Smithson *et al.*
Cognitive Research: Principles and Implications (2023) 8:9
https://doi.org/10.1186/s41235-022-00456-9

Cognitive Research: Principles and Implications

ORIGINAL ARTICLE

Open Access

Object recognition ability predicts category learning with medical images

Conor J. R. Smithson[1*], Quentin G. Eichbaum[2,3] and Isabel Gauthier[1]

Abstract

We investigated the relationship between category learning and domain-general object recognition ability (*o*). We assessed this relationship in a radiological context, using a category learning test in which participants judged whether white blood cells were cancerous. In study 1, Bayesian evidence negated a relationship between *o* and category learning. This lack of correlation occurred despite high reliability in all measurements. However, participants only received feedback on the first 10 of 60 trials. In study 2, we assigned participants to one of two conditions: feedback on only the first 10 trials, or on all 60 trials of the category learning test. We found strong Bayesian evidence for a correlation between *o* and categorisation accuracy in the full-feedback condition, but not when feedback was limited to early trials. Moderate Bayesian evidence supported a difference between these correlations. Without feedback, participants may stick to simple rules they formulate at the start of category learning, when trials are easier. Feedback may encourage participants to abandon less effective rules and switch to exemplar learning. This work provides the first evidence relating *o* to a specific learning mechanism, suggesting this ability is more dependent upon exemplar learning mechanisms than rule abstraction. Object-recognition ability could complement other sources of individual differences when predicting accuracy of medical image interpretation.

Keywords Category learning, Categorisation, Object recognition, Individual differences, Medical images, Radiology

Introduction

Accurate interpretation of medical images plays a crucial role in the diagnosis of many medical conditions. This process often requires the visual detection of abnormalities, such as lung nodules in radiographs or masses in mammograms. Although experts undergo substantial training, they cannot always make the correct decision (Brady, 2017; Graber et al., 2002). For many diagnostic tests, there are substantial discrepancies in accuracy between practitioners, in part due to differences in experience (Itani et al., 2019; Rudolph et al., 2021). Practitioners may even disagree with their own initial judgement when asked to review images a second time (Abujudeh et al., 2010). Although precise estimates of the prevalence of medical imaging errors are difficult to obtain, as errors vary widely based on test, practice setting, and population, estimates of real-world error rates range from < 1% to around 10% (Gergenti & Olympia, 2019; Lamoureux et al., 2021; Lockwood, 2017). Error rates can be higher still when the relevant disease is rare in the studied population (Kolb et al., 2002). These errors have multiple causes at the individual and the system level, including fatigue, communication failure, biased reasoning, failures of visual search, interpretive errors, technological errors, and poor technique (Lee et al., 2013; Waite et al., 2017). A majority of radiological errors are perceptual in nature, with practitioners failing to spot abnormalities, whereas a smaller but still substantial proportion of errors are due to failure to correctly categorise

*Correspondence:
Conor J. R. Smithson
conor.smithson@vanderbilt.edu
[1] Department of Psychology, Vanderbilt University, PMB 407817, 2301 Vanderbilt Place, Nashville, TN 37240-7817, USA
[2] Department of Pathology, Microbiology and Immunology, Vanderbilt University, Nashville, USA
[3] Vanderbilt Pathology Education Research Group, Nashville, USA

abnormalities (Donald & Barnard, 2012; Ferguson et al., 2021; Kim & Mansfield, 2014).

As the accurate interpretation of medical images relies on the detection and categorisation of objects, differences in diagnostic accuracy among practitioners may partially result from individual differences in visual abilities. The existence of such differences is supported by evidence for a domain-general object recognition ability (*o*). Confirmatory factor models demonstrate that diverse measures of object recognition, with differing task demands and differing object categories, load onto a single higher-order factor (Richler et al., 2019). The *o* factor explains variance in scores on object recognition tests beyond that explained by intelligence and visual working memory, and it can do so for both familiar and unfamiliar object categories (Richler et al., 2017; Sunday et al., 2022). In studies that are not concerned with investigating the structure of this visual ability, or that for practical reasons cannot achieve the sample size or the number of tasks required for structural equation modelling, an aggregate approach (Rushton et al., 1983) to measuring object recognition ability has been used (Chang & Gauthier, 2021; Chow et al., 2022). In this approach, *z*-scores on two object recognition tests that differ in format and stimuli are averaged to estimate the level of the underlying *o* ability. This approach provides a valid compromise in estimating *o* in smaller samples and when time is limited (Smithson et al., 2022). Using this approach, Sunday et al. (2018) found that *o* predicts the accurate detection of lung nodules in chest radiographs for both novices and experts, demonstrating a link between *o* and successful abnormality detection. The detection of lung nodules depends on successful visual search, but other radiological tasks rely less on visual search, and more on accurate categorisation.

As *o* captures the ability to learn individual identities, it is unclear whether it will also predict accurate categorisation. There are demonstrated individual differences in both speed and accuracy of category learning, in addition to differences in strategy use. Some people rely more on the abstraction of simple rules, leading to categorisation decisions based on one dimension. Others preferentially rely on judgements of perceptual similarity to category exemplars, which can be measured in a space defined by several relevant dimensions (Little & McDaniel, 2015; Wahlheim et al., 2016). For example, one may learn to categorise a skin mole as cancerous if it is asymmetrical, but one may also rely on comparisons of the mole to remembered examples of cancerous and non-cancerous moles. *o* predicts performance on many visual tasks that require judgements other than individuation, such as visual search, and judgements of summary statistics for ensembles (Chang & Gauthier, 2021; Sunday et al., 2018). Given that *o* can predict such a wide array of visual tasks, it is reasonable to question whether *o* could also predict accurate categorisation in a visual domain. There is some support for a relationship between individual differences in object recognition (measured by one of the tasks that tap into *o*) and accurate categorisation of medical images, at least under some conditions. In one study, participants categorised white blood cells as cancerous or not under conditions emphasising speed or accuracy, or when provided with a biased cue (Trueblood et al., 2018). Performance on an object recognition test predicted accurate categorisation, particularly for categorisation following biased cues. While this suggests that categorisation may rely on visual abilities under some conditions more than others, this work did not have sufficient power to compare correlations across conditions.

The study of domain-general high-level visual abilities is an emerging research area (Gauthier, 2018; Gauthier et al., 2022), and the extent to which these abilities can explain variability in performance on real-world tasks is still unclear. As diagnostic imaging has a heavy visual component, it is plausible that visual abilities may influence performance on these tasks. A good first step in showing this is to demonstrate that *o* can predict accurate categorisation of medical images. To investigate this, we created a three-alternative forced choice test in which participants learn to categorise white blood cell images as cancerous (blast) or non-cancerous (non-blast). We used a novice sample to test the relationship between *o* and categorisation in the absence of extensive pre-existing experience, which could contaminate the relationship.

Study 1

Participants

Thirty-nine Vanderbilt University students participated for course credit. A further sixty-seven adults were recruited on Amazon Mechanical Turk. Recruitment criteria required the use of a US IP address, greater than 50 approved hits, and a greater than 90% approval rating. We used a Bayesian stopping rule, collecting data in batches, until the Bayes Factor for the correlation between *o* and performance on the Blast Test reached a suggested threshold for moderate evidence, $BF_{10} > 3$ or $BF_{10} < 1/3$ (Lee & Wagenmakers, 2014). From our total of 106 participants, we excluded 26 for below-chance performance on either of the two tests used for estimating *o*.[1] This left 80 participants in the final analysis (Mage = 34.7, SD = 14.8; 32 men, 46 women, 2 other).

[1] 19 excluded participants were from the Mechanical Turk sample, and 7 were from the student sample.

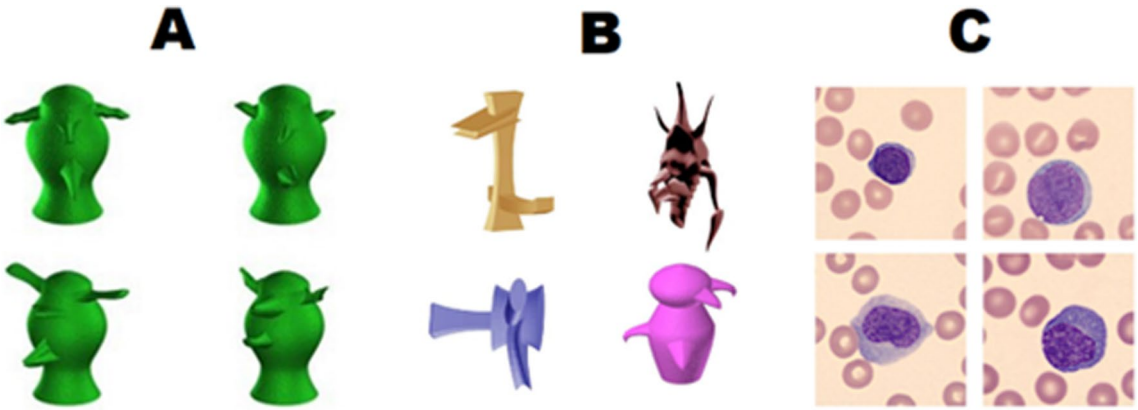

Fig. 1 Example Stimuli. **A:** Symmetrical Greebles used in the NOMT. **B:** Novel objects used in the Object Matching Test. From the top left, anticlockwise: vertical and horizontal Ziggerins, asymmetrical Greebles, and Sheinbugs. **C:** White blood cells used in the Blast Test

Materials and procedure

Participants completed three on-screen tests. First, they completed the Blast Test—a category learning test involving the identification of cancerous cells. After this, they completed the Novel Object Memory Test (NOMT) and the Object Matching Test, which were used to estimate *o*. Example stimuli for all three tests can be seen in Fig. 1. We used a fixed order of trials for all tests to minimise variance due to factors other than individual differences. Informed consent was obtained from all participants, and the study was approved by the Vanderbilt University Institutional Review Board.

Development of Blast Test

We obtained images of blast and non-blast blood cells from peripheral blood smears conducted at Vanderbilt University Medical Center. These images have been used in prior research on medical decision making (Hasan et al., 2021; Trueblood et al., 2018). They were categorised by expert consensus as blast or non-blast. They were additionally sorted into easy or hard categories on the basis of whether each cell image shared features common to the other category (Trueblood et al., 2018). We initially created 100 trials. Each trial was composed of two non-blast cells, and one blast cell. The task on each trial was to identify the blast cell from a side-by-side display of the three images. The use of three cells to choose from on each test trial reduces the importance of response bias and reduces the successful random guessing rate for each trial, compared to using only two. We initially created 25 trials composed of one easy blast image, and two easy non-blast images; 25 trials composed of one easy blast and two hard non-blast images; 25 trials composed of one hard blast and two easy non-blast images; and 25 trials composed of one hard blast, and two hard non-blast images. On the basis of pilot testing, we selected trials from a broad range of difficulty levels to maximise the informativeness of our test across a wide range of ability levels. Trials were also selected for high reliability; we checked item-rest correlations, and internal consistency if an item was dropped, and dropped those that reduced reliability. In the final Blast Test, the trials were ordered from easiest to hardest based on our pilot data. To familiarise participants with the two categories, participants were first shown 6 blast blood cell images, and 6 non-blast blood cell images, with category membership clearly labelled. Participants then completed 60 trials. Feedback indicating whether responses were correct appeared at the top of the screen for 1 s after each of the first 10 trials. No feedback was given for the remaining 50 trials. Percent correct over the 60 trials indexed performance.

Tests to estimate *o*

As in prior work, we used the aggregate of two object recognition tests to estimate *o* (e.g. Chang & Gauthier, 2021; Chow et al., 2022; Sunday et al., 2018). These two tests were chosen from a battery of tests that were good indicators of *o* in confirmatory factor models (Richler et al., 2019; Sunday et al., 2022), on the basis that they have different test constraints and use different categories of novel objects. The aggregation of scores from tests using different object categories and different task demands purifies the measurement of domain-general ability, by reducing the proportion of variance in scores that is due to irrelevant variation specific to particular task demands or stimuli (Rushton et al., 1983). The expected correlation for a pair of such tests is relatively low (0.3–0.4) because superficial features of the tests and stimuli are different. The aggregate of standardised performance on two tests provides a good estimate ($r \approx 0.8$) of *o* measured as a factor score in a confirmatory factor analysis based on six tests (Smithson et al., 2022).

Novel object memory test

The NOMT was developed to assess object recognition ability (Richler et al., 2017). Participants were asked to memorise six exemplars from a category of novel objects (symmetrical Greebles; Gauthier & Tarr, 1997). They then viewed these six targets for as long as they needed, before completing six test trials. On each test trial, one target Greeble appeared alongside two distractor Greebles. Participants had unlimited time to select the target Greeble with their mouse on each trial. Participants then reviewed the targets and completed a further 18 test trials. Participants were then informed that the Greebles could appear in different viewpoints on remaining trials. The targets were presented again for review, prior to the final 24 test trials. Percent correct over the 48 test trials indexed performance.

Object matching test

On each trial participants had to determine whether two serially presented images displayed the same object. The objects were selected from four categories of novel objects: asymmetrical Greebles, Sheinbugs, and two distinct categories of Ziggerins (Richler et al., 2019). Each trial used either one or two objects from the same category. Each trial began with the presentation of a central fixation cross for 500 ms. The target object was then presented for 300 ms before a visual mask composed of scrambled object parts appeared for 500 ms. Finally, another object was presented which was either the same as the target or different. Participants had four seconds to respond by clicking either the same or different buttons on-screen. The target object could change in orientation or size from study to test, but participants were asked to judge only whether the identity of the object was the same. After an initial four practice trials, participants completed 70 test trials. Performance was indexed by a signal detection theory measure of sensitivity (d'). Timed-out responses were not included in the calculation. Less than 1% of all trials had timed-out responses.

Results

Descriptive statistics and reliability for each test can be seen in Table 1. To estimate *o*, *z*-scores for percent accuracy on the NOMT and d' on the Object Matching Test were averaged. Correlational analyses used a Jeffreys-beta prior (Jeffreys, 1961) with a scale of 1 and were conducted with the BayesFactor Package (Morey & Rouder, 2021) in R. BF_{+0} indicates a one-sided test in the positive direction, and BF_{10} is used for two-sided tests. We report highest posterior densities as 95% credibility intervals, and the median of the posterior distribution is used for parameter estimation. Our reported CIs and parameter estimates are always calculated from two-sided analyses.

Table 1 Descriptive statistics

Test	Mean (*SD*)	Reliability
NOMT (percent accuracy)	55.4% (15.2%)	0.75
Object Matching (d')	0.98 (0.53)	0.73
o	0 (0.82)	0.81
Blast (percent accuracy)	64.5% (19.1%)	0.93

Reliability is calculated using Pearson's *r* between two halves composed of alternating trials, with the Spearman–Brown prophecy formula applied. Aggregate reliability of *o* was calculated with equal weighting using a formula adapted from Wang and Stanley (1970)

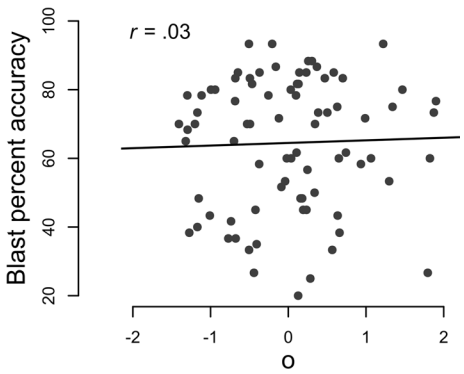

Fig. 2 Correlation between *o* and percent accuracy on the Blast Test

As expected, there was very strong Bayesian evidence for a positive correlation between performance on the NOMT and the Object Matching Test ($r = 0.33$, 95% CI [0.13, 0.52], $BF_{+0} = 31.07$). We obtained moderate evidence against a correlation between *o* and percent accuracy on the Blast Test ($r = 0.03$, 95% CI [− 0.18, 0.25] $BF_{+0} = 0.18$, Fig. 2), although this was somewhat sensitive to the choice of prior, with the Bayes factor rising above 1/3rd for prior scales equalling or below 0.32.

Discussion

Contrary to our hypothesis, *o* did not predict performance on the Blast Test. One possible reason for the lack of a relationship with category learning may be the limited amount of feedback that participants received. Although tests that are used to estimate *o* do not use feedback, and *o* can predict other skills measured in tests without feedback, such as working memory judgements with musical notation (Chang & Gauthier, 2021) or food oddball judgements (Gauthier & Fiestan, 2023), the strategies and mechanisms recruited during category learning may be particularly sensitive to feedback. Early on in category learning, people tend to rely on simple rule-based judgements and then update these rules as they receive further feedback, before shifting to similarity-based

Table 2 Descriptive statistics

	Limited feedback		Full feedback	
Test	**Mean (*SD*)**	**Reliability**	**Mean (*SD*)**	**Reliability**
NOMT (percent accuracy)	59.72% (15.76%)	0.82	57.07% (14.24%)	0.76
Object Matching (d')	1.44 (0.6)	0.81	1.33 (0.6)	0.83
o	0 (0.77)	0.85	0 (0.81)	0.84
Blast (percent accuracy)	61.46% (17.62%)	0.89	66.46% (18.49%)	0.92

Reliability is calculated using Pearson's *r* for two halves composed of alternating trials, with the Spearman–Brown prophecy formula applied. Blast Test analyses here are of the last 50 trials only. Aggregate reliability of *o* was calculated with equal weighting using a formula adapted from Wang and Stanley (1970)

exemplar retrieval as expertise develops (Johansen & Palmeri, 2002). In Study 1, we only provided participants with feedback on the first ten trials of the Blast Test. As earlier trials in the Blast Test are easier, participants did not receive any feedback on more difficult trials. Individual differences in performance may thus result from divergent initial rule choices, or differing success in the application of these rules. Due to the limited feedback, participants may have seen no need to update their initial rules or may have had no basis on which to do so. Additionally, the lack of feedback may have discouraged a switch in strategy to reliance on judgements of perceptual similarity to prior exemplars. Harder trials are presumably more likely to require methods of judgement other than simple rule use. The tests used to estimate *o* require within-category individuation, which also cannot usually rely on the use of simple rules, as objects in a common category will share a basic configuration of parts.

To test whether the lack of association between *o* and Blast Test accuracy was due to the limited feedback, we repeated the study with the addition of a full-feedback condition, wherein participants received feedback for all 60 trials of the Blast Test.

Study 2

Materials and procedure

Participants completed the same three tests as in study one. However, the tests were in a different fixed order: NOMT, Object Matching Test, and Blast Test. In Study 2, we compared the limited feedback and the full-feedback versions of the Blast Test, so placing this test last ensured that performance on the two object recognition tests could not be affected by assignment to either condition of the Blast Test. The NOMT was modified such that participants had a fixed 20 s to familiarise themselves with the six targets on each study trial, reducing differences in study time as a source of individual differences. The Object Matching Test was altered so that for the first 35 trials the study object was presented for 600 ms. This was done to lower difficulty on some trials, as mean d' was low in Study 1 (0.97). For the remaining 35 trials, the study object was presented for 300 ms, as in Study 1. Another alteration was to allow unlimited time to respond, eliminating timed-out responses so that d' was calculated for the exact same trials for all participants.

Participants

Due to the high percentage of participants excluded in Study 1 for below-chance performance (27.4%), we switched recruitment platform to Prolific.co. We recruited 245 participants, with the requirement of English fluency. Our pre-set exclusion criteria excluded one participant who failed more than one of five attention checks spread throughout the study. This method of exclusion allowed us to treat low scores as valid. The attention checks were dummy trials in which participants were instructed to click on a specific response option. Participants were randomly assigned to the limited or the full-feedback condition. Due to error, the first 6 participants were non-randomly assigned to the limited feedback condition. Once we reached 122 participants in the limited feedback condition (64 men, 57 women, 1 other; Mage = 25.4, SD = 6.3), we added 21 participants to the full-feedback condition (50 men, 67 women, 5 other; Mage = 26.3, SD = 8.8) to achieve equal group sizes (which were unequal due to random assignment as well as the initial error). We then ceased collecting data as we were able to find a conclusive Bayes factor for the existence of a correlation between *o* and categorisation accuracy in the full-feedback condition.

Results

Descriptive statistics and reliability for each test can be seen in Table 2. As expected, there was a correlation between the NOMT and the Object Matching Test (r = 0.26, 95% CI [0.14, 0.37], BF_{+0} = 752.82) across all participants. There was inconclusive evidence for an overall difference in accuracy on the last 50 trials of the Blast Test (conditions diverge after the first 10 trials) between the limited and the full-feedback conditions (BF_{+0} = 2.48). For the limited feedback condition, there was inconclusive evidence against a correlation

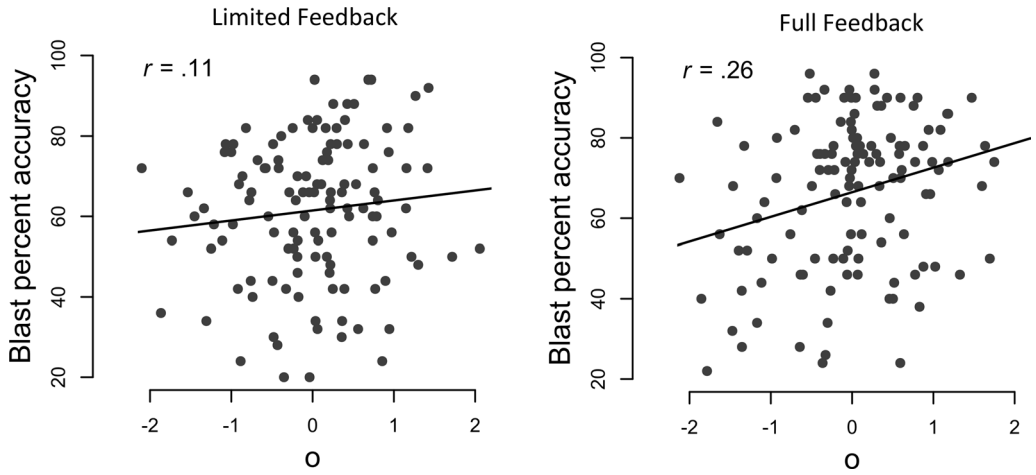

Fig. 3 Correlation between o and percent accuracy on the Blast Test

between *o* and percent accuracy on the Blast Test ($r = 0.11$, 95% CI [− 0.07, 0.28], $BF_{+0} = 0.41$, Fig. 3.). For the full-feedback condition, there was strong evidence for a correlation between *o* and percent accuracy on the Blast Test ($r = 0.26$, 95% CI [0.10, 0.43], $BF_{+0} = 18.75$, Fig. 3.). Using BFpack (Mulder et al., 2019) in R, we obtained moderate Bayesian evidence for the hypothesis that the correlation between *o* and blast percent accuracy is greater in the full-feedback condition than in the limited feedback condition, compared to its complement ($BF_{10} = 8.34$).

Discussion

In support of our main hypothesis, we found a relationship between *o* and accuracy in the full-feedback condition of the Blast Test. We further showed that this relationship was greater than the relationship between *o* and Blast Test accuracy in the limited feedback condition. This suggests a relationship between *o* and category learning. We hypothesised that such a correlation would emerge in the full-feedback condition because greater feedback may allow for a shift from the use of simple rules for categorisation judgements to the use of complex rules, or a change in strategy from rule-based judgements to judgements based on perceptual comparison of cells against prior exemplars. However, because each trial included one blast image, and two non-blast images, one task-specific strategy that participants could have employed is to select the odd one out. Perhaps particularly for the limited feedback condition, participants may have used this strategy instead of trying to learn to categorise blast cells explicitly. To test this possibility, in study 3 we assessed participants in a no-feedback version of the Blast Test.

Study 3

Participants and method

Fifty-one participants (Mage = 24.69, $SD = 6.8$; 24 men, 24 women, 3 other) were recruited on the Prolific.co platform, with a requirement for English fluency. We then assessed if our stopping criteria had been met, which was a Bayes factor < 1/3 or > 3 for a correlation between *o* and Blast Test accuracy. No participants were excluded, as none missed more than one of five attention checks. Participants completed the same tests in the same order as in Study 2. The only differences were that the Blast Test gave no feedback on performance, and no examples of blast and non-blast cells were given prior to testing. Participants were instructed to try their best to choose the cancerous cell on each trial, despite the lack of examples or feedback.

Results and discussion

Descriptive statistics are presented in Table 3. Performance in the no-feedback condition of the Blast Test was slightly below-chance level (chance = 33%; $M = 0.3$, $SD = 0.16$), and we obtained strong Bayesian evidence

Table 3 Descriptive statistics

Test	Mean (*SD*)	Reliability
NOMT (percent accuracy)	59% (15%)	0.79
Object Matching (*d'*)	1.24 (0.71)	0.85
o	0 (0.72)	0.83
Blast (percent accuracy)	30% (16%)	0.90

Reliability is calculated using Pearson's *r* for two halves composed of alternating trials, with the Spearman–Brown prophecy formula applied. Blast Test here is the no-feedback version. Aggregate reliability of o was calculated with equal weighting using a formula adapted from Wang and Stanley (1970)

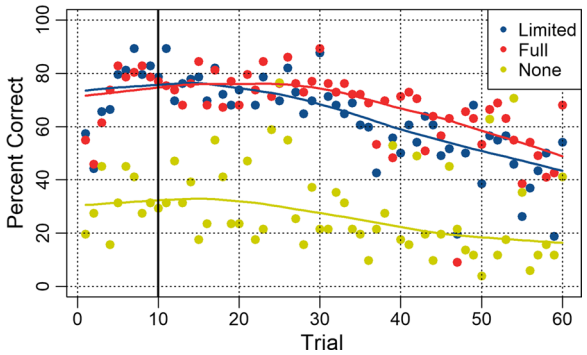

Fig. 4 Percentage of participants who responded correctly per trial in Study 2 and 3. Note. Limited-feedback and full-feedback conditions are from Study 2, No-feedback condition is from Study 3. The solid vertical line indicates trial 10, after which feedback only continues for the full-feedback condition

from a one-sample *t*-test that performance is not above chance ($BF_{+0} = 0.06$). It appears that the presentation of a small number of examples and the presence of limited feedback are required for most people to perform above chance. Given the high performance in Study 1 and 2, it is unlikely that participants in those studies made use of an odd-one-out strategy. There was no correlation between *o* and accuracy in this version of the Blast Test ($r = -0.01$, 95% CI [− 0.28, 0.25], $BF_{+0} = 0.16$).

To determine whether trial difficulty was consistent between the feedback conditions in Study 2 and the no-feedback condition in Study 3, we tested for cross-condition correlations for the percentage of participants who responded correctly on each trial. Figure 4 shows average accuracy per trial. To counter skewness, we applied a log transformation to the limited (− 0.75 to 0.32) and full (− 1.44 to − 0.15) feedback conditions, and a square root transformation to the no-feedback condition (0.82 to 0.25). There was a high correlation between trial accuracy in the full-feedback condition and the limited feedback condition ($r = 0.83$, 95% CI [0.74, 0.91], $BF_{10} = 9.18 \times 10^{13}$). There were much smaller correlations between trial accuracy in the no-feedback condition and the full feedback ($r = 0.24$, 95% CI [− 0.01, 0.48], $BF_{10} = 0.94$) and limited feedback ($r = 0.31$, 95% CI [0.08, 0.53], $BF_{10} = 3.42$) conditions.

Although the Blast Test was not designed to reveal strategy use, we suspected that participants who rely primarily on a simple rule may choose to use size as the basis and select the largest cell as being cancerous. We measured the maximum diameter of the cells in each trial, excluding small appendages (except in the case of ties), and found that size was diagnostic in nine of the first ten trials, and 70% of first-half trials, but only diagnostic in 33% of second-half trials.[2] If participants in the limited feedback condition rely on a simple rule, it is likely that they would form a size-based rule, given that it is so diagnostic for early trials. Indeed, we found a strong correlation between percent accuracy on each trial and whether size was diagnostic in the no feedback ($r = 0.56$, 95% CI [0.37, 0.72], $BF_{+0} = 26{,}208.3$) and limited feedback conditions ($r = 0.4$, 95% CI [0.18, 0.60], $BF_{+0} = 54.78$), but a weaker one in the full-feedback condition ($r = 0.31$, 95% CI [0.08, 0.53], $BF_{+0} = 7.43$).

General discussion

We found evidence in Study 2 that *o* predicts performance on a category learning task requiring the assessment of radiological images. This demonstrates a link between individual differences in individuation and categorisation, adding to a growing corpus of literature suggesting that individual differences in a wide variety of visual tasks are related (Chang & Gauthier, 2021, 2022; Growns et al., 2022). We also demonstrate in Study 1 and 2 that *o* either does not predict or at most predicts to a lesser extent, categorisation accuracy when participants receive feedback on only a very limited number of easy categorisation trials. This is despite the fact that performance in this limited feedback condition was very similar to performance in a condition where feedback was given continuously. Furthermore, participants who received limited feedback in Study 1 and 2 successfully categorised cells as cancerous or not with a success rate approximately double that of participants in Study 3, who received no feedback. In other words, we find the biggest difference in performance on the Blast Test between the no-feedback condition (Study 3) and any of the other conditions in which examples and some feedback were presented. This is not surprising given the extensive literature on the advantages of supervised learning for categories that are not based on a simple verbalisable rule (Ashby et al., 1999, 2002). In contrast, the difference between providing feedback only on ten trials vs. all trials was more modest in terms of average performance on the Blast Test. Nonetheless, this additional feedback made a substantial difference across individuals, leading to an advantage for those participants with higher *o*.

The correlation between *o* and performance in the full-feedback condition was small ($r = 0.26$, or $r = 0.30$ when accounting for attenuation from measurement error). The effect is similar to that in Sunday et al. (2018), who found a correlation of $r = 0.28$ between *o* and decisions on a test of tumour detection in chest radiographs, after controlling for intelligence. The correlation could be

[2] Trial images are available at https://osf.io/yaqs8/.

limited by the fact that the Blast Test is short and allows different strategies. But more importantly, *o* is conceived as a general ability that does not reflect specifics of the domain or the task constraints. Performance on any test is explained by a variety of factors, some of them general and some specific to the test. For instance, two similarly formatted matching tests may correlate more strongly (e.g. Growns et al., 2022), but some of this correlation may be due to specific task requirements. When tests with different formats use similar stimuli (e.g. faces, Wilmer et al., 2014), the resulting strong correlation is partially due to the common domain. But when two tests differ in both format and domain, like the tests we use to estimate *o,* the shared variance is expected to be smaller. Importantly, the advantage is that we can expect a domain-general ability to predict some of the variance in other very different tasks, such as the Blast categorisation test. This is somewhat analogous to intelligence, which is, for instance, a predictor of job performance in many domains, with effect sizes that are comparable to what we observe here (e.g. $r = 0.33$; see Ree & Earles, 1992, for a review). In addition, it is important to note that a small effect size when measured in a single task can translate into large consequences in the long run, in real-world situations where individuals make a very large number of perceptual decisions in the course of their work (Funder & Ozer, 2019).

By crossing the measurement of *o* with an experimental manipulation, our results are the first to speak to the mechanisms that support this ability, because of the extensive literature distinguishing different modes for category learning. One influential model proposes two systems for category learning, one using simple explicit rules, and the other for learning more complex multidimensional categories that are difficult to verbalise (Ashby et al., 1999). A different account suggests that even within a single system, feedback is more critical for more cognitively demanding categorisation tasks (Le Pelley et al., 2019). In the Blast Test, early category learning can plausibly rely primarily on simple rule generation. As difficulty increases, a simple rule will become ineffective and participants need to switch to judgements of similarity to stored exemplars. In summary, many different visual tasks tap into *o*, but this ability may not support categorisation following simple verbalisable rules, or categorisation of a multidimensional nature without mechanisms responsible for supervised learning. Fully supervised learning may not be necessary; semi-supervised learning (e.g. labelling a few exemplars) is the predominant method by which humans learn categories (Gibson et al., 2013; LaTourrette & Waxman, 2019), and we would also expect *o* to predict categorisation abilities that have developed through this method. These conjectures could be more directly addressed by testing for correlations between *o* and accuracy on categorisation tasks that only require simple rules, require a combination of rules to increase cognitive demands, or which require multidimensional judgements which cannot easily be reduced to verbalisable rules. To make strong claims about the underlying categorisation strategies being employed by individual participants would require an analysis of response patterns in tasks that have been designed to reveal strategy use.

The finding that *o* can predict successful categorisation adds to existing knowledge that *o* can predict successful visual search in a radiological task (Sunday et al., 2018). Both perceptual and interpretative skills are fundamental for radiological diagnostics. Therefore, *o* may plausibly predict diagnostic accuracy; although we have not yet tested this in an ecologically valid design. Accuracy on category learning experiments can be influenced by task demands and sequence effects, in which case, performance may reflect the successful use of task-specific strategies that may be hard to account for (Richler & Palmeri, 2014; Stewart et al., 2002). Nevertheless, when combined with the influences of experience and general intelligence, the contribution of *o* may provide a fuller explanation of individual differences in diagnostic accuracy. Further research should explore a relationship between *o* and categorisation accuracy in a radiological task using an expert sample. The contribution of *o* to performance and the conditions necessary for it to be used may well differ in experts. For instance, the contribution of *o* to categorisation may be greater for experts than for novices, because experts have more exemplars in memory and may rely less on simple categorisation rules. In addition, feedback may not be necessary for a relationship with *o* to emerge in experts who can already categorise blast from non-blast cells.

Acknowledgments

Significance statement

Visual abilities have an important role to play in the diagnosis of disease as diagnosis often requires the categorisation of medical images. This study demonstrates that a domain-general high-level visual ability can predict accuracy of medical image categorisation. This is an important first step in establishing a link between visual abilities and diagnostic accuracy. We also contribute to the understanding of the nature of object recognition ability by showing that it relates to category learning. This relationship only appeared in our tasks when continuous feedback on category learning was given. It is plausible that continuous feedback allowed participants to abandon use of a simple rule, and/or to use exemplar comparisons. This suggests new directions to investigate the relationship between categorisation and recognition ability.

Author contributions

All authors read and approved the final manuscript.

Funding

David K. Wilson Chair Research Fund (Vanderbilt University).

Availability of data and materials
Data, analysis code, and materials are available at https://osf.io/yaqs8/. No studies were preregistered.

Declarations

Competing interests
The authors declare no competing interest.

Received: 31 December 2021 Accepted: 18 December 2022
Published online: 01 February 2023

References

Abujudeh, H. H., Boland, G. W., Kaewlai, R., Rabiner, P., Halpern, E. F., Gazelle, G. S., & Thrall, J. H. (2010). Abdominal and pelvic computed tomography (CT) interpretation: Discrepancy rates among experienced radiologists. *European Radiology, 20*(8), 1952–1957. https://doi.org/10.1007/s00330-010-1763-1

Ashby, F. G., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition, 30*(5), 666–677. https://doi.org/10.3758/BF03196423

Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics, 61*(6), 1178–1199. https://doi.org/10.3758/BF03207622

Brady, A. P. (2017). Error and discrepancy in radiology: Inevitable or avoidable? *Insights into Imaging, 8*(1), 171–182. https://doi.org/10.1007/s13244-016-0534-1

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

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

Donald, J. J., & Barnard, S. A. (2012). Common patterns in 558 diagnostic radiology errors. *Journal of Medical Imaging and Radiation Oncology, 56*(2), 173–178. https://doi.org/10.1111/j.1754-9485.2012.02348.x

Ferguson, A., Assadsangabi, R., Chang, J., Raslan, O., Bobinski, M., Bewley, A., Dublin, A., Latchaw, R., & Ivanovic, V. (2021). Analysis of misses in imaging of head and neck pathology by attending neuroradiologists at a single tertiary academic medical centre. *Clinical Radiology, 76*(10), 786.e9-786.e13. https://doi.org/10.1016/j.crad.2021.06.011

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2*(2), 156–168. https://doi.org/10.1177/2515245919847202

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*. 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. *Vision Research, 37*(12), 1673–1682. https://doi.org/10.1016/S0042-6989(96)00286-6

Gergenti, L., & Olympia, R. P. (2019). Etiology and disposition associated with radiology discrepancies on emergency department patients. *The American Journal of Emergency Medicine, 37*(11), 2015–2019. https://doi.org/10.1016/j.ajem.2019.02.027

Gibson, B. R., Rogers, T. T., & Zhu, X. (2013). Human semi-supervised learning. *Topics in Cognitive Science, 5*(1), 132–172. https://doi.org/10.1111/tops.12010

Graber, M., Gordon, R., & Franklin, N. (2002). Reducing diagnostic errors in medicine: What's the goal? *Academic Medicine, 77*(10), 981–992.

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, 29*(3), 866–881. https://doi.org/10.3758/s13423-021-02044-2

Hasan, E., Eichbaum, Q., Seegmiller, A., Stratton, C., & Trueblood, J. S. (2021). *Harnessing the Wisdom of the Confident Crowd in Medical Image Decision-making* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/wkqgs

Itani, M., Assaker, R., Moshiri, M., Dubinsky, T. J., & Dighe, M. K. (2019). Inter-observer variability in the American college of radiology thyroid imaging reporting and data system: In-depth analysis and areas for improvement. *Ultrasound in Medicine & Biology, 45*(2), 461–470. https://doi.org/10.1016/j.ultrasmedbio.2018.09.026

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.

Johansen, M., & Palmeri, T. J. (2002). Are there representational shifts during category learning? *Cognitive Psychology, 45*(4), 482–553. https://doi.org/10.1016/S0010-0285(02)00505-4

Kim, Y. W., & Mansfield, L. T. (2014). Fool me twice: Delayed diagnoses in radiology with emphasis on perpetuated errors. *American Journal of Roentgenology, 202*(3), 465–470. https://doi.org/10.2214/AJR.13.11493

Kolb, T. M., Lichy, J., & Newhouse, J. H. (2002). Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: An analysis of 27,825 patient evaluations. *Radiology, 225*(1), 165–175. https://doi.org/10.1148/radiol.2251011667

Lamoureux, C., Hanna, T. N., Sprecher, D., Weber, S., & Callaway, E. (2021). Radiologist errors by modality, anatomic region, and pathology for 1.6 million exams: What we have learned. *Emergency Radiology, 28*(6), 1135–1141. https://doi.org/10.1007/s10140-021-01959-6

LaTourrette, A., & Waxman, S. R. (2019). A little labeling goes a long way: Semi-supervised learning in infancy. *Developmental Science, 22*(1), e12736. https://doi.org/10.1111/desc.12736

Le Pelley, M. E., Newell, B. R., & Nosofsky, R. M. (2019). Deferred feedback does not dissociate implicit and explicit category-learning systems: Commentary on Smith et al. (2014). *Psychological Science, 30*(9), 1403–1409. https://doi.org/10.1177/0956797619841264

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press.

Lee, C. S., Nagy, P. G., Weaver, S. J., & Newman-Toker, D. E. (2013). Cognitive and system factors contributing to diagnostic errors in radiology. *American Journal of Roentgenology, 201*(3), 611–617. https://doi.org/10.2214/AJR.12.10375

Little, J. L., & McDaniel, M. A. (2015). Individual differences in category learning: Memorization versus rule abstraction. *Memory & Cognition, 43*(2), 283–297. https://doi.org/10.3758/s13421-014-0475-1

Lockwood, P. (2017). Observer performance in computed tomography head reporting. *Journal of Medical Imaging and Radiation Sciences, 48*(1), 22–29. https://doi.org/10.1016/j.jmir.2016.08.001

Morey, R., D., & Rouder, J., N. (2021). *BayesFactor: Computation of Bayes Factors for Common Designs. R package version 0.9.12–4.3.* https://CRAN.R-project.org/package=BayesFactor

Mulder, J., Gu, X., Olsson-Collentine, A., Tomarken, A., Böing-Messing, F., Hoijtink, H., Meijerink, M., Williams, D. R., Menke, J., Fox, J.-P., Rosseel, Y., Wagenmakers, E.-J., & van Lissa, C. (2019). BFpack: Flexible bayes factor testing of scientific theories in R. ArXiv:1911.07728 [Stat]. http://arxiv.org/abs/1911.07728

Ree, M. J., & Earles, J. A. (1992). Intelligence is the best predictor of job performance. *Current Directions in Psychological Science, 1*(3), 86–89.

Richler, J. J., & Palmeri, T. J. (2014). Visual category learning: Visual category learning. *Wiley Interdisciplinary Reviews: Cognitive Science, 5*(1), 75–94. https://doi.org/10.1002/wcs.1268

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

Rudolph, J., Fink, N., Dinkel, J., Koliogiannis, V., Schwarze, V., Goller, S., Erber, B., Geyer, T., Hoppe, B. F., Fischer, M., Ben Khaled, N., Jörgens, M., Ricke, J., Rueckel, J., & Sabel, B. O. (2021). Interpretation of thoracic radiography shows large discrepancies depending on the qualification of the physician—quantitative evaluation of interobserver agreement in a representative emergency department scenario. *Diagnostics, 11*(10), 1868. https://doi.org/10.3390/diagnostics11101868

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

Smithson, C. J. R., Chow, J. K., Chang, T.-Y., & Gauthier, I. (2022). Measuring Object Recognition Ability: Reliability, Validity, and the Aggregate z-score Approach. *Manuscript in Preparation*.

Stewart, N., Brown, G. D. A., & Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(1), 3–11. https://doi.org/10.1037//0278-7393.28.1.3

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

Trueblood, J. S., Holmes, W. R., Seegmiller, A. C., Douds, J., Compton, M., Szentirmai, E., Woodruff, M., Huang, W., Stratton, C., & Eichbaum, Q. (2018). The impact of speed and bias on the cognitive processes of experts and novices in medical image decision-making. *Cognitive Research: Principles and Implications*. https://doi.org/10.1186/s41235-018-0119-2

Wahlheim, C. N., McDaniel, M. A., & Little, J. L. (2016). Category learning strategies in younger and older adults: Rule abstraction and memorization. *Psychology and Aging, 31*(4), 346–357. https://doi.org/10.1037/pag0000083

Waite, S., Scott, J. M., Legasto, A., Kolla, S., Gale, B., & Krupinski, E. A. (2017). Systemic error in radiology. *American Journal of Roentgenology, 209*(3), 629–639. https://doi.org/10.2214/AJR.16.17719

Wang, M. W., & Stanley, J. C. (1970). Differential weighting: a review of methods and empirical studies. *Review of Educational Research, 40*(5), 663–705. https://doi.org/10.2307/1169462

Wilmer, J. B., Germine, L. T., & Nakayama, K. (2014). Face recognition: a model specific ability. *Frontiers in Human Neuroscience*. https://doi.org/10.3389/fnhum.2014.00769

Publisher's Note