al., 2012), we used the aggregate score to estimate *Gf*.

Correlation between VET and SVET with experience. In prior work (e.g., Gauthier et al., 2014) self-report of experience in a domain correlated poorly with visual performance, but we expected a stronger relationship with semantic knowledge, because such knowledge is more observable in other people. First, we consider the relation between VET and SVET scores respectively with general experience individuating objects. VET performance was not strongly correlated with general experience. SVET performance for five of the categories (Transformer, bird, car, dinosaur, and leaf) showed a small but significant correlation with report of general experience individuating objects. Consistent with the idea that experience is largely useful as a domain-specific construct, all VETs (except mushrooms) and SVETs were significantly related to the domain specific report of experience. All such correlations were significantly higher (Steiger's Z, $p<0.05$) for the SVET (mean r=0.44) than the VET (mean r=0.27).

Table 7. Correlations (*r*) of general experience aggregate and category-specific experience aggregate for each category with VET and SVET accuracy as well as VET-Other and SVET-Other accuracy (the average of accuracy on the other 7 categories) in Study 2. Values shown in bolded red are statistically significant ($r_{Crit}(211)=.132$, $p<.05$).

	Correlation (*r*) of General Experience Aggregate with:		**Correlation (*r*) of Category Experience Aggregate with:**			
	VET	SVET	VET	SVET	VET-Other	SVET-Other
Car	0.11	**0.14**	**0.42**	**0.52**	0.03	0.13
Plane	0.10	0.09	**0.17**	**0.32**	-0.06	0.06
Transformer	**0.16**	**0.13**	**0.23**	**0.55**	-0.02	0.05
Dinosaur	0.06	**0.17**	**0.32**	**0.50**	0.08	**0.32**
Shoe	-0.02	-0.01	**0.50**	**0.62**	0.01	**-0.29**
Bird	**0.24**	**0.16**	**0.26**	**0.42**	0.06	**0.26**
Leaf	0.05	**0.14**	**0.15**	**0.36**	0.10	**0.24**
Mushroom	0.00	0.07	0.11	**0.28**	0.13	0.13

These correlations suggest convergent validity for the domain-specific abilities measured by the VET and SVET. These relations are relatively specific to each category: the correlation between VET and SVET with category experience was nearly always significantly greater (Steiger's Z, $p<0.05$) within category (with VET and SVET for the same category) than across category (with VET-Other and SVET-Other, the average of accuracy on the other seven categories for each test – see Table 7) with the exception of the VET-Leaf and VET-Mushroom.

Age and sex effects. Age was only correlated with *Gf* and four of the SVETs (car, bird, leaf and mushroom – see Table 8). The negative correlation between *Gf* and age is very likely due to sampling bias, as the majority of our subjects were from the university community and under 30 years of age, with above average *Gf* relative to the older subjects from the community.

Women performed better than men on the CFMT (Bowles et al., 2009; Cho et al., 2015; Duchaine & Nakayama, 2006). Our expectations for our four typical male-interest and four typical female-interest categories were largely borne out, however sex effect was larger for male-interest categories for SVET than VET. For both VET and SVET, only shoes demonstrated a strong female advantage.

Correlation with face recognition and fluid intelligence. Interestingly, *Gf* and CFMT were positively correlated: while this correlation is small, it has not been observed previously with other measures of intelligence or verbal ability (Davis et al., 2011; Hedley et al., 2011; Wilhelm et al., 2010; Wilmer et al., 2010). *Gf* was also positively correlated with VET accuracy for most categories, although not significantly for cars or shoes. That VETs were more strongly related to *Gf* than to the CFMT could be due to the fact that unfamiliar faces had no names, whereas common objects did. Perhaps surprisingly, the correlation between *Gf* and SVET was generally smaller than with the VET and significant only for planes, Transformers, dinosaurs, and mushrooms. This is surprising given the similarity between some measures used to estimate premorbid intelligence, like the "Spot-the-Word" test (Baddeley, Emslie, & Nimmo-Smith, 1993) and the SVET. This suggests that the vocabulary knowledge tested in the SVET is much more sensitive to domain-specific experience than to domain-general influences.

Even though the CFMT is a domain-specific test we would expect experience to be somewhat saturated for faces relative to other domains, suggesting that CFMT may also be interpreted as an estimate of a domain-general visual ability (potential to learn a visual category given unlimited experience; see Gauthier et al., 2014). As in prior studies (Gauthier et al., 2014; McGugin et al., 2012), CFMT scores were positively correlated with VET performance for all categories, suggesting that a common visual ability contributes to performance on all of these visual tests, but the correlations are not large, presumably because of the influence of domain-specific experience. As found before (Gauthier et al., 2014), the CFMT is not particularly distinct from the other visual tests (VETs): the average correlation between CFMT and VETs was r=0.26 (range = 0.19-0.38) (Table 9), while the mean pairwise correlation among VETs was r=0.33 (range = 0.09-0.46) (Table 10, Panel A). Interestingly, the CFMT's correlation was numerically

stronger with the average performance on all eight VETs (VET-All, r=0.40) than with any single category VET, suggesting that aggregating the VET across categories reduces category-specific contributions. The correlation between CFMT and SVET performance (mean r=0.08), a non-visual measure, was on average weaker than for CFMT-VET correlations (mean r=0.26, two-sided Fisher's Z=-2.69, p=0.007).

Table 8. Correlations of test accuracy with age, sex, *Gf* and CMFT in Study 2. The first column shows the reliability of each measure as Cronbach's alpha. Columns 2-5 show the correlation (r) of each accuracy on each measure with age, sex, *Gf* accuracy, and CFMT accuracy. Values shown in bolded red are statistically significant ($r_{Crit}(211)$=.132, p<.05).

		Correlation (*r*) of test with:			
	Cronbach's α	**Age**	**Sex**	***Gf***	**CFMT**
Gf	0.92	**-0.25**	0.06	-	
CFMT	0.92	-0.07	**-0.15**	**0.14**	-
VET-Car	0.90	-0.07	0.06	0.01	**0.21**
VET-Plane	0.89	0.08	0.09	**0.28**	**0.22**
VET-Transformer	0.89	-0.01	0.06	**0.29**	**0.28**
VET-Dinosaur	0.88	0.07	0.00	**0.29**	**0.26**
VET-Shoe	0.89	0.00	**-0.54**	0.10	**0.38**
VET-Bird	0.93	0.01	-0.05	**0.29**	**0.24**
VET-Leaf	0.84	-0.04	**-0.16**	**0.25**	**0.28**
VET-Mushroom	0.71	0.03	-0.12	**0.21**	**0.19**
SVET-Car	0.89	**0.22**	**0.28**	-0.08	0.06
SVET-Plane	0.72	0.10	**0.36**	**0.14**	-0.01
SVET-Transformer	0.74	0.02	**0.30**	**0.15**	0.07
SVET-Dinosaur	0.73	0.09	**0.24**	**0.25**	**0.14**
SVET-Shoe	0.91	0.04	**-0.55**	-0.11	**0.18**
SVET-Bird	0.52	**0.29**	0.02	0.04	0.06
SVET-Leaf	0.77	**0.39**	-0.01	0.04	0.06
SVET-Mushroom	0.66	**0.16**	0.13	**0.14**	0.10

Domain-specificity of VET-SVET relationship. We considered the relationship between performance on the VET and SVET within and between the eight categories (Table 9). Panel A of Table 10 reveals that nearly all VETs are positively correlated with each other, with

the exception of VET-Bird and VET-Mushroom with VET-Car. The pair-wise correlations between VET accuracy was r=0.32 for all typical male-interest categories (cars, planes, Transformers, and dinosaurs) and r=0.37 for typical female-interest categories (shoes, birds, leaves, mushrooms). However, these correlations were not on average much stronger than the correlations between VET categories across sex-interest: r=0.31 (e.g. VET-Car with VET-Shoe). Contrary to previous findings with the VET (McGugin et al., 2012), in this sample and with these modified tests, we did not observe strong sex effects in VET performance.

Panel C of Table 10 reveals a strong positive correlation between all of the male-interest categories (mean r=0.27), although for female-interest categories, SVET-Shoe performance appears to be unrelated to the living, female-interest categories, birds, leaves, and mushrooms, which were all positively correlated with each other (mean r=0.31). This suggests that while sex may predict domain-specific interest and experience, it is only one of several factors.

In considering the relations between VETs and SVETs, we were most interested in the link between VET and SVET accuracy for the same category, which might indicate how common category experience contributes to both visual and semantic performance (Panel B of Table 10, diagonal). VET and SVET performance within domain was positively correlated for all categories, with 6 out of 8 relationships significant. For seven of the eight categories, excluding birds, we found that the VET-SVET correlation was stronger within than between categories (the average of performance on the other seven SVET categories). This difference was significant using Steiger's Z (p<0.05) for cars, Transformers, and shoes. We also compared the relationship between SVET and VET within category versus across categories (average performance on the other 7 VET categories). For five out of eight categories, excluding birds, leaves, and

mushrooms, the SVET-VET correlation was greater within than between categories, an effect that was significant using Steiger's Z (p<.05) for cars, Transformers, dinosaurs, and shoes.

Table 9. Correlations (r) of VET and SVET accuracy for each category in Study 2. Panel A shows the correlations between each of the VETs. Panel B shows the correlations between each SVET and VET with within category correlations outlined along the diagonal. Panel C shoes the correlations between each of the SVETs. Values shown in bolded red are statistically significant ($r_{Crit}(211)$=.132, p<.05).

A. VETs

		1	2	3	4	5	6	7	8
1	VET-Car	-							
2	VET-Plane	**0.26**	-						
3	VET-Transformer	**0.25**	**0.46**	-					
4	VET-Dinosaur	**0.15**	**0.46**	**0.36**	-				
5	VET-Shoe	**0.24**	**0.28**	**0.25**	**0.25**	-			
6	VET-Bird	0.09	**0.39**	**0.44**	**0.44**	**0.26**	-		
7	VET-Leaf	**0.25**	**0.43**	**0.44**	**0.38**	**0.41**	**0.44**	-	
8	VET-Mushroom	0.09	**0.36**	**0.41**	**0.28**	**0.27**	**0.37**	**0.46**	-

B. SVET and VET

		1	2	3	4	5	6	7	8
9	SVET-Car	**0.45**	**0.25**	0.09	0.02	-0.08	-0.09	0.03	**-0.14**
10	SVET-Plane	0.08	**0.32**	**0.16**	**0.21**	**-0.18**	0.11	0.04	0.02
11	SVET-Transformer	0.05	**0.22**	**0.30**	**0.26**	-0.07	**0.25**	0.05	-0.03
12	SVET-Dinosaur	0.03	**0.14**	**0.18**	**0.31**	-0.11	**0.19**	0.02	-0.07
13	SVET-Shoe	**0.18**	-0.07	-0.10	-0.02	**0.40**	-0.05	0.03	-0.05
14	SVET-Bird	-0.13	0.03	0.02	**0.17**	0.02	**0.16**	-0.01	0.01
15	SVET-Leaf	0.01	0.11	0.11	**0.21**	0.11	**0.16**	0.08	-0.02
16	SVET-Mushroom	0.10	**0.24**	0.11	**0.29**	-0.01	**0.18**	0.07	0.12

C. SVETs

		9	10	11	12	13	14	15	16
9	SVET-Car	-							
10	SVET-Plane	**0.35**	-						
11	SVET-Transformer	**0.17**	**0.28**	-					
12	SVET-Dinosaur	**0.17**	**0.25**	**0.38**	-				
13	SVET-Shoe	0.07	**-0.18**	**-0.22**	-0.02	-			
14	SVET-Bird	0.06	**0.16**	0.03	**0.29**	-0.09	-		
15	SVET-Leaf	**0.31**	**0.22**	**0.19**	**0.30**	0.06	**0.38**	-	
16	SVET-Mushroom	**0.15**	**0.22**	**0.28**	**0.32**	0.01	**0.28**	**0.26**	-

Contribution of experience to within category VET-SVET correlations. One of our goals was to understand the shared variance between visual and semantic performance, which we hypothesized stem from domain-specific experience. To understand the contribution of category experience to the correlation between VET and SVET performance, we performed several multiple regressions that progressively account for more factors that can account for the shared variance between VET and SVET. After removing these domain-general influences, we ask if the self-reports of experience with each category account for the remainder of the shared VET-SVET variance.

The zero-order correlations between VET and SVET performance for each category are shown in Table 10, column A. Column B presents the partial correlations between VET and

SVET performance residuals for each category with age, sex, and *Gf* entered simultaneously as predictors. In most cases, we observed only a slight decrease in the correlation coefficient once the influence of these factors was accounted for. The notable exception is the VET-SVET-Shoe relationship, which was substantially reduced from column A to column B. This reduction is almost entirely due to the contribution of sex, which is strongly correlated with both VET-Shoe and SVET-Shoe performance. Overall, age, sex, and *Gf* do not account for much of the shared variance between VET and SVET performance.

Table 10. Correlations and partial correlations (*r*) of VET and SVET performance for each category in Study 2 – at top using sum scores and all SVET trials and bottom using theta and SVET-Select trials. Column A shows the zero-order correlations (also shown in *Table 9*, Panel B, diagonal boxes). Column B shows the partial correlations with age, sex, and *Gf* regressed out. Column C shows the partial correlations with age, sex, and *Gf* as well as VET-Other and SVET-Other performance regressed out. Column D shows the partial correlations with age, sex, *Gf* , VET-Other, and SVET-Other performance, as well as self-report category experience aggregate regressed out.

		A. Zero-Order	**B. Age, Sex, *Gf* Out**	**C. B + VET-Other, SVET-Other Out**	**D. C + Category Experience Out**
	α=.05:	r_{Crit}=.13	r_{Crit}=.13	r_{Crit}=.14	r_{Crit}=.14
Sum Score	Car	**0.45**	**0.49**	**0.49**	**0.37**
All Trials	Plane	**0.32**	**0.27**	**0.24**	**0.19**
	Transformer	**0.30**	**0.27**	**0.21**	0.08
	Dinosaur	**0.31**	**0.25**	**0.22**	0.13
	Shoe	**0.40**	**0.16**	**0.19**	0.12
	Bird	**0.16**	0.03	**0.17**	0.08
	Leaf	0.08	0.07	0.06	0.01
	Mushroom	0.12	0.09	0.11	0.09

Next, we asked how much of the shared variance between VET and SVET for a category might be the result of a domain-general ability reflected in performance across all tests. Because experience is going to be correlated across some categories, this is likely an overly conservative procedure, removing more than only domain-general influences. For each VET and SVET we created a non-category score for each subject (VET-Other and SVET-Other), which was the average performance on the other seven categories for that test (e.g. SVET-Other for cars is the average of performance on all SVETs except SVET-Car). We then used VET-Other and SVET-

Other as regressors to remove domain-general task performance from the VET-SVET relationship. Table 10, column C shows the partial correlations between VET and SVET performance residuals for each category with age, sex, *Gf* aggregate and, in addition, VET-Other and SVET-Other entered as simultaneous predictors. Again, we observe very little change in the correlation between VET and SVET performance after accounting for domain-general effects.

We hypothesize that the remaining shared variance between VET and SVET comes from domain-specific experience. To test this, we performed another regression including domain-specific experience as a predictor (Table 10, column D). Theoretically, if our hypothesis was correct *and* if our measure of category experience contained no measurement error (an assumption of mediation analyses that is almost universally violated; Baron & Kenny, 1986), we would expect the correlation between VET and SVET to be completely eliminated by regressing out experience. In some cases (e.g., Transformers, dinosaurs, birds), this analysis suggests that more of the VET-SVET shared variance is accounted for by domain-specific experience (drop from Column B and C) than by domain-general effects (drop from Column A and C). But there were also categories for which domain-general factors were more important (shoes) and some for which the shared variance was very small in the first place (leaves, mushrooms).

In two cases (cars and planes) correlation coefficients in Table 10, column D are still statistically significant, suggesting there is variance not explained by any of the regressors including experience. One reason for this could be the imprecision of our experience measure (Gauthier et al., 2014). There may be aspects of domain-specific experience that we did not measure well with our questionnaire. In supplemental materials, we explore one possible factor, the ability to use available verbal labels during the VET.

Study 2 Conclusions

Our analyses provide evidence of acceptable reliability in several of the SVET subtests. Our results also provide evidence for the validity of the SVET subtests, and simultaneously of the VET subtests, as domain-specific measures, by showing that they correlate more strongly within domain than across domains.

We extended use of the questionnaire of domain-general and domain-specific experience created by Gauthier et al. (2014). Measurement of experience in a domain remains in its infancy but should be an important goal for future work in high-level vision because of the undeniable contribution of experience in domain-specific performance on visual and non-visual tasks. We found partial evidence for our hypothesis that domain-specific experience accounts for the correlation between visual and semantic knowledge, but a stronger test will require better measures of experience.

We also found more evidence for the claim in McGugin et al., (2012) that individual differences in visual object recognition cannot realistically be measured with a single object domain (e.g., Dennett et al., 2012). Visual tasks that have equally high reliability can show quite different relations with measures of semantic knowledge that also have equally high reliability, as well as different relationships with fluid intelligence, age, sex, or estimates of experience. Finally, face recognition does not stand out as particularly independent from the recognition of other objects: for instance, car and bird recognition shared less variance than faces with any of the non-face object categories.

Study 3: Testing the SVET-Bird in Expert Birders

Studies 1 and 2 tested the SVET in samples that were not recruited with regards to their expertise for any category. The SVET appears to have sufficient range for these “normal” samples, but the SVET could also be useful for experts. In Study 3 we provide a case study with

the SVET-Bird in subjects with considerably more expertise. In particular, we ask if the SVET had sufficient range to discriminate individuals in such a population, and what the relation would be between the VET and SVET in experts. We were also interested in developing a more extensive questionnaire of experience specifically tailored to measure experience with birds.

We identified as many very good bird experts as we could from our available contacts, which essentially determined the sample size. Nonetheless we have 80% power of detecting an effect size of r=0.35. In the tradition of individual differences work, we do not correct for multiple comparisons but invite readers to focus on effect sizes (r-values) rather then significance per se, especially in using the results to predict further work with the SVET.

Methods

Subjects. We recruited 64 subjects by email to participate in an online study. All subjects self-reported substantial experience in bird watching as a hobby or profession. The study was approved by the Vanderbilt IRB. As compensation, subjects who completed all parts of the study were entered in a lottery to win $50. Two additional subjects who began the study but chose not to complete all parts were not included. One subject who completed all parts was excluded because they misunderstood the VET instructions, resulting in below chance performance. The data reported here are for 63 subjects (29 male) aged 23-82 (mean=50.86, SD=15.16). All subjects reported speaking English and living in the United States or Canada. Our subjects resided in several locations across North America, with clusters in British Columbia, upstate New York, Colorado, Utah, and Tennessee. All subjects specifically reported experience with passerine birds.

Procedure. Subjects completed four tasks in the following order: an extended bird experience questionnaire for birders, the VET-Bird, the SVET-Bird, and the bird image naming

task from Study 2. The birder experience questionnaire first asked the same questions used in Study 2, four domain-general and seven bird specific; all were on a scale from 1-9 except for the duration of interest in birds for which they were asked to enter the number of years. We added 11 additional bird-specific questions, for example how frequently they go birding, how often they plan vacations around birding, if they belong to birding groups, and approximately how many different types of birds they have observed in person while birding during their lifetime (for full set of extended birder questions see Supplemental materials).

These tasks were completed using two online platforms: REDCap survey data collection tools were used for the experience questionnaire and bird naming test as in Study 2, and our own testing website for the VET and the SVET. Subjects indicated their response by clicking on an image or name, instead of typing a number from 1 to 3.

Results and discussion

Accuracy on VET-Bird, SVET-Bird, and bird naming. The birders performed very well on the VET-Bird, SVET-Bird, and bird naming (see mean and SD in Table 11). No subjects were excluded due to catch trials. In this dataset, age and sex were correlated (r=0.31, p=0.02). This renders any relations between performance and age or sex difficult to interpret, so we will not consider these variables here.

Self-reported bird experience. Responses on the extended bird experience questionnaire for birders revealed high levels of self-reported experience with birds including many years birding, much time spent birding, looking at birds, and reading about birds (Table 11). Based on reports of the number of birds sighted during their lifetimes, frequency of birding including on bird-related trips and vacations, and involvement in birding organization and events, we can be fairly certain that we sampled a group of truly experienced birders.

The seven bird-specific questions used in Study 2 again demonstrated high internal consistency (Cronbach's alpha=0.82). We computed a bird-experience aggregate for each subject (basic bird aggregate) as the average of the Z-scored reports for each of the questions. The eleven extended bird questions specifically for birders also demonstrated good internal consistency (Cronbach's alpha=0.59). The basic and extended questions were also correlated (average r=0.40, Cronbach's alpha=0.80).

One question, age at which interest in birds/birding became intense, was poorly correlated with the other birder questions, while a similar question, age at which you became interested in birds/birding, was more consistent, so we removed the age intensity question from later analyses and did not include it in the aggregate measure. Using the other ten birder-specific questions and the seven basic bird experience questions we computed an extended bird experience aggregate score for each subject as the average of the Z-scored reports for each of the 17 questions.

Interestingly, the domain-general experience aggregate score (the average of ratings on the four domain-general experience questions) was highly correlated with the extended bird experience aggregate (r=0.69, $p\leq$0.0001). It is possible that with an expert population, asking about general experience with all objects still leads subjects to reflect on experience with their primary category of expertise. Still, domain-specific aggregate scores outperformed this domain-general experience measure in predicting VET and SVET scores.

Table 11. Results from birders in Study 3. The first column shows the mean and standard deviation for accuracy on each task, age, and self-reports of experience. The second and third columns give the correlations (*r*) between VET-Bird and SVET-Bird accuracy and each measure. Correlation coefficients shown in bolded red are statistically significant ($r_{Crit}(62)=.25$, $p<.05$).

	Mean (SD)	**Corr. VET-Bird**	**Corr. SVET-Bird**
VET-Bird	0.96 (0.08)	-	-
SVET-Bird	0.96 (0.07)	**0.43**	-
Bird Naming	14.95 (3.77)	**0.67**	**0.55**
General Experience (1-9):			
General Experience Aggregate	6.79 (1.09)	0.20	0.22
Bird-specific Experience (1-9):			
Overall Expertise	6.90 (1.36)	**0.48**	**0.45**
Importance	7.81 (1.27)	0.13	0.16
Duration Interest (years)	27.17 (17.95)	0.08	0.08
Visual Memory	7.00 (2.26)	0.10	0.12
Image Frequency	7.90 (1.59)	**0.32**	**0.30**
Text Frequency	7.79 (1.85)	**0.37**	**0.26**
Essay	6.21 (2.06)	0.24	**0.33**
Basic Bird Experience Agg. (Z-score)	-	**0.48**	**0.45**
Birder Extended Questions:			
Age Started (age)	20.33 (13.43)	**-0.44**	-0.14
Age Intense (age)	26.46 (13.53)	**-0.42**	-0.16
Birding Frequency (1-7)	5.86 (1.28)	**0.25**	**0.41**
Travel (1-5)	3.60 (1.36)	**0.26**	**0.41**
Vacation (1-6)	3.71 (1.68)	-0.13	**-0.26**
Log of Sightings (1-3)	2.49 (0.69)	**0.30**	**0.33**
Birds Sighted (number)	714.37 (811.89)	0.23	**0.32**
Local Expertise (1-7)	4.14 (1.29)	**0.32**	**0.28**
Periodicals (number)	1.46 (1.38)	0.25	-0.01
Organizations (number)	2.16 (1.61)	0.20	0.10
Events (1-7)	2.43 (1.28)	0.20	**0.28**
Extended Bird Experience Agg. (Z-score)	-	**0.44**	**0.43**

Correlations between SVET-Bird and VET-Bird and experience. Several of the category-specific experience questions (both in the original and extended sets) were related to VET and SVET performance individually. The aggregate experience scores for the original seven questions, as well as for the extended set, significantly predicted both VET and SVET performance (Table 11). Interestingly, the category-specific overall expertise question ("Rate your expertise with *(category)* considering your interest in, years of exposure to, knowledge of,

and familiarity with *(category)*") did just as well on its own as did these aggregates. This is consistent with previous findings (Gauthier et al., 2014; McGugin et al., 2012) that this omnibus question of expertise is remarkably informative given its simplicity, and here even in experts.

Correlations between VET-Bird and SVET-Bird and bird naming. Besides measures of experience, relationships with the bird naming test can also provide further convergent validity. The number of birds correctly named was positively correlated with both VET-Bird and SVET-Bird performance (Table 11). Bird naming was also strongly correlated with extended bird experience aggregate scores (r=0.72, $p\leq$0.0001). To investigate the possible overlap between naming and these visual and semantic measures, we asked if VET and SVET contribute independently to naming performance for birds. We performed a multiple regression on bird naming performance entering VET-Bird and SVET-Bird simultaneously as predictors (N=59 after removing three subjects who had very large externally studentized residuals (>2.5) in the correlation between VET and SVET; see blue X's in Figure 5). The VET-Bird and SVET-Bird each made independent contributions to bird naming, and together accounted for 60.8% of the variance (Table 12). Adding the bird aggregate experience scores to this model leads to an impressive R^2 adjusted of 71%.

Table 12. Results of a simultaneous multiple regression predicting bird naming performance with VET-Bird and SVET-Bird performance for birders (N=59) in Study 3.

Model and predictor	β	*SE*	*t*	*p*
Bird Naming (R^2 adj = 60.8%)				
Intercept	-53.079	7.127	-7.450	≤0.0001
VET-Bird	25.412	7.942	3.200	0.002
SVET-Bird	44.637	8.824	5.060	≤ 0.0001

It is also useful to consider this expert data together on a continuum with non-expert data to observe the spread of performance from novice to expert. Image naming is a task that has been used in prior expertise work (Barton et al., 2009). Figure 4 shows VET-Bird and SVET-Bird

accuracy with bird naming scores for both the large sample of subjects collected at Vanderbilt University in Study 2 and the birders collected online in Study 3. Considering both groups of subjects together (N=275), the correlations between each measure and naming were high (VET-Bird and bird naming: $r(273)=0.74$, $p<0.0001$; SVET-Bird and bird naming: $r(273)=0.93$, $p<0.0001$), although the strength of these relationships may be driven primarily by large group differences between those who could not name any birds and those who could. Nevertheless, it is interesting to note variability in performance for both a general and an expert sample.

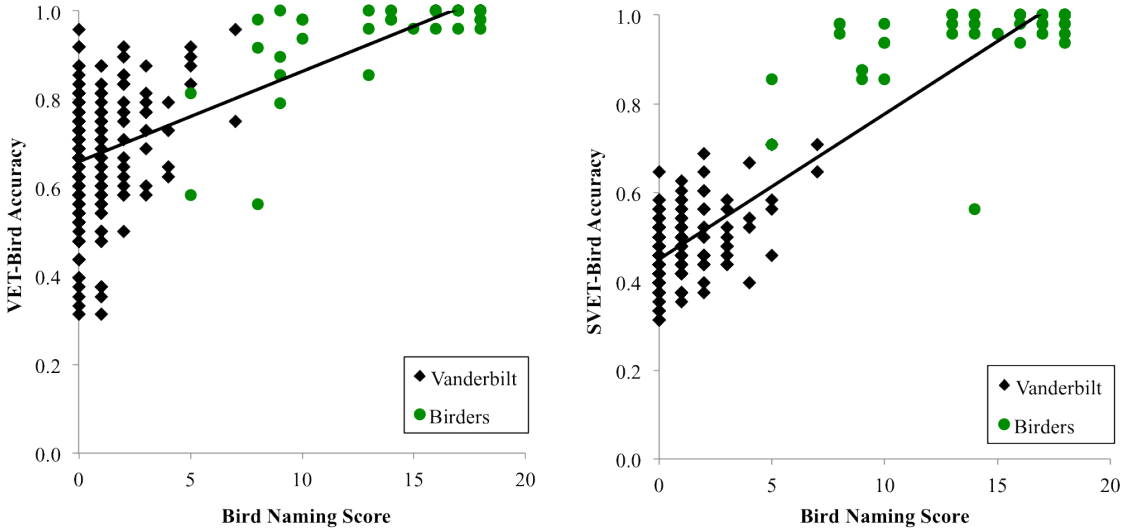

Figure 4. Scatterplots showing the relationship between bird naming score and VET-Bird (left; $r(273)=0.74$, $p<0.0001$) and SVET-Bird (right; $r(273)=0.93$, $p<0.0001$) in Vanderbilt subjects from Study 2 (black diamonds) and Birders from Study 3 (green circles).

Correlations between VET-Bird and SVET-Bird. Among birders, performance on the VET-Bird and SVET-Bird was positively correlated, $r(61)=0.43$, $p=0.005$, suggesting that the tests are reasonably discriminative despite approaching ceiling. We can also consider this relation for non-expert subjects from Study 2 together with birders in Study 3 to ask if the two tasks separate novices from experts. Considering both groups together (N=278), the SVET-VET correlation for birds was high, $r(276)=0.70$, $p<.0001$, but this was largely due to a separation of the two groups (Figure 5).

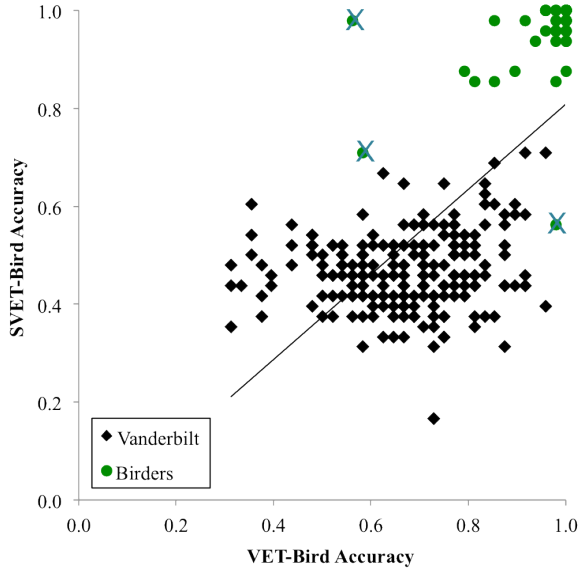

Figure 5. Scatterplots showing the relation between SVET-Bird and VET-Bird accuracy for Vanderbilt subjects from Study 2 (black diamonds) and birders from Study 3 (green circles). Three birders marked with blue X's denote subjects with very high externally studentized residuals (>2.5).

Study 3 Conclusions

In Study 3 we tested the SVET-Bird in an expert population, as well as an extended questionnaire of experience with birds. While birders' performance was high on all of the bird relevant measures, there was sufficient individual variability to observe strong relationships between performance on each test and experience. For future work with experts, extended versions of the VET, SVET, and the naming test with additional difficult trials to reduce ceiling effects for experts could help better resolve individual differences at the high end of performance. Critically, adding more difficult trials to the VET and SVET would be relatively easy, while trying to design an easier naming task that can better discriminate among novices would be much more difficult. Thus, if the goal is to measure visual and semantic skills along a very wide continuum, the VET and the SVET formats may be more easily adapted than a naming task.

General Conclusions

Our goal was to create a battery of tests to measure semantic knowledge across a number of object categories for use in parallel with measures of visual expertise like the VET (McGugin et al., 2012). We wanted to provide evidence for these measures' reliability and validity, and demonstrate how they can be used to study how domain-general abilities and domain-specific experience contribute to knowledge about objects.

The SVET 1.0 includes scales for eight object categories, four living categories and four non-living, attempting to balance domains for which men vs. women may be expected to show a relative advantage. The concise task measures only one aspect of semantic knowledge, knowledge of object names and labels for a category and can be completed by subjects who have low, moderate, or high levels of expertise in a domain. The SVET format is adaptable to many categories that have relevant nomenclature, and the tests could also be easily adapted for use in expert populations with the addition of more difficult trials. Researchers studying populations outside of the US or Canada may not be able to use the current tests, but could develop new measures in the same format.

The SVET 1.0 demonstrated acceptable internal consistency both in the laboratory in a young adult sample (mean age = 22.49, SD=6.31), and online in a sample that was on average older and also covered a larger age range (mean age = 35.82, SD=12.45). Five of the subtests (shoe, car, leaf, plane, and Transformer) had internal consistency over .7 in both samples, and the other three reached .7 in one of the two samples.

We provided evidence for the validity of the SVET (and the VET) in a number of ways: i) SVETs showed domain-specific correlations with their corresponding VETs, a result that combines both convergent and discriminant validity; ii) similarly, SVETs showed domain-

specific correlations with reports of experience; iii) the SVET discriminated experts from novices in the bird domain; and iv) SVETs were more related to domain-specific measures than to fluid intelligence or face recognition. Thus, measuring performance with several object categories allows the validation of the domain-specificity of each test.

We hypothesized that the shared variance between the VET and SVET should reflect primarily domain-specific experience. In other words, after we removed the influence of domain-general variables (age, sex, *Gf*, and general ability in the VET and the SVET), we found that experience as measured by our self-report measure contributed independently and significantly to the VET-SVET relation for six of the eight categories we tested. Indeed, in several cases, regressing out experience left very little shared variance between VET and SVET, less than 3% (the exception was cars, where the correlation still accounted for 16% of the variance). Whether the differences between categories we found here replicate or depend on properties of our specific sample remains to be seen. With a category for which there was great variability in the ability to name objects at the subordinate-level, birds, we found that while bird naming may account for some of the variance shared by VET and SVET, this variance was independent of self-reported experience with birds. In sum, we expect that as the measurement of experience improves, this construct will consistently and fully account for the overlap between visual and semantic knowledge.

This research program represents a new direction in individual differences work in high-level vision. We applied a framework in which domain-specific measures can be used to estimate domain-specific effects after partialing out domain-general variance. Similar methods have been used in cognitive domains (Hambrick, 2003; Hambrick et al., 2007; 2008; Stanovich & Cunningham, 1992) but have not been applied to questions in object recognition. Our main goal

was to provide a means to measure non-visual domain-specific knowledge in a reliable manner, thereby demonstrating that individuals acquire at least two different kinds of knowledge (visual and non-visual) from experience with a category. The creation of the SVET, and its use with the VET, should have many applications. They could be useful in the evaluation of patients with visual or semantic deficits (e.g., Bozeat, Ralph, Patterson, Garrard, & Hodges, 2000; Behrmann & Avidan, 2005), in testing various explanations of sex differences in different domains (e.g., Gainotti, 2005; Laiacona et al., 2006), and they should facilitate the development of models describing how different aspects of knowledge and experience are related, via more sophisticated methods like structural equation modeling.

Acknowledgments

This work was supported by NSF (SBE-0542013), the Vanderbilt Vision Research Center (P30-EY008126), and the National Eye Institute (R01 EY013441). We thank Jackie Floyd, Yimin Qiu, and David Nelwan for their assistance with data collection, and Simen Hagen, Jianhong Shen, Ryan Farrell, and Tom McKay for assistance recruiting birders. Thanks to Tom Palmeri, Jim Tanaka, Tim McNamara, and Jenn Richler for comments on the work.

References

Baddeley, A., Emslie, H., & Nimmo-Smith, I. (1993). The Spot-the-Word test: a robust estimate of verbal intelligence based on lexical decision. *British Journal of Clinical Psychology*, *32*, 55–65.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182.

Barton, J. J. S., Hanif, H., & Ashraf, S. (2009). Relating visual to verbal semantic knowledge: the evaluation of object recognition in prosopagnosia. *Brain*, *132*(12), 3456–3466. doi:10.1093/brain/awp252

Behrmann, M., & Avidan, G. (2005). Congenital prosopagnosia: face-blind from birth. *Trends in Cognitive Sciences*, *9*(4), 180-187. doi:10.1016/j.tics.2005.02.011

Bilalic, M., Langner, R., Erb, M., & Grodd, W. (2010). Mechanisms and neural basis of object and pattern recognition: a study with chess experts. *Journal of Experimental Psychology: General*, *139*(4), 728–742. doi:10.1037/a0020756

Bilalic, M., Langner, R., Ulrich, R., & Grodd, W. (2011). Many Faces of Expertise: Fusiform Face Area in Chess Experts and Novices. *Journal of Neuroscience*, *31*(28), 10206–10214. doi:10.1523/JNEUROSCI.5727-10.2011

Bowles, D. C., McKone, E., Dawel, A., Duchaine, B., Palermo, R., Schmalzl, L., et al. (2009). Diagnosing prosopagnosia: effects of ageing, sex, and participant-stimulus ethnic match on the Cambridge Face Memory Test and Cambridge Face Perception Test. *Cognitive Neuropsychology*, *26*(5), 423–455. doi:10.1080/02643290903343149

Bozeat, S., Ralph, M. A. L., Patterson, K., Garrard, P., & Hodges, J. R. (2000). Non-verbal

semantic impairment in semantic dementia. *Neuropsychologia*, *38*(9), 1207-1215. doi: 10.1016/S0028-3932(00)00034-8

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*(3), 305–327.

Cho, S.-J., Wilmer, J., Herzmann, G., McGugin, R., Fiset, D., Van Gulick, A. E., et al. (2015). Item Response Theory Analyses of the Cambridge Face Memory Test (CFMT). *Psychological Assessment*. doi:10.1037/pas0000068

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE*, *8*(3), e57410. doi:10.1371/journal.pone.0057410.t001

Davis, J. M., McKone, E., Dennett, H., O'Connor, K. B., O'Kearney, R., & Palermo, R. (2011). Individual differences in the ability to recognise facial identity are associated with social anxiety. *PLoS ONE*, *6*(12), e28800. doi:10.1371/journal.pone.0028800

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

Duchaine, Brad, & 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. doi:10.1016/j.neuropsychologia.2005.07.001

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for kit of factor-*

referenced cognitive tests. Princeton, NJ: Educational Testing Services.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence-Erlbaum.

Gainotti, G. (2005). The influence of gender and lesion location on naming disorders for animals, plants and artefacts. *Neuropsychologia*, 43, 1633-1644. doi:10.1016/j.neurospcyhologia.2005.01.016

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), 1–12. doi: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. doi:10.1016/j.cognition.2010.11.002

Hambrick, D. Z. (2003). Why are some people more knowledgeable than others? A longitudinal study of knowledge acquisition. *Memory & Cognition*, *31*(6), 902–917.

Hambrick, D. Z., Meinz, E. J., & Oswald, F. L. (2007). Individual differences in current events knowledge: Contributions of ability, personality, and interests. *Memory & Cognition*, *35*(2), 304–316.

Hambrick, D. Z., Pink, J. E., Meinz, E. J., Pettibone, J. C., & Oswald, F. L. (2008). The roles of ability, personality, and interests in acquiring current events knowledge: A longitudinal study. *Intelligence*, *36*(3), 261–278. doi:10.1016/j.intell.2007.06.004

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap) - A metadata-driven methodology and workflow process for providing translational research information support. *Journal of*

Biomedical Informatics, *42*(2), 377–381.

Hedley, D., Brewer, N., & Young, R. (2011). Face recognition performance of individuals with Asperger syndrome on the Cambridge Face Memory Test. *Autism Research : Official Journal of the International Society for Autism Research*, *4*(6), 449–455. doi:10.1002/aur.214

Humphreys, G. W., Lamote, C., & Lloyd-Jones, T. J. (1995). An interactive activation approach to object processing: Effects of structural similarity, name frequency, and task in normality and pathology. *Memory*, *3*, 535–586.

Laiacona, M., Barbarotto, R. & Capitani, E. (2006). Human evolution and the brain representation of semantic knowledge: is there a role for sex differences? *Evolution and Human Behavior*, 27, 158-168. doi:10.1016/j.evolhumbehav.2005.08.002

McGugin, R. W., Newton, A. T., Gore, J. C., & Gauthier, I. (2014a). Robust expertise effects in right FFA. *Neuropsychologia*, 63, 135-144. doi:10.1016/j.neuropsychologia.2014.08.029

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. doi:10.1016/j.visres.2012.07.014

McGugin, R. W., Van Gulick, A. E., Tamber-Rosenau, B. J., Ross, D. A., & Gauthier, I. (2014b). Expertise effects in face-selective areas are robust to clutter and diverted attention, but not to competition. *Cerebral Cortex*. doi:10.1093/cercor/bhu060

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science*, *23*(3), 184–188. doi:10.1177/0963721414531598

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices and*

Vocabulary Scales. New York, NY: Psychological Corporation.

Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., et al. (2012). No Evidence of Intelligence Improvement After Working Memory Training: A Randomized, Placebo-Controlled Study. *Journal of Experimental Psychology: General*. doi:10.1037/a0029082

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

Stanovich, K. E., & Cunningham, A. E. (1992). Studying the consequences of literacy within a literate society: The cognitive correlates of print exposure. *Memory & Cognition*, *20*(1), 51–68.

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

Thurstone, L. L. (1938). *Primary Mental Abilities*. Chicago, IL: University of Chicago Press.

Wilhelm, O., Herzmann, G., Kunina, O., Danthiir, V., Schacht, A., & Sommer, W. (2010). Individual differences in perceiving and recognizing faces-One element of social cognition. *Journal of Personality and Social Psychology*, *99*(3), 530–548. doi:10.1037/a0019972

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., et al. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences*, *107*(11), 5238–5241.

Zell, E., & Krizan, Z. (2014). Do People Have Insight Into Their Abilities? A Metasynthesis. *Perspectives on Psychological Science*, *9*(2), 111–125. doi:10.1177/1745691613518075