Measuring object recognition ability: reliability, validity, and the aggregate z-score approach.

Conor J. R. Smithson[1]*, Jason K. Chow[1]*, Ting-Yun Chang[1] & Isabel Gauthier[1]

[1]Department of Psychology, Vanderbilt University

*indicates equal co-first-authorship

Keywords:

Object Recognition, Individual Differences, Measurement, High-level vision

Corresponding Author:

Conor J. R. Smithson, Email: conor.smithson@vanderbilt.edu

Abstract

Measurement of domain-general object recognition ability (*o*) requires minimization of domain-specific variance. One approach is to model *o* as a latent variable explaining performance on a battery of tests which differ in task demands and stimuli; however, time and sample requirements may be prohibitive. Alternatively, an aggregate measure of *o* can be obtained by averaging z-scores across tests. Using data from Sunday et al. (2022), we demonstrate that aggregate scores from just two such object recognition tests provide a good approximation (r = .79) of factor scores calculated from a model using a much larger set of tests. Some test combinations produced correlations of up to r = .87 with factor scores. We then revise these tests to reduce testing time, and develop an odd-one-out task, using a unique object category on each trial, to increase task and stimuli diversity. To test our measures, 163 participants completed the object recognition tests on two occasions, one month apart. Providing the first evidence that *o* is stable over time, our short aggregate *o* measure demonstrated good test-retest reliability (r = .77). The stability of *o* could not be completely accounted for by intelligence, perceptual speed, and early visual ability. Structural equation modelling suggests our tests load significantly onto the same latent variable, and reveals that as a latent variable, *o* is highly stable (r = .93). Aggregation is an efficient method to estimate *o*, allowing investigation of individual differences in object recognition ability to be more accessible in future studies.

Introduction

Despite a long history of individual differences research in psychology (Cronbach & Meehl, 1955; Spearman, 1904), there has remained a divide between the experimental and the correlational tradition (Cronbach, 1957) in vision research, with the focus primarily on the experimental (Wilmer, 2008). However, in recent years there has been an increase in interest in individual differences in high-level vision, particularly in the area of object recognition (Dennett et al., 2012; Gauthier et al., 2022; McGugin et al., 2012; Richler et al., 2019). Vision science has primarily focused on effects at the group level, often neglecting the value of variation between individuals. The study of individual differences can help validate measures (Vogel & Awh, 2008) and test hypotheses about mechanisms (Mollon et al., 2017), including their functional organization and their utility (Wilmer, 2008). Measures of visual abilities designed to capture individual differences allow researchers to answer novel and important theoretical questions that cannot be assessed at the group level and may have practical applications in identifying individuals of exceptional ability, or of unusual inability.

Object recognition ability is the ability to discriminate between visually similar objects, or the ability to make object category judgments. We focus particularly on within-category discrimination, due to its relevance to many important but difficult human activities. Individual differences in this ability can be specific to individual domains (e.g. bird recognition), sometimes as a result of differential levels of experience, but can also reflect domain-general differences in ability (Gauthier, 2018). Measures of object recognition ability will always tap into the domain-general ability but may tap more or less into abilities specific to particular domains. Researchers who are interested in domain-general abilities must always consider the domain-generality of their measures. In this paper we discuss an aggregate approach to measuring domain-general object recognition ability and provide a new set of object-recognition tests for research use.

The measurement of individual differences in object recognition is rooted in the development of tests targeting specific domains, in particular, that of face recognition. One of the

most widely used measures is the Cambridge Face Memory Test (CFMT; Cho et al., 2015; Duchaine & Nakayama, 2006), but several similar tests now exist (see White & Burton, 2022 for a review). Building on the success of the CFMT, several measures of object recognition ability were developed for other object categories, such as The Cambridge Car Memory Test (Dennett et al., 2012). The Vanderbilt Expertise Test (VET; McGugin et al., 2012) was developed to test for expertise across many categories. The existence of substantial correlations between recognition ability on the VET for different object categories suggested that there may be a domain-general ability. To identify a truly domain-general object recognition ability, it is necessary to account for variance caused by 1) experience with specific object categories; 2) the kinds of features diagnostic for objects in specific categories; and 3) the specific demands of the tasks used to measure performance. To eliminate experience-related variance, novel categories of objects such as Greebles (Gauthier & Tarr, 1997) can be used, such as is done in the Novel Object Memory Test (Richler et al., 2017a). Alternatively, latent variable modelling or the aggregation of scores across tests can combine performance for multiple familiar object categories, minimizing the influence of variation in experience and also any other property of a specific category (e.g., whether texture is diagnostic, or whether the objects are symmetrical). The same approaches can be used to reduce the influence of diagnostic feature types and task demands to ensure that what is measured is not specific to strategies or skills that are only beneficial for certain tasks, such as matching strategies (Growns et al., 2022) or reliance on visual short term memory (Vogel & Awh, 2008).

Using structural equation modelling (SEM), Richler et al. (2019) showed that a higher order factor explained 89% of variance in performance across memory and matching tasks for different object categories, strongly supporting the existence of a domain-general object recognition ability (*o*). Further work demonstrated that this ability was common to both novel and familiar object categories, as a higher order factor for novel object recognition correlated nearly perfectly (r = .98) with a higher order factor for familiar object recognition (Sunday et al., 2022). Additional investigation has begun to position *o* in relation to other abilities: *o* can predict performance in

medical imaging tasks requiring visual search, such as the identification of lung nodules (Sunday et al., 2018), and it can do so separately from intelligence and experience. It can also predict categorization of music notes (Chang & Gauthier, 2021), or judgements of whether blood cells are cancerous (Smithson et al., 2023). *O* also predicts the accuracy with which people can estimate summary statistics (e.g. mean) of groups of objects (Chang & Gauthier, 2022), and ability to recognize different types of food (Gauthier & Fiestan, 2023), whereas it does not predict experience with visual arts (Chow et al., 2022b). Beyond visual abilities, *o* is also related to haptic object recognition accuracy (Chow et al., 2022a), and has a strong correlation with auditory object recognition accuracy (Chow et al., 2023). Neural correlates of *o* have been identified, and are distributed across the visual cortex, ventral visual areas of the occipitotemporal cortex, and parietal areas that are also implicated in shape perception (McGugin et al., 2022).

In many of the recent studies exploring relationships between *o* and other variables, researchers did not use SEM with several tasks and categories, as was used in the initial identification of the construct. An SEM approach has several advantages, as it can estimate latent variables that are free from construct-irrelevant variance and measurement error (Bollen, 1989). However, the approach requires several indicators for each latent variable and large enough samples to fit models, which is why many studies have used an aggregate approach. To efficiently obtain estimates of *o,* aggregates of two tests (one memory task and one matching task) were used, with the tests varying in their task requirements and in the object categories of their stimuli. This approach has typically led to correlations around .25-.4 between the two object recognition tests, an effect size that is expected because the two tests do not have much overlap other than via the higher-level ability. By the principle of aggregation, the influence of task-specific variance on final scores is reduced, and the influence of the construct of interest that is partly measured by both tasks is increased (Rushton et al., 1983). While this approach was assumed to measure the same construct as that identified in SEM research using the same tests, it had not been directly compared to it. In this work, we have several goals: 1) to compare, using an existing dataset, the aggregate estimation

of *o* using two tests with factor scores based on a confirmatory factor model that includes a larger number of tests (here 6); 2) to increase the efficiency (in terms of testing time) of the aggregate method by modifying the matching test; 3) to expand the construct coverage of a general *o* by including a new task with task demands and stimuli that differ from the existing tasks; 4) to test the convergent validity of our measures; 5) to measure the test-retest reliability of the aggregate method and to estimate the stability of the *o* construct; and 6) to further map the nomological net surrounding *o*, using new measures that have not been related to the construct before, and to assess whether longitudinally stable variance in *o* survives the partialling out of measures of several theoretically related abilities.

Part I - Comparing the two-test aggregates and six-test factor scores with data from Sunday et al. (2022).

First, we reanalyzed existing data to explicitly compare the aggregate method using just two tests with factor scores calculated from a larger confirmatory factor model. This dataset contains participants' scores on two sets of six object recognition measures, each set using either novel or familiar object categories. The dataset provides an opportunity to compare aggregate scores calculated from one set of tests with factor scores calculated from the other set. Sunday et al. (2022) recruited 294 participants from the Vanderbilt University community. Participants completed 18 tests across two 2-hour sessions, no more than a week apart. Participants completed 3 tasks for each of 6 object categories. Three object categories were novel (symmetrical Greebles or s-Greebles, vertical Ziggerins or v-Ziggerins, and Sheinbugs[1]), and 3 categories were familiar (birds, planes, and Transformer toys in their robot form). Participants completed two types of object recognition tasks – learned exemplars (LE) and matching (MA). Participants also completed ensemble perception tasks for each category, but we do not use these results in the present analyses. Session one included MA

[1] There are two sets of Greebles, symmetrical and asymmetrical, and two sets of Ziggerins, vertical and horizontal, that were used in the original SEM study on *o* (Richler et al. 2017).

tests for Transformer toys, birds, and s-Greebles, and LE tests for birds and Sheinbugs. Session two included MA tests for v-Ziggerins, Sheinbugs, and planes, and LE tests for Transformer toys, planes, s-Greebles, and v-Ziggerins. The test sequence and trial order for each test were fixed for all participants to avoid disparate order effects which may impair measurement of individual differences.

Learning Exemplars Task

For each test, participants had to learn and then try to recognize six target objects from one object category. Participants studied 6 exemplars shown simultaneously, and they were given the chance to review them after trials 6 and 24. Each test included 48 three-alternative forced choice trials, in which they had unlimited time to select the object they thought was one of the six target objects from an array of three. Performance was calculated as percent accuracy (further details of the tests are described in Sunday et al., 2022)

Matching Task

On each trial participants were asked to compare two sequentially presented objects and determine if they have the same identity. Each test for any given category included 360 trials, with equal numbers of same and different trials. Each trial began with the presentation of an object for 300 ms for the first half of the test, and 150 ms for the remaining trials. This was followed by a 500 ms scrambled mask specific to the object category presented on that trial, and then a probe object to which participants were given 3000 ms to respond. A fixation cross was displayed for 500 ms as an interstimulus interval. Performance was calculated as sensitivity (*d'*). Further details of these tests are described in Sunday et al. (2022).

Analyses and Discussion

To assess the criterion validity of aggregate *o* scores calculated from two measures, we compared these scores to factor scores calculated using the maximum a posteriori method from the full model (Figure 1). This model closely resembles that used in figure 5 from Sunday et al. (2022), without the ensemble processing variable and indicators. Factor scores were calculated separately for novel and familiar objects. Aggregate scores were calculated from every combination of one MA task and one LE task that use different object categories, within both subsets of tasks for novel and familiar categories. There are therefore 6 sets of possible aggregate scores each for novel and familiar objects. To avoid comparing scores calculated from the same data, we compare aggregate scores from two *o* tasks using categories from either novel or familiar objects, with factor scores calculated for the opposite type of object. A schematic example of the approach is in Figure 1. Although in principle, correlating across different object-familiarity levels could reduce the correlation between aggregate *o* scores and factor scores, Sunday et al. (2022) found that the correlation between o_N and o_F was $r = .98$, so the two abilities are practically the same and an attenuation of the correlations for this reason is unlikely.

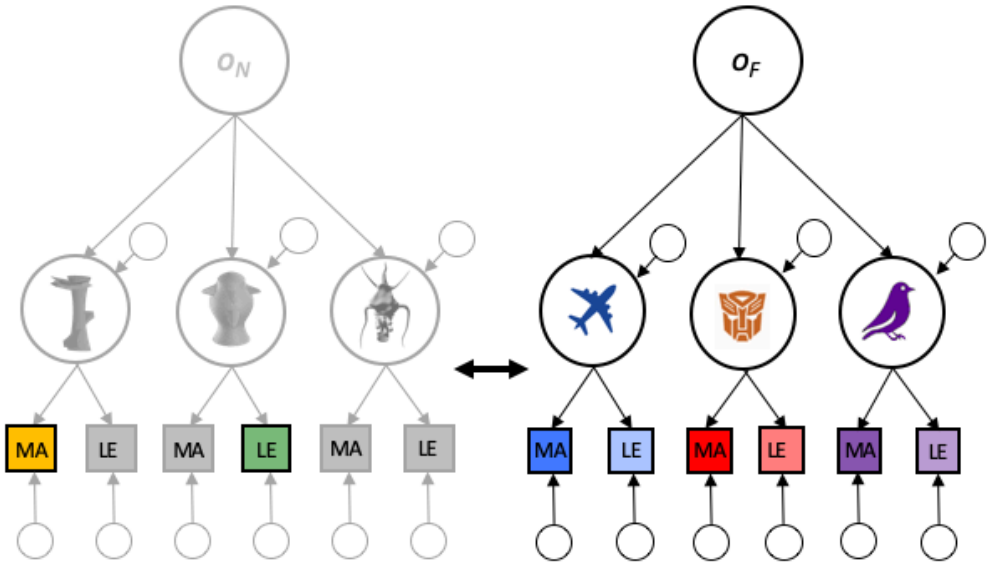


Figure 1. Diagram showing an example of how tests were selected from Sunday et al.'s (2022) model for our comparisons. In this example, the two novel object tests that are shown in color on the left are used to create an aggregate score (MA from one object category, LE from another object category, while the greyed-out tests are not used), whereas all six of the familiar object tests on the right are used to create a factor score. The aggregate score uses equally weighted tests, whereas the weights for the factor score calculation were obtained from the SEM model. The correlation between the aggregate score and the factor score is then calculated.

Although Sunday et al. (2022) collected data from 294 participants and factor scores were calculated on the entire sample (using methods robust to missing data or outliers), for aggregate scores we excluded participants with missing data because they would affect aggregate calculations. This left 210 participants. Table 1 shows correlations between factor scores and aggregate scores. We Fisher transformed the correlations before calculating the mean, and then inverse transformed the mean back to *r*, as is recommended practice to obtain less biased estimates (Corey et al., 1998). The mean correlation between aggregate scores and factor scores was $r = .79$. For comparison, the correlation between the aggregate of all 6 novel object tests and the familiar object factor score was $r = .95$). Interestingly, the aggregates of two novel object tests were better estimates of familiar object factor scores ($r = .82$) than aggregates of familiar object tests were for novel object factor scores ($r = .75$), a difference that cannot be attributed to the different factor scores as they were

nearly identical (*r* = .99). Although a two-sided Fisher's z-test (.82 – .75; z = 1.878, *p* = .06; 95% CI [-0.03, 0.15]) was not significant, the difference is consistent with previous evidence that novel object recognition tests correlate more highly with each other than do familiar object tests (Richler et al., 2017), potentially due to a greater role for the domain-general ability in tests of novel objects. Given these results, we can be confident that for both theoretical and empirical reasons, using novel objects to estimate *o* may be better than using familiar objects, or no different, but should not provide worse estimates. In prior work using the aggregate approach to estimate *o*, novel object tasks have been used to avoid any contribution from experience with familiar objects. A large and varied set of familiar categories may converge on an unbiased estimate of domain-general ability (Richler et al., 2017), but measurement with novel objects may provide a more direct way to avoid this potential source of bias.

The high correlation between aggregate estimates of *o* and factor scores from a model with 6 tasks validates the aggregate approach that has been used in prior research. There is no question that there are some advantages to an SEM approach, including the ability to calculate correlations among variables, free from the attenuating effects of measurement error. An alternative approach to handling measurement error when correlating aggregate scores with other variables is to disattenuate correlations based on the reliability of the measures (Nunnally, 1994; Wang & Stanley, 1970). The ability to estimate *o* quickly makes possible the use of this construct in a broader set of research circumstances (e.g., when time is limited, often because many constructs must be estimated, or when limits in sample size will not allow an SEM approach).

Table 1

Correlations between aggregate scores and factor score estimates

Matching Category	Learning Exemplar Category	$o_{Familiar}$	o_{Novel}
v-Ziggerins	s-Greebles	.85	
v-Ziggerins	Sheinbugs	.71	
s-Greebles	v-Ziggerins	.87	
s-Greebles	Sheinbugs	.76	
Sheinbugs	v-Ziggerins	.84	
Sheinbugs	s-Greebles	.85	
Birds	Planes		.77
Birds	Transformers		.80
Planes	Birds		.70
Planes	Transformers		.83
Transformers	Birds		.63
Transformers	Planes		.73

Note. Aggregates of novel object tests are compared with the familiar object factor scores, whereas aggregates of familiar object tests are compared with the novel object factor scores.

Part II – A Trio of Measures

In Part I we demonstrated that it is possible to obtain good estimates of *o* by aggregating across only two tests with differing task demands and stimuli. In Part II, we further refined this approach by revising our existing tests to measure *o* more efficiently and adding an additional test to measure *o* more comprehensively. Then we assessed the validity and reliability of these measures.

Our motivations for developing a new set of tests are multiple. The matching task used in Sunday et al. (2022) had 360 trials, and is thus impractical for many research projects. Shorter measures of *o* will make measurement more efficient and allow the inclusion of the construct in research where time and resources are limited. Furthermore, the matching task was not initially designed for individual differences work, and so trials were not previously selected on the basis of

item analyses, nor was trial difficulty manipulated to ensure a good range (high reliability was instead achieved because of a large number of trials). Additionally, because the matching task is an old/new task, the decision to respond 'old' on a given trial depends on whether the memory strength for the object is stronger than a decision threshold that varies across participants (some participants need to be very sure before they say they recognize an object, others do not). To account for the influence of response bias, a measure from signal detection theory known as *d'* is used to index accuracy (Stanislaw & Todorov, 1999). However, in attempting to correct for response bias, the calculation of *d*' as commonly performed relies on unrealistic assumptions about the distribution of memory strengths that may lead to incorrect conclusions. It is assumed that memory strengths for old and new items follow equivalent normal distributions, albeit with different means. Brady et al. (2022) demonstrate that the violation of this assumption can lead to cases where an individual who has better memory than another person will have a *d*' suggestive of worse memory than that person. These issues are especially concerning when *d*' is used to measure individual differences. Therefore, it is recommended to use a forced choice design, because instead of measuring whether a person's memory strength for an object is above or below a particular threshold which is unique to them, an individual can compare memory strengths for each item on a trial and choose the item with the strongest one.

Another motivation for revising measurement methods is to further refine the construct of *o*. As *o* is a newly identified construct, there is still substantial uncertainty about the specific dimensions that must be measured to capture the construct well. Construct theories are transient and evolve as measurement helps us to better align theoretical predictions and observations (Kuhn, 1961; Stenner et al., 2022). When selecting indicators for a construct, it is necessary to consider whether the dimensions of importance have been well characterized by prior work. Constructs which have been well characterized may be confidently measured by a small number of indicators that capture the known dimensions of relevance. However, for novel constructs there is great uncertainty about which specific dimensions are of importance. Only through an iterative process

can researchers arrive at greater certainty about how to best capture a latent construct. Simulations suggest that under such conditions of uncertainty, it is beneficial to use a greater number of diverse indicators, which capture multiple dimensions of potential importance (Little et al., 1999). By adding another test to our set of measures that varies significantly from our existing tests, we may be better able to triangulate *o*. Our existing two tests (MA and LE) both allow for within-test learning of categories, as participants gain experience with objects from the same category across trials. It is currently uncertain how important this learning element is to the *o* construct, so we designed an additional test – the Many Objects Oddball test (MOO), which presents a new object category on each trial and does not allow for learning. The shared variance between this test and the existing tests should tap into an ability that applies regardless of whether learning is possible across trials.

Our proposed trio of measures therefore consists of the LE test used in Part I, a revised version of the MA test used in part I, and the new MOO test. We describe the development and features of these tasks in the task development section. To test the psychometric properties of our new trio of tests, we administered them to a large sample on two occasions spaced a month apart. Using this design, not only can we assess the convergent validity of our measures, but we can also calculate test-retest reliability for each measure, and for the overall aggregate. While *o* may be assumed to be a stable trait, no prior work has directly assessed the degree of stability, which could have implications for any potential applications of the construct in real world contexts. This study thus represents the first longitudinal investigation of *o*. Additionally, we tested participants on other cognitive and perceptual abilities, including perceptual speed, intelligence, and low-level visual perception. This allows us to investigate the relationship between *o* and these other abilities and determine whether the longitudinal stability of performance on *o* tasks can be explained by these other abilities. We deliberately chose measures of other perceptual abilities which that most similar in task design to our *o* tests so that we would have the best chance of accounting for *o* test variance with these measures. Many psychological constructs are merely rehashed versions of existing

constructs (Kelley, 1927), and it is therefore useful to test whether *o* can be accounted for by theoretically similar constructs and existing task designs.

Participants

We recruited 200 participants online from Prolific to complete two sessions, spaced 30 days apart. Based on previous work (e.g. Richler et al., 2019; Sunday et al., 2022), we expected that correlations between our object-recognition tests would range from .3 to.4. Such correlations should, with 80% confidence, reach a critical point of stability with a half-width of .1 at around 200 participants (Schönbrodt & Perugini, 2013). We restricted recruitment to participants whose first language was English and had an approval rate above 95% on the site. To ensure high-quality data, we embedded attention checks into most of our tests that did not have strict time limits (power tests, in contrast to speed tests). These attention checks were simple instructions intended to ensure that participants were attentive to the task. For the first session, if a participant failed two or more attention checks embedded throughout our tasks, we replaced them. In this first session, we collected 200 complete datasets (mean age = 41.2 years, SD = 13.9; 103 women, 92 men, 5 other). In the second session, we also did not use data from participants who failed two or more attention checks. Of the initial 200 participants, 163 participants (mean age = 41.4 years, SD = 13.6; 87 women, 72 men, 4 other) successfully completed the second session. For analyses concerning *o* tests only, all 163 participants who completed both sessions were included. For analyses involving additional variables, four participants were excluded for scores that were more than three standard deviations below the mean or scored negatively on perceptual speed tasks.

Task Development

Three-Alternate Forced Choice Matching Task

We designed a new version of the matching task, switching to a three-alternative forced choice design. There are several benefits to this format relative to the same/different format used in prior work. First, changing the task from one which asks participants if an object is one they have

previously seen to a task where they have to determine which one of a set of objects is one they have previously seen relieves us from making strong assumptions about unobservable memory strengths that are necessary to calculate a measure of discriminability like *d*'. Recent best practices recommendations suggest that a forced-choice format should be used whenever possible (Brady et al., 2022). In addition, using a 3AFC format allows us to reduce the probability of being correct if guessing at random to one third from one half, which reduces measurement error and allows achievement of higher reliability using fewer trials. Finally, a 3AFC format provides us with more degrees of freedom to manipulate the difficulty of the discrimination. A large variability in the difficulty of the trials is helpful, as easy trials can help discriminate individuals at the lowest end of the ability, while harder trials help discriminate individuals at the highest end of the ability. These changes allowed us to reduce the number of trials from 360 in the old format to 51 in the new format, raising reliability in a shorter test. It can also be scored more easily using percent accuracy.

On each trial, participants were presented with an object briefly, then asked to select the matching object amongst an array of three objects. This test included 51 trials using asymmetrical Greebles, rendered in two views approximately 30 degrees apart, rotated around the vertical axis. The first three trials have correct/incorrect feedback, and feedback is not included in later trials. Following practice there were four blocks of trials increasing in difficulty. Target objects were repeated only once during the entire test, in a different view each time. Target objects were never used as foils on other trials. All trials began with the fixation cross for 500 ms, followed by the presentation of the target object for 300-1000 ms (dependent on trial difficulty). This was followed by a 500 ms mask of scrambled object parts with visual noise, and an array of three objects. The trial ended when the participant made a response by clicking on one of the objects, indicating the match. The first block of 8 trials showed the target for 300 ms and the options in the same view as the target. The second block of 16 trials showed the target for 500 ms and the options in a different view from the target. Participants were then informed that the targets would appear in visual noise with an example. The noise was Perlin noise, which is a pseudo-pattern gradient noise used by visual

effects artists to increase realism in computer graphics. A two-tone pseudo-pattern appeared behind the Greebles and with 80% transparency over them. The third block of 12 trials showed the target in noise and presented for 750 ms, followed by the options without noise in the same view as the target. The fourth block of 12 trials showed the target in noise for 1000 ms, followed by the options without noise in a different view as the target. Percent accuracy was used as a measure of performance.

Many Objects Oddball

We also developed a new object recognition test – the Many Objects Oddball Test (MOO). In this test, participants pick the odd one out from three objects from the same category. These objects are complex, varying along multiple dimensions, but crucially, because no object category is repeated, there is no possibility for within-test learning about the dimensions that are relevant for any given category. This is unlike in the MA and the LE in Sunday et al. (2022), where object categories are repeated across trials. In some of our work we have used an MA task that mixed objects from five novel categories (Sunday et al., 2018), but even these tasks allowed learning at the category level as several objects within any one category were presented over the course of 200 trials. For both the revised MA task and the MOO task, we initially created a set of trials intended to capture a range of difficulty. We collected data from independent samples and then calculated trial difficulty, item-rest correlation, and internal reliability. For each task, we reordered trials based on empirical difficulty and then replaced trials with low item-rest correlation. After these modifications, we collected data again from another sample and continued to improve the tasks. We repeated this data collection from independent samples followed by task modifications until we achieved high internal reliability.

On each trial, participants were presented with three objects from the same object category and asked to choose which object was different from the other two. This test included 45 trials using

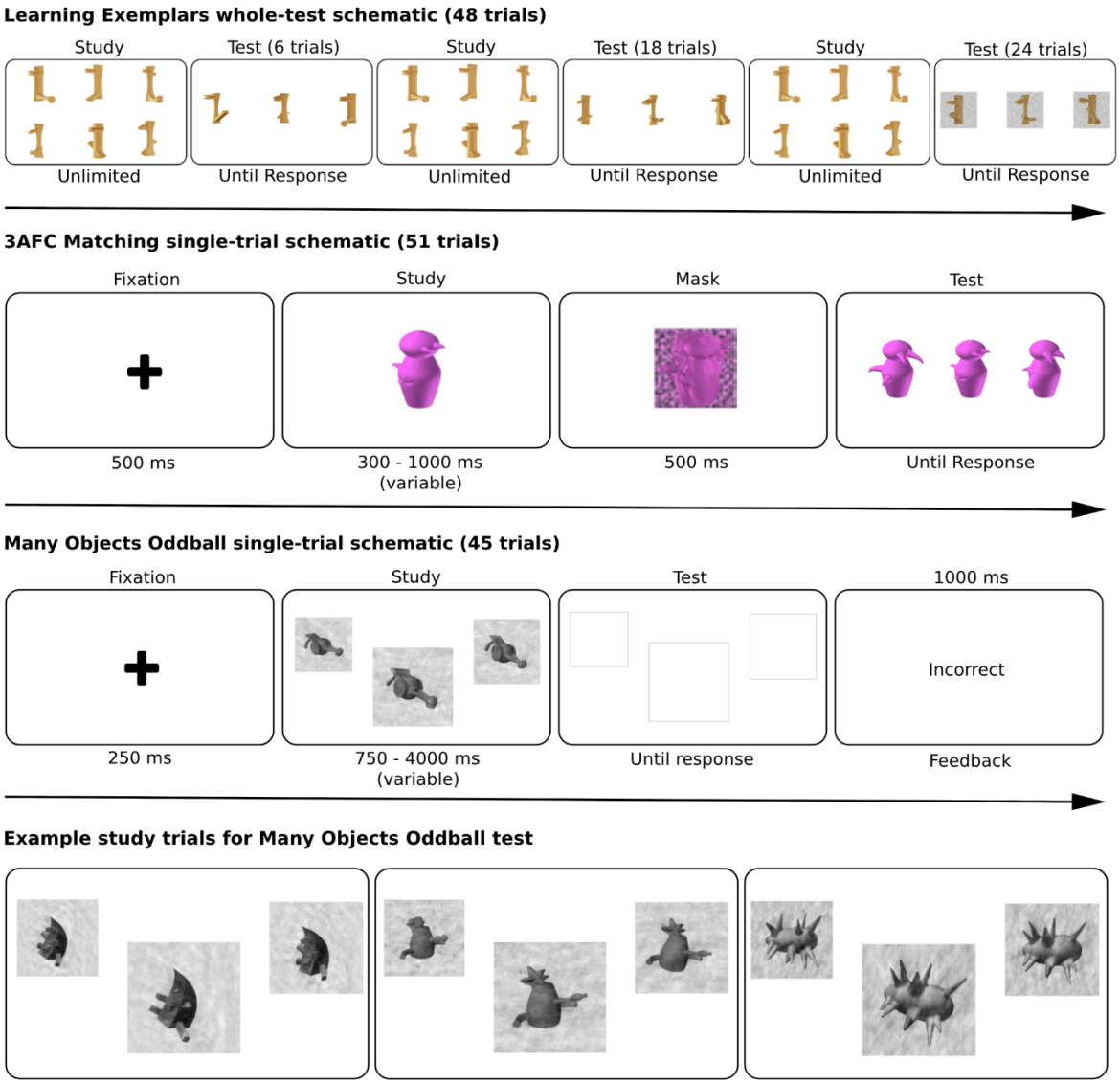
objects from many novel object categories intermixed throughout the task – specifically, while objects were highly similar within a trial, they were from highly different categories across trials. The objects came from a set collected by McGugin et al. (2022) from various sources. Each trial began with the presentation of a fixation cross in the center of the screen for 250 ms, followed by the presentation of three objects for 750-4000 ms to vary difficulty. This was followed by a prompt for a response with three empty squares where the objects had previously appeared. The trial would only end when the participant made a response by clicking on one of the squares, indicating the object that appeared there was the oddball. During the presentation of the objects, each image was a slightly different size and vertically offset such that it was difficult to determine the oddball using low-level feature comparisons. The two images showing the same object showed it in a slightly different view. All images were transformed to greyscale and equated for low-level image properties using the Matlab SHINE toolbox (Willenbockel et al., 2010). In addition, based on pilot data, Perlin noise was added to some trials to increase difficulty. Feedback was given on every trial, but due to the wide variety of object categories used in this task, we assumed very little learning in this test. Percent accuracy was used as a measure of performance.

Test Battery

o tests

Our trio of *o* tests consisted of one of the previously used LE tests (with vertical Ziggerins; details described in Part I), the 3AFC version of the MA test with asymmetrical Greebles, and the newly developed MOO test (Figure 2). Aggregate *o* scores were calculated for each participant by averaging z-scores from each test. We implemented these tests using jsPsych (version 7.2; de Leeuw, 2015) and we provide these tests in an easy-to-use offline format that can be adapted for online use in our GitHub repository (https://github.com/OPLabVanderbilt/Ojs/tree/main/standalone).

Figure 2. Trio of object recognition tasks used to measure *o* scores.



Note. The object category differs for each trial of the MOO test.

Matrix Matching Intelligence Test

The Matrix Matching Test (Pluck, 2019) asks participants to select the missing piece of a visuo-spatial pattern, or to select the item that best matches the semantic meaning of a given set of images. Of the total 24 trials, 12 trials measured visuospatial intelligence, and 12 measured semantic reasoning. Previous work suggests this task correlates highly with other established intelligence tests such as the Weschler Adult Intelligence Scale-IV (r = .89; Pluck, 2019).

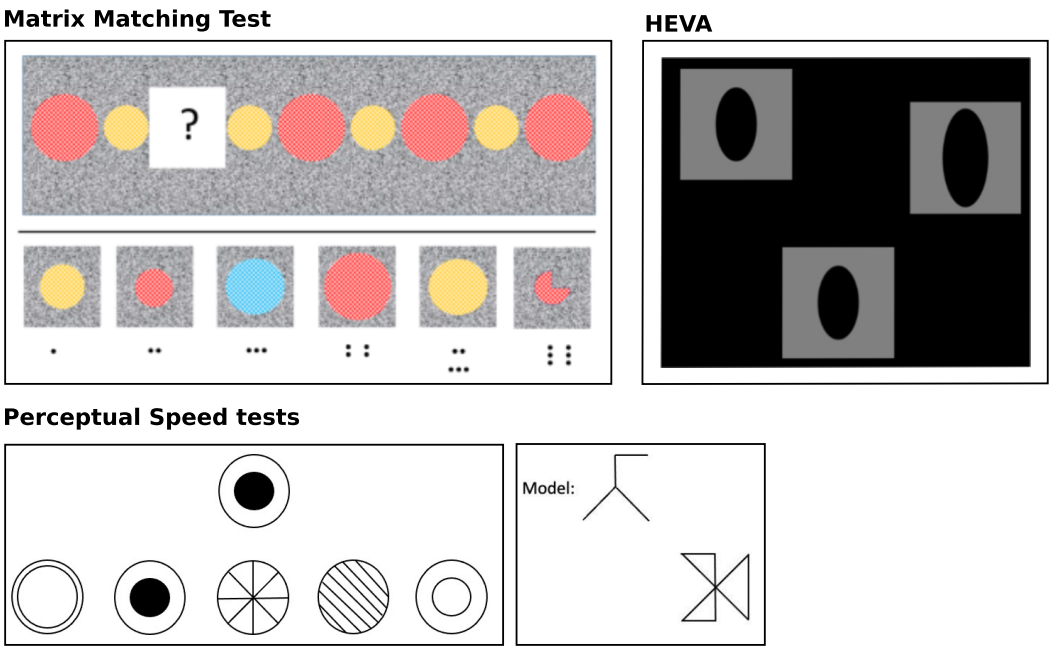
Perceptual Speed

Perceptual speed was measured using a pair of speeded tests (Ekstrom et al., 1976): Hidden Patterns, in which participants had to verify whether one line drawing was contained within another; and Identical Picture matching, in which participants saw a picture and had to select the exact matching picture amongst foils. In both tests participants complete as many trials as possible and incorrect trials are subtracted from the total. Participants were given three minutes for the Hidden Patterns test, and one and a half minutes for the Identical Pictures test. We aggregated z-scores on these tests to create an overall perceptual speed score.

Hanover Early Vision Assessment

The Hanover Early Vision Assessment (HEVA; Kieseler et al., 2022) measures low-level visual ability, where on each trial participants had to select the oddball amongst three images based on a low-level feature such as line length or orientation. There were 120 trials in total made up of 20 blocks of six trials each. All six trials within a block tested comparisons of the same feature type. The five feature comparisons were dot distance, circle size, ellipsoid size, angle size, and line length.

Figure 3. Example trials for tests of abilities other than object recognition



Note. In the Matrix Matching Test the goal is to select the option that fits the pattern. In the Hanover Early Vision Assessment participants pick out the odd shape. There are two perceptual speed tests, in order from left to right: Identical Pictures, and Hidden Patterns. In the Identical Pictures task, participants pick the object which matches the target object, in the Hidden Patterns task participants judge whether the model is contained in the other line drawing.

Procedure

The study was split across two sessions and was completed entirely online. In the first session participants completed the LE test, the 3AFC MA test, the MOO test, the perceptual speed tests, the Matrix Matching test, and then HEVA. This first session took approximately 45 minutes. Participants were invited to the second session 30 days after they completed the first session and were given seven days to complete it. This second session only included the LE test, the 3AFC MA test, and the MOO test. The trio of *o* tasks took approximately 15 minutes.

Results

Reliability & Validity

The three *o* tests all had significant positive correlations with one another at both sessions (see Table 2 for correlations between tests). We calculated single session λ_2 reliability estimates for the *o* tests, which suggested good reliability (see Table 3 for descriptive statistics and reliability). However, the single-session reliability of aggregate *o* scores was higher still, as is often the case when measures correlate (Lord & Novick, 1968). The test-retest reliability of the three *o* tests were all acceptable, though somewhat lower than single-session reliability. However, the test-retest reliability of the aggregate measure of *o* was considerably higher than that of the individual tests, demonstrating that the aggregate measure of *o* is a reliable measure of individual differences even between sessions. As correlation estimates are limited by the reliabilities of the variables (Lord & Novick, 1968), by disattenuating for measurement error we can obtain estimates of the stability of the underlying construct of interest (Spearman, 1907). The disattenuated correlation (r = .89) suggests that the *o* construct is highly stable over a month.

Table 2.
Correlations between tasks

Task	Age	MOO	LE	3AFC MA	Identical Pictures	Hidden Patterns	Matrix Intelligence
MOO	-.26**						
LE	.10	.18*					
3AFC MA	-.20*	.44***	.35***				
Identical Pictures	-.37***	.52***	.19*	.46***			
Hidden Patterns	-.22*	.32***	.27***	.44***	.56***		
Intelligence	-.02	.29***	.5**	.4***	.28***	.42***	
HEVA	-.01	.24**	.37***	.4***	.19*	.31***	.45***

Note. *p < .05; ** p <.01; *** p <.001, using the Holm-Bonferroni method of controlling family-wise error rate (Holm, 1979).

Table 3

Descriptive statistics

	Session 1		Session 2		
Test	***Mean* (*SD*)**	**Reliability**	**Mean (*SD*)**	**Reliability**	**Test-retest reliability**
Aggregate *o*	0.07 (0.73)	.85	0 (0.76)	.86	.77 [.70, .83]
Learning Exemplars	58.5% (15.8%)	.85	60.8% (17.6%)	.88	.62 [.52, .71]
3AFC Matching	69.1% (11%)	.73	71.5% (11.1%)	.72	.66 [.56, .74]
Many Objects Oddball	71% (10.6%)	.66	71.8% (11%)	.68	.66 [.56, .74]
Perceptual Speed	0.05 (1.29)	.94			
Identical Pictures	47.7 (8.3)	.86			
Hidden Patterns	97.5 (37.2)	.96			
Matrix Intelligence	72% (11.7%)	.69			
Hanover Early Vision Assessment	74.8% (11.3%)	.92			

Note. λ_2 reliability was calculated for Learning Exemplars, 3AFC Matching, Many Objects Oddball, and Hanover Early Vision Assessment, as it is more robust than Cronbach's α (Callender & Osburn, 1979). Aggregate *o* reliability was calculated with equal weighting for each subtest following Wang and Stanley (1970). For the Matrix Matching intelligence test, Lambda 2 was first calculated separately for visuospatial trials and semantic trials, and then aggregate reliability was calculated for the final score. As the perceptual speed tests are speeded, the split-half reliabilities for the identical pictures and hidden patterns tests were calculated as the Spearman-Brown prophecy corrected correlation between the scores from the first half of allotted time and the scores from the second half of allotted time. These estimates were then used to calculate the aggregate perceptual speed reliability. 95% CIs are reported for test-retest reliability estimates.

The high reliability of aggregate *o* both within-session and across-sessions suggests our measures are well suited for individual differences research. However, it is possible that the stability of our measures may be explained by other abilities these aggregate scores may capture. Table 4 shows that all of the abilities we measured correlated significantly with one another, suggesting that there is substantial shared variance between them. To test whether our measures capture a stable object recognition ability that is independent of these related cognitive abilities, we conducted a hierarchical regression (Table 5) predicting aggregate *o* at session two. In the first step we added age as a predictor variable, as age is a subject factor that could impact many cognitive abilities. In the second step we added perceptual speed, intelligence, and early visual abilities as predictor variables. Finally, to determine the extent to which our object recognition measures capture *o* independently of these other tests, we added aggregate *o* from session one in the final step. Age did not significantly predict *o* at session two**.** A large portion of *o* score variance was predicted by the addition of intelligence, perceptual speed, and early visual abilities in step 2 ($\Delta R^2 =$ .41), which all independently predicted *o*. In the final step, the addition of aggregate *o* at session one significantly

increased the total variance explained in aggregate *o* at session two (ΔR^2 = .17). Thus, the relationship between aggregate *o* at session one and session two cannot be completely explained by perceptual speed, intelligence, and low-level visual abilities. In the final model early visual abilities and perceptual speed were significant unique predictors of *o*, whereas intelligence no longer uniquely predicted *o*. Interestingly, this is true even though we used measures of intelligence and processing speed that included objects and shapes (i.e. we used measures of perceptual speed, which is a facet of processing speed concerned with visual stimuli) which may themselves include a contribution from *o.* One explanation for this lack of a unique relationship may be that general intelligence contributes to all cognitive ability tests so this source of variance in *o* tests may also be captured by the other perceptual tests.

Table 4
Correlations between abilities

Task	Age	*o*	Perceptual Speed	Matrix Intelligence
o	-.16 [-.29, -.02]			
Perceptual Speed	-.36*** [-.47, -.23]	.56*** [.45, .65]		
Matrix Intelligence	-.02 [-.16, .12]	.54*** [.43, .63]	.36*** [.23, .48]	
HEVA	-.02 [-.15, .13]	.45*** [.33, .56]	.26** [.12, .38]	.45*** [.33, .55]

Note. *p < .05; ** p <.01; *** p <.001, using the Holm-Bonferroni method of controlling family-wise error rate (Holm, 1979). 95% CIs are displayed.

Table 5
Hierarchical Regression

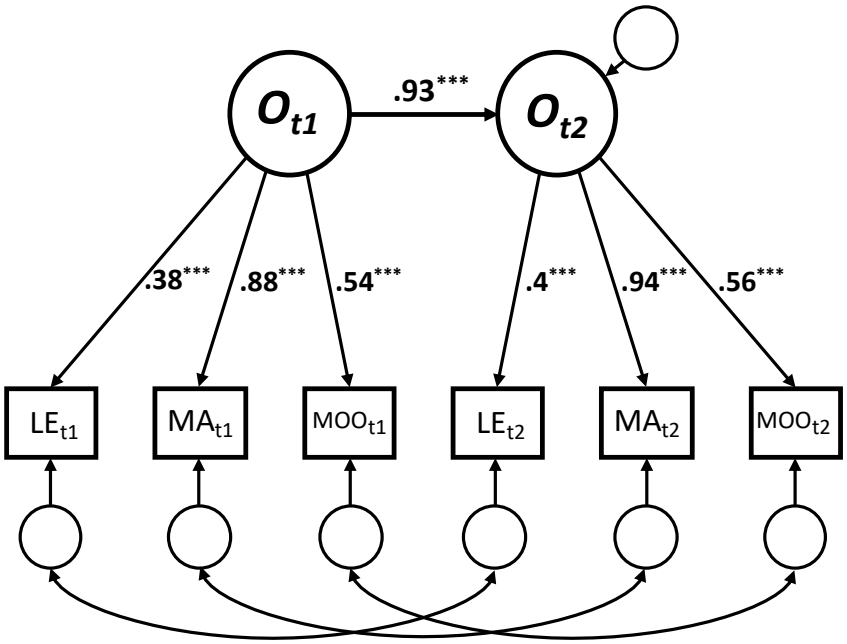
Predictor	β	*t*	*F*	adj R^2	ΔR^2
Step 1 (age)			3.213	.01	
Age	-0.14	-1.837			
Step 2 (other abilities)			29.19	.42***	.41***
Age	-0.02	-0.362			
Intelligence	0.19	2.653**			
Perceptual Speed	0.13	5.234***			
HEVA	0.32	4.769***			
Step 3 (*o* at time 1)			46.91	.59***	.17***
Age	0	-0.055			
Intelligence	0.03	0.534			
Perceptual Speed	0.13	2.07*			
HEVA	0.21	3.603***			
o Session 1	.57	8.212***			

Note. *p < .05; ** p <.01; *** p <.001. Standardized beta weights and *t* values are for simultaneous regression, change in R^2 is the change between each step.

Structural Equation Models – Convergent Validity and Stability

To estimate the stability of *o* over the course of a month, we modelled *o* as a latent variable in a structural equation model. This approach is advantageous as it accounts for measurement error, in addition to more effectively removing the influence of irrelevant task-specific variance (Bollen & Bauldry, 2011). Our model indicated a very high stability coefficient (γ = 1.08, 95% CI [.68, 1.47]; r = .93, 95% CI [.85, 1]), demonstrating that *o* is almost perfectly correlated across sessions. In both sessions, all three tests had highly significant path coefficients from *o,* suggesting the measures have convergent validity. These path coefficients were particularly high for the 3AFC Matching Task (See Figure 3 for standardized coefficients). We also tested for metric measurement invariance across the two sessions and found no significant reduction in model fit, but we did find a significant reduction in fit for a scalar invariant model. (See Supplementary Materials).

Figure 3. Longitudinal Structural Equation Model of *o* at time 1 and time 2.



Note. Standardized solution. Significant paths have coefficients in Bold, *** indicates $p < .001$**.** Fit statistics: χ^2 (5) = 1.4, p = .92; RMSEA = 0, 90% CI [0, .035]; NNFI = 1.03; SRMR = .009.

Discussion

The aggregate *o* measure was highly reliable for both single-session and between-session measurement, including across a month-long time interval (r = .77). Classical Test Theory assumes that correlations between observed scores for different tests occur due to correlations between the true score portion of the variance of each test, and not the error variance. If multiple tests measure the same construct, their true-score variances will correlate, increasing the reliability of their aggregate. Because we aim to measure domain-general ability, not all true-score variance in an object recognition test is relevant to the construct, as some true-score variance is due to domain-specific factors, such as stimuli or task demands. An inherent limitation of the aggregate approach is that it is does not separate domain-general and domain-specific sources of true-score variance, so some of the increased reliability of the aggregate measure may result from irrelevant true-score variance that is shared between only some of the constituent tasks (for a further explanation see Gerbing & Anderson, 1984). Because composite measures like aggregate *o* assume that a linear combination of scores define the construct, they partially consist of measurement error and irrelevant domain-specific variance. These unwanted sources of variance can contaminate aggregate measures and may bias estimated relationships between the composite and other variables (Bollen & Lennox, 1991) as these irrelevant sources of variance may themselves covary, or fail to covary, with other constructs. However, these issues may be mitigated by aggregation of measures with diverse task demands. With diverse measures, construct-irrelevant sources of variance should be unique to individual tests, and domain-general ability should predominate over other sources of variance, as domain-general ability contributes to performance on every test. The resulting aggregate should therefore largely reflect domain-general ability as opposed to measurement error and task-specific abilities (See Lubinski, 2004, p.99).

Although an aggregate measure of *o* can better capture the domain-general construct than a single-format test, *o* is properly conceived of as a latent variable which causes performance on the

object-recognition tests, rather than a variable composed from test scores. This is synonymous with the distinction between effect and cause indicators as has been discussed in the structural equation modelling literature (Bollen & Bauldry, 2011). To obtain a purer measure of *o*, our trio of measures can also be used as indicators of *o* modelled as a latent variable, thus removing measurement error and further limiting the influence of task-specific variance. This difference between approaches explains the difference in stability estimates between *o* modelled as a latent variable (r = .93) and *o* measured as an aggregate (r = .77; disattenuated r = .89). However, the aggregate approach is a useful compromise between the analysis of single measure variables, which can be more biased, and the use of structural equation models, which require large samples and often more complex research designs.

Whenever a new construct is proposed, it is worth establishing its independence of, or dependence on, other theoretically related constructs. The location of the construct within the nomological net allows for a fuller understanding of the identified construct, and for the validation of new measures (Cronbach & Meehl, 1955). In the case of *o*, we are still beginning to understand its place among its peers. The construct is likely to be related to measures of more specific aspects of object recognition, for instance visual memory (Brady et al., 2011) or object matching (Growns et al., 2023). Critically, both the latent variable and aggregate approaches we compared here target a more general ability, as the individual tests vary on both dimensions of object category and task demands. Previous work has demonstrated that *o* is related to, but largely independent of, fluid IQ and visual working memory (Richler et al., 2017, 2019; Sunday et al., 2018). We find moderate to strong relationships between aggregate *o,* perceptual speed, intelligence, and early visual abilities. The interpretation of these relationships based solely on first-order correlations can be difficult, as much of the variance may overlap between multiple variables. Using hierarchical regression with aggregate *o* at session two as the dependent variable, and all other variables as predictors, we found that a considerable proportion of aggregate *o* could be explained by intelligence, perceptual speed, and early visual abilities (R^2 = .41). But crucially, even after partialling the other abilities out, there

was still a substantial unique relationship between *o* at session one and *o* at session two (R^2 = .17), which demonstrates that the aggregate *o* measure captures something distinct from intelligence, perceptual speed, and early visual ability. Furthermore, as we are not modelling *o* as a latent variable in this analysis, the unique relationship between each session's aggregate *o* is likely an underestimate, as it is constrained by measurement error (Charles, 2005).

Although the variance in aggregate *o* at session two uniquely explained by aggregate *o* at session one was smaller than the total variance explained by other abilities, it is worth considering the similarity of the tests used in measuring these other abilities to those used in measuring *o*. For example, HEVA requires that participants detect the odd one out of three images of low-level features. These task demands are very similar to those in the MOO test, except that the HEVA test uses low-level features such as line length, rather than high-level complex objects. The Matrix Matching Intelligence test is a visual measure of intelligence, rather than verbal. Perceptual speed is the facet of the more general processing speed that is explicitly concerned with the rapid search for, and comparison of, specifically visual stimuli (Flanagan & Dixon, 2014). The Identical Pictures perceptual speed test is a matching task requiring participants to pick which of three images is the same as the target image, which is a very similar format to the 3AFC MA test. The Hidden Patterns perceptual speed test also requires matching, as it asks participants to judge target line drawings that may be contained within another line-drawing. Given the similarities in task demands and the heavy visual component of all tasks, the existence of a sizeable unique relationship between aggregate *o* at session one and session two suggests that there may be a substantial degree of independence between *o* and these other abilities. Our approach was to deliberately pick those aspects of other constructs which were most theoretically close to *o*, and those tests of other constructs which shared the most task demands with our tests, in order that there would be the best chance possible of accounting for the variance in *o* tests with these tests of other abilities.

A limitation of this approach is that it is not possible to separate shared method variance from shared construct variance when analyzing correlations between these abilities. Future research should use more balanced sets of measures of other constructs in order to precisely estimate the degree of relationship between *o* and other constructs through the use of SEM methods. For example, the aggregate measure of perceptual speed correlated highly with aggregate *o* (r = .56), but it is unclear whether this is due to an inherently strong relationship between the two constructs, or because the tests for both abilities have similar task demands. Indeed, for the LE, which is the most dissimilar in task demands from the perceptual speed tests, there is a much lower relationship with perceptual speed (r = .24) than there is for the 3AFC MA (r = .5) and the MOO (r = .5). The perceptual speed tests require rapid visual comparison of simultaneously presented stimuli, and the 3AFC MA and the MOO both have limits on encoding time and have short time intervals between encoding and retrieval. However, the LE has no limit on encoding time, and has longer time intervals between encoding and retrieval. By using a more diverse set of tests as indicators for these other abilities, we could use an SEM approach to obtain a clearer picture of how closely the underlying abilities are related when accounting for the shared variance due to similarity of task demands. It is likely that there is a substantial relationship between *o* and these other abilities at a construct level, even if shared method variance were accounted for. This might be expected if *o* is theoretically conceived of as existing within a hierarchical model of cognitive abilities such as that described by the Cattell-Horn-Carrol (CHC) model (Schneider & McGrew, 2018). In such a model, all abilities are expected to correlate due to the shared influence of general intelligence, which occupies the top of the hierarchy, but abilities are also expected to correlate more highly with one another if they are theoretically related due to the shared influence of some broad ability. For example, if *o* were conceived of as being a narrow facet of a broader visual processing ability, it would be expected to correlate with other visual abilities to a greater degree than with non-visual abilities, as the other visual abilities would also be strongly influenced by the broad visual processing ability. The position that *o* should occupy in a model like CHC is currently unclear, but we expect strong correlations

between *o*, which fundamentally concerns individuation, and abilities that require perceptual discrimination – especially of complex stimuli. Perceptual speed requires rapid visual comparison, but perceptual speed tests require simpler judgements than *o* tests do, and scores on perceptual speed tests depend on the rapidity of these simple judgements rather than on the accuracy of difficult judgements. Because both constructs describe an ability concerning perceptual comparison, we would expect some overlap between these abilities, despite the theoretical distinction. Likewise, we would expect some overlap between *o* and the ability to make similarity judgements for low-level stimulus properties that is tested by HEVA. However, our analysis suggests that these related abilities do not explain all of the variance in the *o* ability. Recent evidence suggests that the relationship between visual *o* and the ability to individuate complex auditory stimuli is higher than the relationship between *o* and visuospatial abilities (Smithson et al., 2024; see also: Chow et al., 2023). Given these results, *o* seems unlikely to rely primarily on a more general visual ability.

We find a large relationship between the aggregate *o* measure and general intelligence (r = .54). In part, this reflects an already known relationship between *o* tasks and fluid-IQ; previous research has indicated relationships of small to moderate size (Richler et al., 2017, 2019). However, this relationship may also reflect the nature of aggregate measures, in that the 'general factor' can be expected to contribute to performance on all ability tests, and so aggregates of different tests will contain a component of variance that can be attributed to general intelligence. Of course, this is not unique to aggregate measures, but while aggregate measurement can minimize the influence of domain-specific variance, domain-general influences will remain influential on scores. By implication, the aggregate *o* measure will, at least in part, correlate with other abilities for the simple reason that it correlates substantially with intelligence. For example, the correlation between aggregate *o* and perceptual speed (r = .56 to $r_{partial}$ = .46) and HEVA (r = .43 to $r_{partial}$ = .26) are both reduced when we partial out scores on the matrix matching intelligence test. This underscores the importance of controlling general intelligence when using aggregate measures to predict other variables, if the hypothesis concerns the predictive ability of *o* specifically and not general intelligence. In previous

work using aggregate *o* measures, general intelligence has been partialed out when predicting criterion variables (e.g. Chang & Gauthier, 2021; Sunday et al., 2018). In addition, intelligence can also be measured using tests that do not require shape processing, and this may alter the size of the relationship with *o*.

As the measurement of *o* is still in its infancy, there are many unanswered questions about the best task designs for measuring it. All path coefficients from the latent variable *o* to the three object recognition tests were highly significant, but varied greatly in size, in particular the 3AFC MA predominated over the other two tests. However, in recent large-sample research using these measures, the loadings onto *o* from these three tasks has been more balanced (Chow et al., 2023; Smithson et al., 2024). For a well-established construct with a long history of psychometric research, each measure can be well designed to reflect the latent construct to a high degree. However, for novel constructs, it may be favourable to choose indicators that differ to a greater extent, such that although the path coefficients are lower, the location of the construct in multivariate space may be more accurately triangulated (Little et al., 1999). Only through an iterative process of adjusting, testing, and seeing what breaks measurement, are we likely to develop a clearer picture of which dimensions are of importance for capturing the construct more accurately with each measure. Furthermore, when measuring a domain-general ability, it may be particularly important that indicators are not too similar, as covariance between three quite different tasks is more likely to represent an ability which is truly domain-general than covariance between three very similar tasks which covary very strongly. It is possible that in such a situation, if each test is not designed to give full coverage of a construct, individual measures of different constructs could correlate more highly than measures of the same construct if task demands were shared between the measures of different constructs, but not shared between the measures of the same construct, especially if the relationship between constructs is non-zero (Bollen & Lennox, 1991). Future work could also aim to develop additional measures of *o* such that the aspects of task design that are crucial for measurement can be determined. By systematically varying features of object recognition tests and

comparing how different variants load onto the latent *o* variable, it may be possible to develop a clearer picture about what is necessary and what is not necessary to capture *o*. The development of this trio of *o* tasks represents an important step forward in the measurement of domain-general object recognition ability, and we provide a valid, reliable, and quick measure of general object recognition for all researchers to use. We consider this a good starting point for future task development and for the exploration of the *o* construct. By making these measures publicly available, we hope that other researchers will be able to include *o* in their research designs.

Open Practices Statement

Data are available at https://osf.io/m9xa5/

Tests are available at https://github.com/OPLabVanderbilt/Ojs/tree/main/standalone

References

Bollen, K. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118619179

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods*, *16*(3), 265–284. https://doi.org/10.1037/a0024448

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*(2), 305–314. https://doi.org/10.1037/0033-2909.110.2.305

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of Vision*, *11*(5), 4–4. https://doi.org/10.1167/11.5.4

Brady, T. F., Robinson, M. M., Williams, J. R., & Wixted, J. T. (2022). Measuring memory is harder than you think: How to avoid problematic measurement practices in memory research. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-022-02179-w

Callender, J. C., & Osburn, H. G. (1979). An empirical comparison of coefficient alpha, Guttman's Lambda—2, and msplit maximized split-half reliability estimates. *Journal of Educational Measurement*, *16*(2), 89–99. https://doi.org/10.1111/j.1745-3984.1979.tb00090.x

Chang, T.-Y., & Gauthier, I. (2021). Domain-specific and domain-general contributions to reading musical notation. *Attention, Perception, & Psychophysics*, *83*(7), 2983–2994. https://doi.org/10.3758/s13414-021-02349-3

Chang, T.-Y., & Gauthier, I. (2022). Domain-general ability underlies complex object ensemble processing. *Journal of Experimental Psychology: General*, *151*(4), 966–972. https://doi.org/10.1037/xge0001110

Charles, E. P. (2005). The Correction for Attenuation Due to Measurement Error: Clarifying Concepts and Creating Confidence Sets. *Psychological Methods*, *10*(2), 206–226. https://doi.org/10.1037/1082-989X.10.2.206

Cho, S.-J., Wilmer, J., Herzmann, G., McGugin, R. W., Fiset, D., Van Gulick, A. E., Ryan, K. F., & Gauthier, I. (2015). Item response theory analyses of the Cambridge Face Memory Test (CFMT). *Psychological Assessment*, *27*(2), 552–566. https://doi.org/10.1037/pas0000068

Chow, J. K., Palmeri, T. J., & Gauthier, I. (2022a). Haptic object recognition based on shape relates to visual object recognition ability. *Psychological Research*, *86*(4), 1262–1273. https://doi.org/10.1007/s00426-021-01560-z

Chow, J. K., Palmeri, T. J., & Gauthier, I. (2022b). Visual object recognition ability is not related to experience with visual arts. *Journal of Vision*, *22*(7), 1. https://doi.org/10.1167/jov.22.7.1

Chow, J. K., Palmeri, T. J., Pluck, G., & Gauthier, I. (2023). Evidence for an amodal domain-general object recognition ability. *Cognition*, *238*, 105542. https://doi.org/10.1016/j.cognition.2023.105542

Corey, D. M., Dunlap, W. P., & Burke, M. J. (1998). Averaging Correlations: Expected Values and Bias in Combined Pearson *r* s and Fisher's *z* Transformations. *The Journal of General Psychology*, *125*(3), 245–261. https://doi.org/10.1080/00221309809595548

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*(11), 671.

Cronbach, L. J., & Meehl, P. E. (1955). Construct Validity in Psychological Tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Dennett, H. W., McKone, E., Tavashmi, R., Hall, A., Pidcock, M., Edwards, M., & Duchaine, B. (2012). The Cambridge Car Memory Test: A task matched in format to the Cambridge Face Memory Test, with norms, reliability, sex differences, dissociations from face memory, and expertise effects. *Behavior Research Methods*, *44*(2), 587–605. https://doi.org/10.3758/s13428-011-0160-2

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, *44*(4), 576–585. https://doi.org/10.1016/j.neuropsychologia.2005.07.001

Ekstrom, R. B., French, J. W., & Harman, H. H. (1976). *Kit of factor-referenced cognitive tests*. Educational Testing Service.

Flanagan, D. P., & Dixon, S. G. (2014). The Cattell-Horn-Carroll Theory of Cognitive Abilities. In C. R. Reynolds, K. J. Vannest, & E. Fletcher-Janzen (Eds.), *Encyclopedia of Special Education* (p. ese0431). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118660584.ese0431

Gauthier, I. (2018). Domain-Specific and Domain-General Individual Differences in Visual Object Recognition. *Current Directions in Psychological Science*, *27*(2), 97–102. https://doi.org/10.1177/0963721417737151

Gauthier, I., Cha, O., & Chang, T.-Y. (2022). Mini review: Individual differences and domain-general mechanisms in object recognition. *Frontiers in Cognition*, *1*. https://doi.org/10.3389/fcogn.2022.1040994

Gauthier, I., & Fiestan, G. (2023). Food neophobia predicts visual ability in the recognition of prepared food, beyond domain-general factors. *Food Quality and Preference*, *103*, 104702. https://doi.org/10.1016/j.foodqual.2022.104702

Gauthier, I., & Tarr, M. J. (1997). Becoming a "Greeble" Expert: Exploring Mechanisms for Face Recognition. *Vis. Res.*, *37*(12), 1673–1682. https://doi.org/10.1016/S0042-6989(96)00286-6

Gerbing, D. W., & Anderson, J. C. (1984). On the Meaning of within-Factor Correlated Measurement Errors. *Journal of Consumer Research*, *11*(1), 572. https://doi.org/10.1086/208993

Growns, B., Dunn, J. D., Mattijssen, E. J. A. T., Quigley-McBride, A., & Towler, A. (2022). Match me if you can: Evidence for a domain-general visual comparison ability. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-021-02044-2

Growns, B., Towler, A., & Martire, K. (2023). The novel object-matching test (NOM Test): A psychometric measure of visual comparison ability. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02069-6

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Kelley, T., Lee. (1927). *Interpretation of educational measurements*. World Book Company.

Kieseler, M., Dickstein, A., Krafian, A., Li, C., & Duchaine, B. (2022). HEVA – A new basic visual processing test [Poster Presentation]. *Journal of Vision*, *22*(14). https://doi.org/10.1167/jov.22.14.4109

Kuhn, T. S. (1961). The Function of Measurement in Modern Physical Science. *Isis*, *52*(2), 161–193. https://doi.org/10.1086/349468

Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When 'good' indicators are bad and 'bad' indicators are good. *Psychological Methods*, *4*(2), 192–211. https://doi.org/10.1037/1082-989X.4.2.192

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley.

Lubinski, D. (2004). Introduction to the Special Section on Cognitive Abilities: 100 Years After Spearman's (1904) '"General Intelligence," Objectively Determined and Measured'. *Journal of Personality and Social Psychology*, *86*(1), 96–111. https://doi.org/10.1037/0022-3514.86.1.96

McGugin, R. W., Richler, J. J., Herzmann, G., Speegle, M., & Gauthier, I. (2012). The Vanderbilt Expertise Test reveals domain-general and domain-specific sex effects in object recognition. *Vision Research*, *69*, 10–22. https://doi.org/10.1016/j.visres.2012.07.014

McGugin, R. W., Sunday, M. A., & Gauthier, I. (2022). The neural correlates of domain-general visual ability. *Cerebral Cortex*, bhac342. https://doi.org/10.1093/cercor/bhac342

Mollon, J. D., Bosten, J. M., Peterzell, D. H., & Webster, M. A. (2017). Individual differences in visual science: What can be learned and what is good experimental practice? *Vision Research*, *141*, 4–15. https://doi.org/10.1016/j.visres.2017.11.001

Nunnally, J. C. (1994). *Psychometric theory 3E*. Tata McGraw-hill education.

Pluck, G. (2019). Preliminary Validation of a Free-to-Use, Brief Assessment of Adult Intelligence for Research Purposes: The Matrix Matching Test. *Psychological Reports*, *122*(2), 709–730. https://doi.org/10.1177/0033294118762589

Richler, J. J., Tomarken, A. J., Sunday, M. A., Vickery, T. J., Ryan, K. F., Floyd, R. J., Sheinberg, D., Wong, A. C. –N., & Gauthier, I. (2019). Individual Differences in Object Recognition. *Psychological Review*, *126*(2), 226–251. https://doi.org/10.1037/rev0000129

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

Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral Development and Construct Validity: The Principle of Aggregation. *Psychological Bulletin*, *94*(1), 18–38. https://doi.org/10.1037/0033-2909.94.1.18

Schneider, W. J., & McGrew, K. S. (2018). The Cattell-Horn-Carroll Theory of Cognitive Abilities. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 73–163). Guilford Publications.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Smithson, C. J. R., Chow, J. K., & Gauthier, I. (2024). *Visual and auditory object recognition and their relations to spatial abilities*. Manuscript in Preparation.

Smithson, C. J. R., Eichbaum, Q. G., & Gauthier, I. (2023). Object recognition ability predicts category learning with medical images. *Cognitive Research: Principles and Implications*, *8*(1), 9. https://doi.org/10.1186/s41235-022-00456-9

Spearman, C. (1904). 'General Intelligence,' Objectively Determined and Measured. *The American Journal of Psychology*, *15*(2), 201–292.

Spearman, C. (1907). Demonstration of Formulæ for True Measurement of Correlation. *The American Journal of Psychology*, *18*(2), 161–169.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149. https://doi.org/10.3758/BF03207704

Stenner, A. J., Smith III, M., & Burdick, D. S. (2022). Toward a theory of construct definition. In *Explanatory Models, Unit Standards, and Personalized Learning in Educational Measurement: Selected Papers by A. Jackson Stenner* (pp. 45–55). Springer Nature Singapore.

Sunday, M. A., Donnelly, E., & Gauthier, I. (2018). Both fluid intelligence and visual object recognition ability relate to nodule detection in chest radiographs. *Applied Cognitive Psychology*, *32*(6), 755–762. https://doi.org/10.1002/acp.3460

Sunday, M. A., Tomarken, A., Cho, S.-J., & Gauthier, I. (2022). Novel and familiar object recognition rely on the same ability. *Journal of Experimental Psychology: General*, *151*(3), 676–694. https://doi.org/10.1037/xge0001100

Vogel, E. K., & Awh, E. (2008). How to exploit diversity for scientific gain: Using individual differences to constrain cognitive theory. *Current Directions in Psychological Science*, *17*(2), 171–176.

Wang, M. W., & Stanley, J. C. (1970). Differential Weighting: A Review of Methods and Empirical Studies. *Review of Educational Research*, *40*(5), 663–705.

White, D., & Burton, A. M. (2022). Individual differences and the multidimensional nature of face perception. *Nature Reviews Psychology*, *1*(5), 287–300.

Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: The SHINE toolbox. *Behavior Research Methods*, *42*(3), 671–684. https://doi.org/10.3758/BRM.42.3.671

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. https://doi.org/10.1163/156856808786451408

Supplementary Materials

Table 1

Model Fit Indices – longitudinal measurement invariance.

Model	Description	*df*	χ^2	$\Delta\chi^2$	RMSEA [90% CI]	NNFI	SRMR	AIC	BIC
1	No Equality Constraints	5	1.4 *p* = .92		0 [0, .035]	1.030	.008	7398.24	7466.30
2	Metric Invariance: equal path coefficients and across sessions.	8	2.74 *p* = .95	1.34 *p* = .72	0 [0, .005]	1.028	.020	7393.58	7452.36
3	Scalar Invariance: equal path coefficients and intercepts across sessions.	11	17.28 *P* = .1	14.54 *P* = .002	.059 [0, .11]	0.976	.042	7402.12	7451.62

Table 2

Model Fit Indices – one factor vs two factor model.

Model	Description	*df*	χ^2	$\Delta\chi^2$	RMSEA [90% CI]	NNFI	SRMR	AIC	BIC
1	Two *o* latent variables: one for each timepoint, with correlated errors.	5	1.4 *p* = .92		0 [0, .035]	1.030	.008	7398.24	7466.30
2	One *o* latent variable for all measures, with correlated errors, implying perfect correlation in latent ability from time 1 to time 2.	6	3.79 *p* = .71	2.39 *p* = .12	0 [0, .005]	1.016	.011	7398.63	7463.6
3	One *o* latent variable, with equal path coefficients across sessions	9	5.35 *P* = .8	1.56 *P* = .67	0 [0, .057]	1.017	.021	7394.19	7449.88