Standardized database of 400 complex abstract fractals

Rebecca Ovalle-Fresa[1,2] · Sarah V Di Pietro[2,3,4] · Thomas P Reber[1] · Eleonora Balbi[1] · Nicolas Rothen[1,2]

Abstract

In experimental settings, characteristics of presented stimuli influence cognitive processes. Knowledge about stimulus features is important to manipulate or control the influence of stimuli. To date, there are a lack of standardized data incorporating such information for complex abstract stimuli. Thus, we provide norms for a database of 400 abstract and complex stimuli. Grey-scaled fractals were rated by 512 participants on the stimulus features of abstractness, animacy, verbalizability, complexity, familiarity, favorableness, and memorability. Moreover, 111 participants labeled the fractals, enabling us to calculate indices of naming agreement and modal names. Overall, the results confirmed high abstractness and low verbalizability of the provided stimuli. To establish external validation for selected stimulus features, we evaluated (a) classifier probability of a deep neural network labeling the fractals, negatively correlated with ratings of abstractness and positively with verbalizability and naming agreement; (b) data compression rate of fractal image files, positively correlated with the rating of complexity; and (c) performance of 212 participants in a recognition-memory task, positively correlated with the rating of memorability. The present work fills the gap of a standardized database for abstract stimuli and provides a database with valid norms for abstract and complex stimuli based on ratings and external validation measures. This database can be used to control and manipulate these stimulus features in experimental settings using abstract stimuli. Such a database is essential in experimental research using abstract stimuli for instance to control for verbal influence and strategy or to control for novelty and familiarity.

Keywords Norms · Standardization · Abstract stimuli · Complex stimuli · Visual stimuli

Visual stimuli are essential to experimentally investigate perceptual or cognitive processes. Such stimuli show high individual variation in various features such as perceptual salience and shape on a perceptual level, but also in familiarity or meaningfulness on a conceptual level. These features are known to influence perceptual and cognitive processes (cf. Brodeur et al., 2010). To control or manipulate such influence in experimental designs, several collections with

visual stimuli have been standardized by providing information about individual stimulus features (e.g., Brodeur et al., 2010; Moreno-Martínez & Montoro, 2012; Nishimoto et al., 2010; Snodgrass & Vanderwart, 1980). To investigate cognitive processes without eliciting semantic concepts, abstract visual stimuli are useful, partly because they impede verbal strategy-use. To our knowledge, no standardized database for complex abstract stimuli currently exists. Thus, the aim of the present study was to fill this gap by standardizing a set of 400 complex abstract fractals.

One of the most used databases with concrete, meaningful visual stimuli consist of black-and-white line drawings. They are standardized for the stimulus features familiarity, image agreement, name agreement and visual complexity (Snodgrass & Vanderwart, 1980). Several recent databases with more realistic stimuli provide information about stimulus features (i.e., norms) for photo images (e.g., Brodeur et al., 2010; Brodeur et al., 2014; Moreno-Martínez & Montoro, 2012). For example, Moreno-Martinez and Montoro additionally standardized typical linguistic features of

✉ Rebecca Ovalle-Fresa
  rebecca.ovallefresa@gmail.com

[1] Faculty of Psychology, UniDistance Suisse, Brig, Switzerland

[2] Institute of Psychology, University of Bern, Fabrikstrasse 8, 3012 Bern, Switzerland

[3] Department of Child and Adolescent Psychiatry and Psychotherapy, University Hospital of Psychiatry Zurich, University of Zurich, Zurich, Switzerland

[4] Neuroscience Center Zurich, University of Zurich and ETH Zurich, Zurich, Switzerland

the images, such as age of acquisition, lexical frequency, and manipulability. Brodeur et al. (2010, 2014) provided further norms especially essential for visual stimuli, such as object agreement and viewpoint agreement.

In contrast to concrete stimuli, almost no standardized information for abstract stimuli features exists. A dataset of standardized abstract line drawings (i.e., droodles) was created by Nishimoto et al. (2010). The droodles were standardized along four features which were assessed by rating (1) the appropriateness of given labels, (2) the appropriateness of generated labels, (3) label variability (participants were asked to label the droodles, then the response disagreement across participants was calculated), and (4) relationships between pairs of droodles. Memory performance was then assessed in a cued recall task. Results showed that memory performance was correlated with the ratings for relationships between pairs, but with no other features.

As is the case for concrete line drawings and complex pictures, abstract line-drawings and abstract complex stimuli may also be differentially processed (i.e., Bellhouse-King & Standing, 2007; Brodeur et al., 2014). Thus, the droodles standardized in Nishimoto et al. (2010) can serve as an optimal abstract counterpart to concrete line drawings (i.e., Snodgrass & Vanderwart, 1980) but it might be difficult to compare and contrast them with concrete complex images (i.e., Brodeur et al., 2010). Other studies used more complex abstract stimuli such as photographs of snowflakes of Bentley and Humphreys (1962; e.g., Bellhouse-King & Standing, 2007; Maguire et al., 2003). However, no norms are provided for these databases.

Abstract stimuli are useful in various fields of research. The use of verbal strategies influences visual processing on perceptual and cognitive levels (i.e., Lupyan, 2012; Schooler, 2002). Thus, one possible field of application of abstract stimuli is the investigation of visual processing without verbal influence, such as in mental rotation (e.g., Smith & Dror, 2001) or in memory (e.g., Murphy & Hutchinson, 1982; Ovalle-Fresa et al., 2021; Ward et al., 2013). In such studies, abstract stimuli are assumed to be meaningless. However, when stimuli are intended to be abstract, they can nevertheless be perceived as meaningful to different degrees during perceptual or cognitive tasks (cf. Lupyan, 2012). In one study, geometric shapes varying in the degree of meaningfulness were presented (Voss et al., 2010). Conceptual priming effects and enhanced amplitude of FN400 brain potentials (i.e., electrophysiological frontocentral brain activity presumably related to semantic processing and familiarity) were found for stimuli previously rated as more meaningful, but not when they were rated as meaningless. These results underline the importance of knowledge about individual features of abstract stimuli, such as the degree to which a stimulus elicits meaning. Thus, for research utilizing abstract

stimuli, information about stimulus features such as abstractness or verbalizability are essential.

Another field of application of abstract stimuli is research on the development of familiarity for certain stimuli by repeatedly presenting completely novel stimuli. Here, abstract stimuli are often used to control for novelty across participants (e.g., Chen et al., 2006). Similarly, when the involvement of long-term memory in specific tasks is the target of investigation, abstract stimuli are used to control for pre-existing memory representation of a stimulus (e.g., Nishiyama & Kawaguchi, 2014). Abstract stimuli have also been used to explore the role of familiarity in recognition-memory performance to control for familiarity and novelty (cf. Leynes et al., 2019; Voss et al., 2012). In such studies, it is assumed that abstract stimuli are unfamiliar to participants. Based on the high individual variation of familiarity in concrete stimuli (cf. Brodeur et al., 2010), it can be assumed that perceived familiarity varies between abstract stimuli. However, in contrast to concrete stimuli, no database including norms about perceived familiarity of individual complex abstract stimuli exists.

In the present study, we aimed to respond to the need for a standardized database of abstract and complex stimuli. Thus, we created a new database with images of abstract fractals and provided norm values for each fractal (e.g., Mandelbrot, 1983). To reduce the amount of semantic or verbalizable information in the stimuli from color, luminance, contrast and spatial frequency, the fractals were grey-scaled, approximately isoluminant and similar in spatial frequency. Example fractals are shown in Fig. 1.

First, we assessed norms for seven stimulus features via a rating questionnaire and given names in a labeling questionnaire. Norms for *verbalizability* and *abstractness* are of particular interest in abstract stimuli. Additionally, we assessed norms for the stimulus features *animacy, complexity, familiarity, favorableness and memorability*, which were previously applied in existing databases with concrete visual stimuli (cf. Brodeur et al., 2010). Given names from the labeling questionnaire were used to evaluate indices of naming agreement (modal agreement): the *modal name agreement* (MNA, cf. Brodeur et al., 2010), the *Shannon's index H* (i.e., *H value,* cf. Brodeur et al., 2010; Shannon, 1948), and the *Simpson's Diversity Index* (i.e., *D value*, Simpson, 1949). The indices of naming agreement conceptualize actual verbalizability and thus provide additional insight about verbalizability of the fractals. We expected higher agreement in choosing a specific label for fractals (i.e., higher naming agreement) when fractals are easier to verbalize and lower naming agreement when fractals are harder to verbalize. Accordingly, we expected higher naming agreement for fractals with higher verbalizability ratings and vice versa.

Moreover, the names from the labeling questionnaire allowed us to obtain the *modal name* (i.e., label chosen with
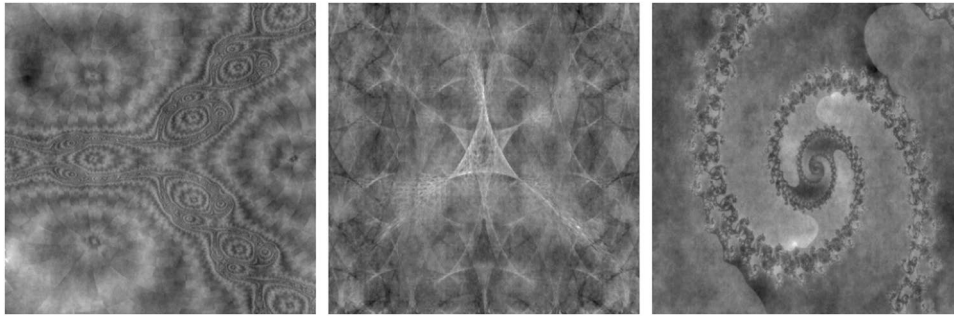
Fig. 1 Examples of fractals provided in the database

highest percentage per fractal). In line with our assumption that the fractals are abstract and thus not contain semantic content per se, we were not specifically interested in the semantic content of the modal names. Rather, the fractals provide more room for semantic interpretation than concrete objects and we assumed that participants associated salient configurational content of the fractals with preexisting semantic concepts to label them. Hence, modal names point to salient visual information of the stimuli. If several fractals result with equal modal names, it allows to detect underlying categorical structures in the database, based on visual similarity.

For external validation of abstractness and verbalizability, we obtained classifier probability resulting from classification of the fractal images with a deep neural network (DNN; cf. King et al., 2019). The applied DNN, AlexNet, is trained to recognize and label objects in images[1] (Krizhevsky et al., 2012). The rationale behind this approach is based on evidence that AlexNet encodes similar information as human brain regions further down the ventral stream (i.e., object recognition in the inferotemporal cortex) (cf. Horikawa & Kamitani, 2017a; Horikawa & Kamitani, 2017b; Jozwik et al., 2017; Mohsenzadeh et al., 2019; Wen et al., 2018). We expected classifier probability of the DNN to correlate negatively with abstractness and positively with verbalizability to externally validate the ratings of the two stimulus features. Correspondence between the classifier probability of the DNN with the indices of naming agreement would further support the use of the DNN for external validation of verbalizability and abstractness.

To externally validate complexity, we calculated the data-compression ratio using a standard zip algorithm for each fractal. This approach takes advantage of the fact that a file with highly structured and repetitive content can be compressed to a lower size than a file with unstructured content

(cf. Casali et al., 2013; Lempel & Ziv, 1976; Sarasso et al., 2014). Implementing the zipping procedure on image files thus addressed the visual complexity of the fractals. We expected a positive correlation between the zip ratio and complexity if the rating provides a valid measure.

To externally validate memorability, we conducted a recognition-memory task in which 60 of the standardized fractals were presented. According to the meta-memory literature, adults are generally good at predicting their memory performance (e.g., Rhodes & Tauber, 2011). Since actual memory performance for the fractals might be of interest in memory research, we used empirical data of performance in a memory task as testable measure. We expected a positive correlation between recognition-memory performance and the rating of memorability to confirm its external validity.

Method

Participants

For the rating questionnaire, we aimed for a minimum of 20 ratings per fractal (i.e., stopping criterion). A total of 793 questionnaires contained data. From this dataset, we first excluded incomplete questionnaires (i.e., not all presented fractals rated, $N = 268$), then duplicates ($N = 5$; i.e., only the first questionnaire from the same participants identified by initials and date of birth was included), participants younger than 16 years ($N = 3$) and, finally, complete test runs by the experimenter ($N = 5$). The final sample for the rating questionnaire included in the analyses thus consisted of 512 German-speaking participants with an average age of 36 years (SD = 14 years, min = 16, max = 78), 352 were women and 461 were right-handed.

For the labeling questionnaire, we set the stopping criterion in terms of time and limited data collection to one month. We obtained a new sample of 313 questionnaires in total, from which we excluded duplicates (i.e., in case of several attempts from the same participant, oldest attempts were removed, $N = 178$, because older attempts often only

---

[1] AlexNet was trained across subsets of the ImageNet (ILS-VRC-2010) with 1000 categories consisting of about 1000 images each (cf. Russakovsky et al., 2015).

contained few responses), incomplete questionnaires (i.e., less than 30% valid responses, where 100% corresponded to at least 200 fractals $N = 24$). The final sample for the labeling questionnaire consisted of 111 German-speaking participants with an average age of 41 years (SD = 12 years, min = 20, max = 72), 85 were women and 26 men. Three were ambidexter and seven left-handed.

For external validation of memorability, we employed a recognition-memory task with a new sample. Sample size was determined by the number of participants who took part in a master thesis project.[2] A total of 381 participants completed the recognition memory task. We first excluded 26 duplicates, then data from 52 participants with a response accuracy lower than 60% and, finally, data from 91 participants completing non-German versions of the task. The final sample for the recognition-memory task included in the analyses consisted of 212 German-speaking participants with an average age of 37 years (SD = 17 years, min = 19, max = 82), 132 were women and 195 were right-handed.

The study was approved by the local ethics committees of the University of Bern and the UniDistance Suisse. All participants were informed about the goal of the experiments and that they could withdraw at any time during the experiment before they consented to participate. For the rating questionnaire and the recognition memory task, participants were recruited in social media channels or via e-mail. For the labeling questionnaire, participants were recruited from the participants-pool of the UniDistance Suisse (including students and non-students) and acquaintances of the authors or research assistants. Psychology students from the UniDistance Suisse received a contribution to course credits for participation.

Materials

An initial pool of 800 fractals was created by an Internet search applying the search-term 'fractal'. The fractals were first resized to 380 x 380 pixels and then grey-scaled with Irfan View (Version 4.41; see www.irfanview.com). Next, spatial frequency and luminance were averaged across the initial pool by means of matching luminance and spatial frequency with "The SHINE toolbox" (Willenbockel et al., 2010). This procedure was applied across each whole image (including fore and background) in ten iterations and included iteratively structural similarity (SSIM) index optimization. For the current standardization study, 400 fractals were randomly chosen from the initial pool. For external validation with the recognition task, 60 fractals were randomly

chosen from the 400 fractals and randomly assigned to two lists. A zip-folder with the 400 fractals is available for download on OSF (direct link to the stimuli: https://osf.io/c8atx/).

The rating questionnaire was programmed and presented online in SosciSurvey (www.soscisurvey.de). Each fractal was accompanied with seven statements (in German) related to the stimulus features abstractness ("The image is abstract. It seems to depict something, that doesn't exist."), animacy ("The image is animated. It seems to depict something alive [compared to something lifeless]"), complexity ("The image is complex. It has a lot of details and a complicated structure."), familiarity ("The image is familiar. It seems familiar, as if I would already know it."), favorableness ("The image is beautiful. I like it."), memorability ("The image is memorable. I can imagine that I could remember it well."), verbalizability ("The image is verbalizable. I can name the stimulus with a clear term."). The instruction was to rate each statement on a Likert scale from 1 ("not at all") to 7 ("completely") with the corresponding radio button.

The labeling questionnaire was programmed and presented online with lab.js (www.lab.js.org/; Henninger et al., 2021). Each fractal was accompanied by an empty text field. Participants were instructed to enter for each fractal the first name that came to their mind, but to refrain from entering "fractal", "pattern", or "abstract picture".

The recognition task was programmed and presented online with lab.js (Henninger et al., 2021). It consisted of a learning and a recognition phase. During the learning phase, either the 30 fractals from list A or list B, respectively, were presented. List selection was randomized between participants. The instructions were to rate symmetry of each fractal with a keypress ("Y" = symmetric, "N" = asymmetric) and to memorize the fractals for subsequent recognition. During the recognition phase, all fractals from both lists A and B were presented. The recognition question was to indicate by keypress if a fractal was old ("Y"; i.e., presented during learning) or new ("N"; i.e., not presented during learning). The confidence rating was assessed with the respective number key ("1" = guess, "2" = relatively confident, "3" = very confident).

Procedure

The rating questionnaire, the labeling questionnaire, and the recognition task were conducted in three independent online studies and based on different participant samples. All studies began with an information page where participants confirmed their consent and completed a short demographic questionnaire. The recognition task was followed by an additional questionnaire which is not relevant for the present study and hence not further discussed.

In the rating questionnaire, 20 out of the 400 fractals were randomly selected for each participant and presented

---

[2] Due to the correlational main hypothesis (not related to the present project), the aim of the master thesis was to test 320 participants or as many as possible over the duration of 2 months.

in randomized order. A trial consisted of the presentation of one fractal accompanied by the seven statements about the stimulus features. Positions of the statements were randomized in each trial. There was no time restriction and participants could not click on the "next" button before all statements were rated.

In the labeling questionnaire, participants chose to label either 200 or 400 fractals. When 200 fractals were chosen, they were randomly selected out of the 400 fractals without replacement. Presentation order was always randomized. One fractal was presented per trial, accompanied by a text field. Participants had to enter a suitable name into the text field and could move to the next trial by clicking on the "next" button or pressing the "enter" key. When no name was entered, the next trial was initiated after 15 s.

The recognition task started with the learning phase followed by the recognition phase. During the learning phase, 30 fractals were presented in randomized order. A trial started with the central fixation cross displayed for 1 s, followed by the fractal presented at the center of the screen for 1 s. Next, the learning question (i.e., is the fractal symmetric or not) was presented for a maximum of 10 s, or until participants made a keypress ("Y" = yes, "N" = no). During the recognition phase, the 30 old fractals (i.e., presented during learning) and 30 new fractals (i.e., not presented during learning) were presented in a random temporal sequence. Each trial started with a central fixation cross presented for 1 s. After that, the fractal appeared at the center of the screen for 1 s. Next, the recognition question was presented, and participants indicated by keypress if the fractal was old ("Y") or new ("N"). After keypress or a maximum of 10 s, the confidence rating was shown and participants pressed a key to indicate how confident they were about the response in the recognition question ("1" = guess, "2" = relatively confident, "3" = very confident). The next trial was initiated by keypress or after 10 s elapsed.

Analyses

The alpha-level was set to .05 for all statistical analyses. All correlational analyses are based on Pearson correlations if not mentioned otherwise. T tests and Mann–Whitney U tests are two-tailed. As a measure of effect size, we report Cohen's d or r. We interpreted the effect sizes r according to Cohen (1988): an r larger than .1 was interpreted as low effect, between .3 and .5 as medium effect and greater than .5 as large effect. R-packages and Matlab toolboxes are reported in the corresponding parts of the analyses. Scripts are available via Open Science Framework (OSF), https://doi.org/10.17605/OSF.IO/CKFMV or directly on Bitbucket, https://bitbucket.org/refresa/fractals_standardization/src/master/.

Labeling questionnaire To obtain maximal consensus between responses of the labeling questionnaire, we applied an automated text-mining approach. The exact preprocessing steps are listed in the supplemental information (SupplementalInformation.pdf). We replaced different expressions of "have no name" or "no image presented" (in case of technical problems) with the unique labels "noidea" and "noimg", respectively. No responses, "noidea" and "noimage" responses were regarded as invalid responses, participants with more than 70% invalid responses were removed from the analyses (100% relating to at least 200 fractals). We preprocessed the valid responses by removing stop words (i.e., words with only grammatical but no contextual function) and unifying the spelling. Spelling correction was completed by comparing each word with the Leipzig Corpora (gsw-ch_web_2017, including Swiss German words, deu_wikipedia_2016 and deu-com_web_2018 including German words; (cf. Goldhahn, Eckart, & Quasthoff, 2012), https://wortschatz.uni-leipzig.de/de). The first preprocessed word of a response was then used as label. We considered singular and plural forms of the same word, such as "flower" and "flowers", as two different labels, because they point to perceptually distinct identities of the fractals.

We then used the labels to calculate indices of naming agreement and to evaluate the modal name per fractal. For each fractal, the *MNA* was calculated as percentage of participants using a specific label after excluding invalid responses (cf. Brodeur et al., 2010). The highest percentage per fractal was used as MNA, higher MNA thus indicates higher naming agreement. The *H value* (Brodeur et al., 2010; Shannon, 1948), which in contrast to the MNA considers all alternative labels and their frequencies, was calculated per fractal according to the equation $H = \sum_{i=1}^{k} P_i \log_2\left(\frac{1}{P_i}\right)$. In this equation, k represents the number of different labels provided for a specific fractal after excluding invalid responses and P_i represents the proportion of responses for a specific label of a given fractal without invalid responses. The H value ranges from 0 to 1: if all participants chose the same label for a fractal, its H value is 0, if almost all participants chose the same labels and only few alternative labels are chosen with low frequencies, the H value approaches 0. The H value is 1 when a fractal is labeled with two equally frequent labels. Moreover, the H value decreases for fractals with less and less frequently chosen alternative labels and increases for fractals with more alternative labels and more frequently chosen alternative labels. Lower H values thus correspond with higher naming agreement. We also calculated the Simpson's Diversity Index (i.e., *D value*, Simpson, 1949) per fractal, which again considers the alternative responses, according to the equation $D = \sum \frac{n(n-1)}{N(N-1)}$. In this equation, n is the total number of responses for a specific label of a particular fractal and N is

the total number of responses across all labels for a particular fractal. The *D* value ranges from 0 to 1 and can be interpreted as probability that two randomly picked labels are the same. Lower *D* values thus correspond to lower naming agreement. An index of 0 reflects that every response for a fractal represented a unique label. An index of 1 reflects that all responses represented the same label. The label with the highest MNA (the label chosen with highest percentage) was then identified as *modal name.* In case of several labels with equal MNA, we chose the label with highest appearance across all fractals as modal name to optimize categorical information from similar modal names. In the norm table (file fractals_norms.csv on OSF, https://doi.org/10.17605/OSF.IO/CKFMV), the indices and the MNA are provided for each fractal. In addition, the labels with equal MNA but lower appearance in the whole database are listed as "alternative modal names".

DNN classifier probability To externally validate abstractness and verbalizability, we obtained classifier probability resulting from classification of the fractal images with AlexNet, a DNN trained to recognize and label objects in images (Krizhevsky et al., 2012). AlexNet consists of eight layers, starting with five convolutional layers followed by three fully connected layers. Critical for our purpose is the output of the last fully connected layer, which is fed into a 1000-way softmax regression. AlexNet selects the label with the highest log-probability across all 1000 class labels as 'correct label' and we use this highest log-probability value as classifier probability (cf. Krizhevsky et al., 2012). Probability values range from 0 to 100. A probability of 100 would thus reflect that this class label was selected with a maximum probability, whilst the other class labels were selected with null probability. A probability of 70, for example, would reflect that this specific class label was selected with a relatively high probability, whilst some of the other labels were selected with a total probability of 30 (i.e., the uncertainty of 100–70 is distributed across the other 999 class labels).

Zip ratio To externally validate complexity, we calculated the zip ratio for each fractal as the compressed file size of each fractal image, divided by the file size of the uncompressed fractal image. A zip ratio of one means that a fractal could not have been compressed because of its highly unstructured content, indicating high complexity. A zip ratio near zero indicates that a fractal could have been compressed to a high extend, due to redundant information in the file, indicating low complexity. To this end, we zipped the fractal images as .bmp files using the R-package zipr (Csárdi et al., 2020) with the highest compression level of nine. Eventually, the deflate algorithm, an improvement of the Lempel-Ziv-77 algorithm (Ziv & Lempel, 1977), was applied.

Table 1 Descriptive statistics of the averaged ratings and of indices of naming agreement (across 400 fractals)

Questionnaire	Feature / Index	M	SD	Min	Max
Rating	Abstractness	4.65	0.62	2.60	5.92
	Animacy	3.83	0.70	2.08	6.00
	Complexity	4.56	0.69	2.28	6.20
	Familiarity	3.12	0.68	1.70	5.33
	Favorableness	3.74	0.69	2.04	5.30
	Memorability	3.85	0.76	1.61	5.83
	Verbalizability	3.18	0.77	1.62	5.92
Labeling	MNA	11.4	8.60	2.33	81.10
	H value	1.92	0.19	0.48	2.19
	D value	0.03	0.05	0.001	0.66

Modal name agreement is abbreviated with MNA. *H* value is also known as Shannon's index, *D* value is also known as Simpson's diversity index

Recognition memory Performance in a recognition memory task served to externally validate memorability. We thus calculated *d*-prime for a given fractal across subjects. *D*-prime was obtained by calculating the normalized hit rate (HR) minus the normalized false alarm rate (FAR). To avoid HR of 1 and FAR of 0, we calculated corrected hit rates, HR = (Hits + 0.5) / (total old + 1), reflecting the rate of fractals correctly recognized as old, and false alarm rates, FAR = (FAs + 0.5) / (total new + 1), reflecting the rate of fractals incorrectly identified as old (cf. Snodgrass & Corwin, 1988). Higher *d*-prime values indicate that a given fractal was better recognized. We also calculated criterion *C* for each fractal, *C* = (normalized HR + normalized FAR) / 2, to obtain a measure of response bias. A criterion *C* of zero indicates no bias, negative values reflect a bias to "no" responses and positive values to "yes" responses.

Results

In the current article, we provide an overview of the norms from the rating and the labeling questionnaires, as well as of external validation measures, as averages across all 400 fractals. Stimulus-specific norms are provided in the file *fractals_norms.csv*, which can be downloaded from OSF, https://doi.org/10.17605/OSF.IO/CKFMV.

Rating questionnaire

Descriptive statistics of the averaged ratings per stimulus feature are shown in Table 1. Each fractal was rated across seven features based on a Likert scale from 1 ("not at all") to 7 ("completely"). The highest mean scores were observed for abstractness (M = 4.65, SD = 0.62) and complexity (M =

4.56, SD = 0.69), confirming the high abstractness and complexity of the fractal stimuli in the database. Complementary, verbalizability (M = 3.18, SD = 0.77) and familiarity (M = 3.12, SD = 0.68) were the stimulus features with the lowest mean ratings, indicating that the fractals are difficult to label and depict novel (unfamiliar) content. The density distributions and boxplots of the averaged ratings are shown in Fig. 2a. Approximately normal density distributions for animacy, favorableness and memorability were observed. However, density distribution of abstractness and complexity was skewed towards higher scores, familiarity, and verbalizability towards lower scores.

Next, we calculated inter-item correlations (Pearson's r) to investigate how ratings for individual stimulus features were related. Correlation coefficients with their 95% confidence intervals are shown in Table 2, the corresponding scatterplots can be found in Supplemental Fig. 1 (SupplementalInformation.pdf). The analyses revealed weak to strong correlations for all stimulus features, except for the negligible correlation between favorableness and complexity, $r(400) = -.04$, $p = .404$. Abstractness was correlated with all six other features, whereas the strongest correlations were observed with verbalizability, $r(400) = -.73$, $p < .001$, and with familiarity, $r(400) = -.68$, $p < .001$, indicating that fractals rated as more abstract were perceived as less verbalizable and less familiar. For verbalizability, we observed strong correlations with familiarity, $r(400) = .84$, $p < .001$, favorableness, $r(400) = .60$, $p < .001$, and memorability,
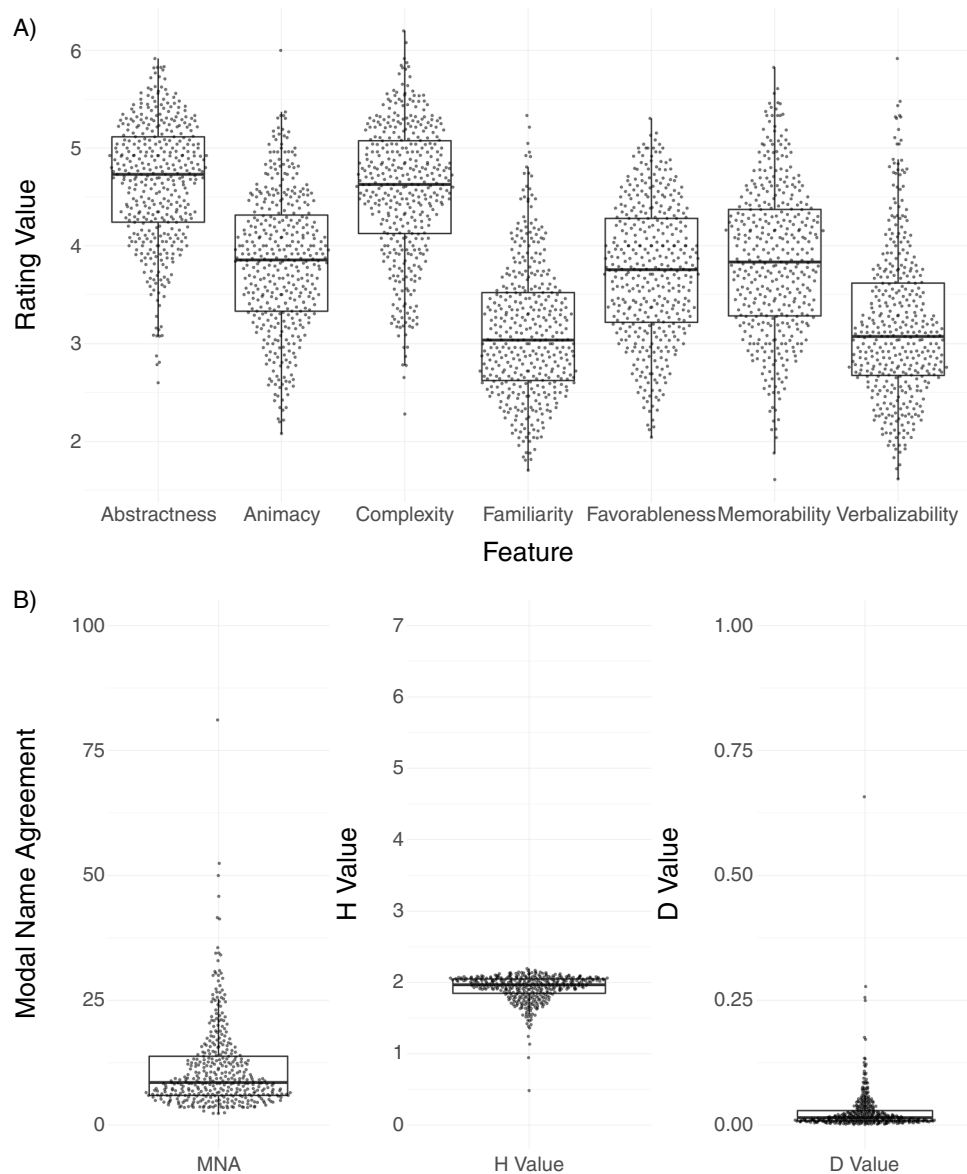

Fig. 2 **a** Density distributions and boxplots from the ratings grouped by the seven stimulus features. Individual datapoints represent averaged ratings per fractal. **b** Density distributions and boxplots for the naming agreement indices. Individual datapoints represent index values per fractal

Table 2 Correlations of the ratings for stimulus features (1–7) and indices of naming agreement (8–9)

Questionnaire	Variable	1	2	3	4	5	6	7	8	9
Rating	1. Abstractness									
	2. Animacy	−.40** [−.48, −.31]								
	3. Complexity	.37** [.29, .46]	.14** [.04, .23]							
	4. Familiarity	−.68** [−.73, −.62]	.41** [.32, .49]	−.27** [−.36, −.18]						
	5. Favorableness	−.41** [−.49, −.33]	.51** [.44, .58]	−.04 [−.14, .06]	.68** [.63, .73]					
	6. Memorability	−.55** [−.61, −.47]	.29** [.20, .38]	−.44** [−.51, −.36]	.75** [.70, .79]	.65** [.59, .70]				
	7. Verbalizability	−.73** [−.78, −.69]	.39** [.31, .47]	−.37** [−.45, −.28]	.84** [.81, .87]	.60** [.54, .66]	.81** [.77, .84]			
Labeling	8. MNA	−.24** [−.33, −.14]	.15** [.05, .24]	−.17** [−.26, −.07]	.29** [.19, .37]	.18** [.08, .27]	.35** [.26, .43]	.38** [.29, .46]		
	9. H value	.33** [.24, .40]	−.18** [−.27, −.08]	.22** [.13, .31]	−.35** [−.43, −.26]	−.23** [−.32, −.14]	−.41** [−.49, −.33]	−.45** [−.54, −.36]	−.90** [−.91, −.88]	
	10. D value	−.30** [−.39, −.21]	.17** [.07, .26]	−.22** [−.31, −.12]	.34** [.25, .43]	.22** [.13, .32]	.42** [.33, .50]	.44** [.36, .51]	.96** [.95, .96]	−.97** [−.98, −.97]

Modal name agreement is abbreviated with MNA. Spearman's *rho* is reported for correlations including indices (8–10), otherwise the values represent Pearson's *r* (1–7). *Values in square brackets indicate the 95% confidence interval for each correlation.* ** indicates *p* < .01
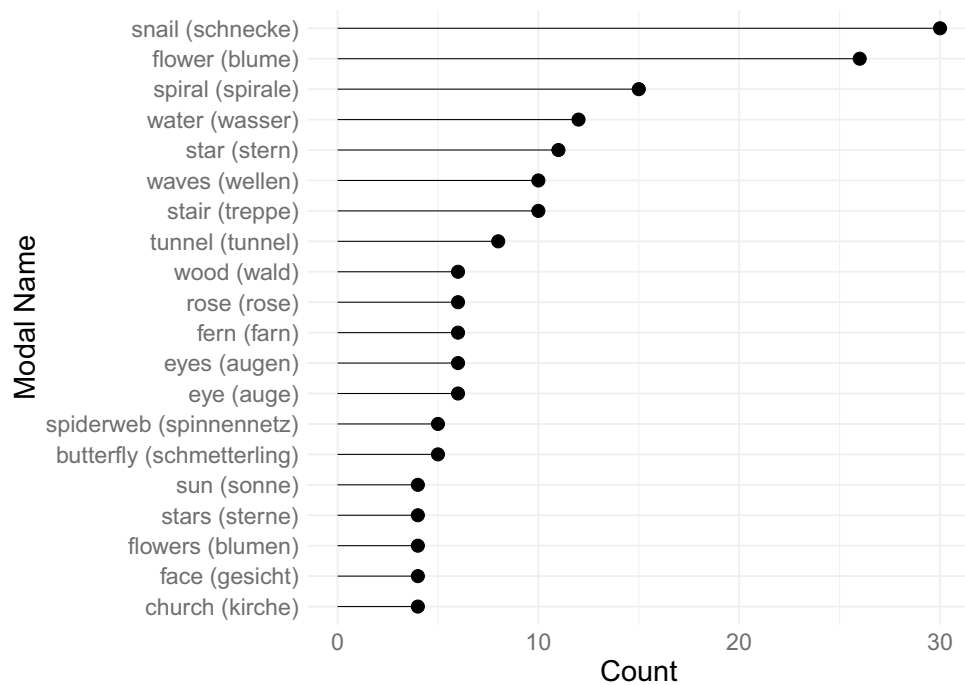
Fig. 3 The 20 most frequent modal names in the database. Modal names were identified as the label with highest MNA (i.e., given with highest percentage) per fractal. Note that we considered singular and plural forms of the same word as two different labels, because these point to visually distinct identities of the fractals. Modal names on the y-axis reflect English translations with the German word in parentheses

r(400) = .81, p < .001, indicating that fractals with higher verbalizability ratings were also rated as more familiar, more favorable, and more likely to be remembered.

Reliability for the features was estimated with an internal consistency analyses using a permutation split-half correlation procedure by means of the R package splithalf (Parsons, 2020). Using 5000 random splits, based on the responses per fractal and feature (i.e., splitting across the participants), the Spearman-Brown corrected reliability estimates were r_{SB} 0.67, 95% CI [0.63, 0.72] for abstractness, r_{SB} 0.79, 95% CI [0.76, 0.81] for verbalizability, r_{SB} 0.79, 95% CI [0.77, 0.72] for complexity, r_{SB} 0.74, 95% CI [0.70, 0.77] for familiarity, r_{SB} 0.74, 95% CI [0.70, 0.78] for favorableness, r_{SB} 0.73, 95% CI [0.69, 0.77] for animacy, r_{SB} 0.80, 95% CI [0.77, 0.83] for memorability. Overall, these values indicate good internal consistency for all features.

Labeling questionnaire

In the labeling questionnaire, we assessed a total of 33883 valid labels. Each fractal was presented on average 92 times (median = 92, min = 82, max = 101), resulting in an average of 84.7 valid labels (median = 85, min = 69, max = 96) and an average of 7.32 invalid labels (i.e., no responses, no idea, no image, median = 7, min = 0, max = 23) per fractal. A total of 8655 different labels were used across all fractals.

The most common labels across all fractals were snail ("Schnecke", N = 551), flower ("Blume", N = 521), and spiral ("Spirale", N = 445). Correspondingly, from totally 178 different modal names (i.e., the most often chosen label for a fractal) in the database, the most frequent modal names were snail (N = 30), flower (N = 26), and spiral (N = 15). The twenty most frequent modal names are depicted in Fig. 3. As groups of fractals resulted in equal modal names, the modal names can be used to detect underlying categorical structure in the database.

Descriptive statistics for the naming agreement indices across all fractals are provided in Table 1, density distributions and boxplots are shown in Fig. 2b. All indices indicated low naming agreement on average. As visible in Fig. 2b, one fractal (SHINEd_fractal_1738) resulted in remarkably higher naming agreement than the others. All results reported in this study include that fractal but were virtual identical when the fractal was excluded.

Naming agreement for the fractals in our database was lower compared to the pictures of 467 meaningful objects in the BOSS database reported in Brodeur et al. (2010): Mann–Whitney U tests indicated significantly higher MNA (indicating lower naming agreement) for the fractals (median = 8.52, min = 2.33, max = 80 81.11) than for the objects in the BOSS database (median = 59, min = 13, max = 100), U = 2999.5, z = 24.59, p < 0.001, r = 0.84. The

same analysis revealed significantly higher H values (indicating lower naming agreement) for the fractals (median = .97, min = 0.48, max = 2.19) than the meaningful objects in the BOSS database (median = 1.58, min = 0, max = 6.04), $U = 111305$, $z = 4.87$, $p < 0.001$, $r = 0.17$. These results indicate that, compared to pictures of meaningful objects, the abstract fractals in the database receive lower agreement on the labels.

Next, we were interested if the indices of naming agreement were related to the responses of the rating questionnaire. As verbalizability, abstractness and familiarity presumably reflect how well a fractal can be named, we expected greater naming agreement to be substantially related to higher verbalizability, lower abstractness, and higher familiarity. Higher MNA and H values indicate higher naming agreement, we thus expected a positive correlation for these two indices with verbalizability and familiarity and a negative correlation with abstractness. As higher D values indicate lower naming agreement, we expected D values to correlate negatively with verbalizability and familiarity and positively with abstractness. Coefficients of Spearman correlations with their 95% confidence intervals are shown in Table 2, the corresponding scatterplots are provided in Supplemental Fig. 2 (SupplementalInformation.pdf). The analyses revealed medium correlations between the different indices and verbalizability, $rho_{MNA}(400) = .38$, $rho_{H value}(400) = -.44$, $rho_{D value}(400) = .44$, all $ps < .001$, and small to medium correlations between the indices and abstractness, $rho_{MNA}(400) = -.24$, $rho_{H value}(400) = .32$, $rho_{D value}(400) = -.29$, all $ps < .001$, and between the indices and familiarity, $rho_{MNA}(400) = .32$, $rho_{H value}(400) = -.39$, $rho_{D value}(400) = .17$ all $ps < .001$. As expected, the correlations indicate higher naming agreement for fractals with higher ratings for verbalizability and familiarity and for fractals with lower ratings for abstractness. The same analyses further revealed small to medium correlations between indices of naming agreement and all other ratings, ranging from small effects for animacy ($rhos \geq .15$) to medium effects for memorability ($rhos \geq .35$), cf. Table 2.

## External validations

DNN classifier probability We computed DNN classifier probability with the AlexNet on https://pjreddie.com/darknet/imagenet/ (Krizhevsky et al., 2012). For comparison with meaningful stimuli, we additionally estimated classifier probability of AlexNet for two sets of naturalistic object images: we selected 930 grey-scaled stimuli from the BOSS database (Brodeur et al., 2010, 2014) and 2400 colored stimuli from the THINGS database (Hebart et al., 2019). For comparison with naturalistic but meaningless stimuli, we used stimuli from the Brodatz's texture database (Abdelmounaime & Dong-Chen, 2013). Here, we used the

112 normalized grey-scaled texture images. We expected lower classifier probability for the fractals than for the other stimuli. Distribution densities and boxplots of classifier probability of a DNN are depicted in Fig. 4a. Due to deviations from normal distribution, we report non-parametric statistics for analyses including the classifier probability values. The DNN labeled the 400 fractals with a mean classifier probability of 22.52 (median = 17.5, SD = 17.76, min = 3.39, max = 96.55). A Mann–Whitney U test indicated that classifier probability for the labeling of fractals were significantly lower than for the naturalistic objects from the BOSS database (median = 33.18, min = 6.35, max = 99.99), $U = 97246$, $z = 13.82$, $p < .001$, $r = 0.38$, and from the THINGS database (median = 38.73, min = 4, max = 100), $U = 229184$, $z = 16.76$, $p < .001$, $r = 0.32$. Classifier probability for fractals was also significantly lower than for the naturalistic textures from the Brodatz database (median = 36.39, min = 3.86, max = 98.94), $U = 12530$, $z = 7.13$, $p < .001$, $r = 0.32$. These results indicate that the fractals are significantly more difficult to label for a DNN than images of naturalistic objects and textures. Classifier probability also differed between the two object databases, $U = 989118$, $z = 5.10$, $p < .001$, $r = 0.09$, with lower classifier probability for the BOSS than for the THINGS images.

We were interested in a potential relation between DNN classifier probability and naming agreement. If higher DNN classifier probability was related to indices pointing to greater naming agreement, the DNN and the human sample would identify similar fractals as harder to label and similar fractals as easier to label. We observed small Spearman correlations between the DNN classifier probability and the indices, with $rho_{MNA}(400) = .12$, $p = .015$, $rho_{H value}(400) = -.15$, $p = .002$, and $rho_{D value}(400) = .15$, $p = .002$, indicating that fractals classified with lower classifier probability also resulted in lower naming agreement and vice versa. Scatterplots of the correlations are shown in Supplemental Fig. 3 (SupplementalInformation.pdf). Only three fractals (< 1%) received the same label from the DNN and the human sample, indicating that virtually no overlap existed between the labels given by the DNN and the modal names resulting from the labeling questionnaire.

Next, we aimed to validate the ratings of abstractness and verbalizability. To do this, we calculated Spearman correlations between DNN classifier probability and the stimulus features rated in the questionnaire. Scatterplots can be found in Supplemental Fig. 3 (SupplementalInformation.pdf). We observed the expected positive correlation with verbalizability, $rho(400) = .26$, $p < .001$, and negative correlation with abstractness, $rho(400) = -.10$, $p = .042$. These results indicate that fractals with high ratings for verbalizability were also classified with high probability by the DNN and fractals with lower abstractness ratings were classified with lower probability. The same analyses with the other stimulus

A)

B)
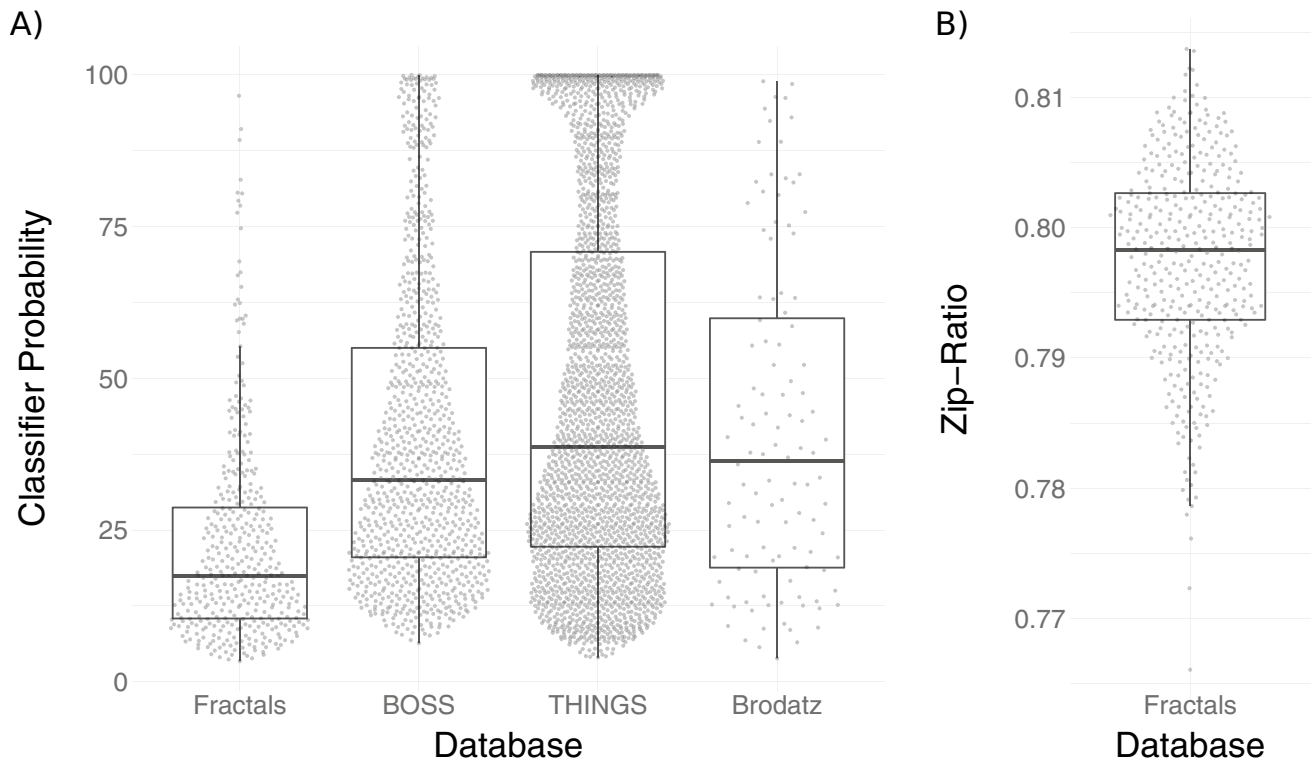
Fig. 4 **a** Density distributions and boxplots for classifier probability of the deep neural network (AlexNet) grouped by database. Fractals include the 400 fractals standardized in the present study. The BOSS and THINGS databases include naturalistic images of objects (i.e.,

Brodeur et al., 2010; Hebart et al., 2019), the Brodatz database of naturalistic images of textures (Abdelmounaime & Dong-Chen, 2013). **b** Density distribution and boxplot for the zip ratio of the fractal image files. Individual datapoints in both figures reflect individual stimuli

features revealed weak correlations with classifier probability, all $rhos(400) \geq .15$, all $ps \leq .003$, except for animacy, $rho(400) = - .01$, $p = .874$ (cf. Fig. 5a).

Zip ratio The zip ratio was used as external validation of complexity. The boxplot and density distribution of the zip ratios are depicted in Fig. 4b. The average zip ratio across all fractals was .80 (SD = .01), indicating predominantly complex visual structures in the fractals. Visual inspection confirmed that fractals with lower zip ratio consisted of larger contiguous areas of the surface filled in the same shade. Complementary, fractals with higher zip ratio generally consisted of smaller contiguous areas of the surface filled in the same shade.

To externally validate the rating for complexity, we calculated Pearson correlations with the zip ratio. Scatterplots are shown in Supplemental Fig. 4 (SupplementalInformation. pdf). We were mainly interested in the relationship between the zip ratio and complexity, where we observed a correlation as predicted, $r(400) = .25$, $p < .001$, indicating higher zip ratios in fractals with higher ratings for complexity. The zip ratio further correlated negatively with abstractness, $r(400) = - .17$, $p < .001$, and positively with familiarity, $r(400) = .13$, $p = .008$, and animacy, $r(400) = .13$, $p = .009$,

indicating that fractals with higher zip ratios revealed lower ratings for abstractness and higher ratings for familiarity and animacy. No significant correlation with favorableness, memorability and verbalizability was observed, all $rs(400) \leq .09$, all $ps \geq .074$, indicating that the ratings for these stimulus features were not significantly related to the zip ratio (cf. Fig. 5b).

Recognition memory To obtain a measure for external validation of memorability, we used 60 of the fractals as stimuli in a yes/no recognition-memory task. Here, we report d-prime and criterion C, reflecting memory performance and response bias, respectively. Further analyses on HR and FAR are provided in the Supplemental Information (SupplementalInformation.pdf, pp. 7). Correlation coefficients and confidence intervals including memory measures are shown in Table 3 and Fig. 5c, the scatterplots can be found in Supplemental Fig. 5 (SupplementalInformation.pdf).

Averaged d-prime across fractals was 1.13 (SD = 0.51), indicating more correctly recognized fractals than false alarms. To externally validate the rating of memorability, we calculated Pearson correlations. We were mainly interested in the correlation between the rating for the stimulus feature memorability and actual memory performance (d-prime),

A)

B)

C)
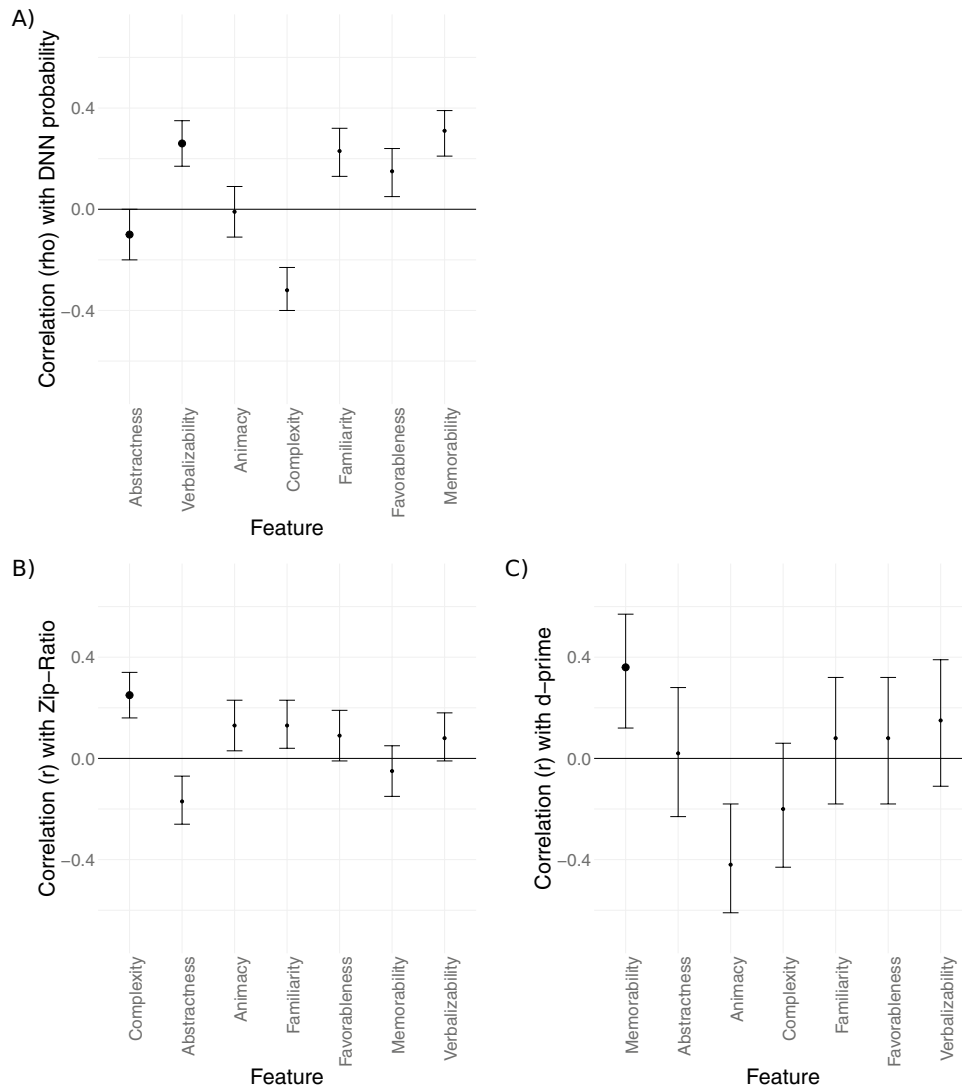
Fig. 5 Correlation coefficients of the ratings for abstractness and verbalizability (***a***), complexity (***b***), and memorability (***c***) with the ratings for the stimulus features. The features of interest for external valida-tions are represented with larger points. *Error bars* represent 95% confidence intervals

where we expected a positive correlation. Our prediction was confirmed with a positive correlation between memorability and memory performance, $r(60) = .36$, $p = .004$, indicating that higher ratings for memorability are indeed associated with higher performance in a recognition task. Furthermore, memory performance correlated negatively with animacy, $r(60) = -.42$, $p = .001$, indicating higher memory performance for fractals with lower animacy rating, but not with the rest of the stimulus features, all rs(60) $\leq .10$, all ps $\geq .126$.

Averaged criterion C was 0.27 (SD = 0.37), indicating a slight overall tendency for "yes" (i.e., "old") responses. No correlation between criterion C and memorability was observed, $r(60) = -.23$, $p = .078$. Criterion C was negatively correlated with familiarity, $r(60) = -.28$, $p = .033$, animacy,

$r(60) = -.29$, $p = .024$, and complexity, $r(60) = -.29$, $p = .026$. These results indicate that fractals with higher ratings for familiarity, animacy and complexity were slightly associated with higher tendency for "no" responses and fractals with lower ratings in these stimulus features with a lower tendency for "yes" responses.

Discussion

The aim of the present study was to create a standardized database of complex abstract stimuli. We provide 400 grey-scaled fractals with norms which are specifically relevant for meaningless stimuli, namely verbalizability and

Table 3 Means, standard deviations, and correlations with confidence intervals for memory measures and ratings for the stimulus features (averaged across the 60 fractals)

Feature	M	SD	1	2
1. *d*-prime	1.13	0.51		
2. Criterion *C*	0.27	0.37	.28*	
			[.02, .50]	
3. Memorability	3.90	0.65	.36**	− .23
			[.12, .57]	[− .46, .03]
4. Abstractness	4.56	0.63	.02	.11
			[− .23, .28]	[− .15, .36]
5. Animacy	3.82	0.71	− .42**	− .29*
			[− .61, − .18]	[− .51, − .04]
6. Complexity	4.50	0.68	− .20	− .29*
			[− .43, .06]	[− .50, − .04]
7. Familiarity	3.19	0.68	.08	− .28*
			[− .18, .32]	[− .49, − .02]
8. Favorableness	3.70	0.61	.08	− .24
			[− .18, .32]	[− .47, .01]
9. Verbalizability	3.19	0.71	.15	− .19
			[− .11, .39]	[− .42, .07]

M and SD are used to represent mean and standard deviation, respectively. The correlation values represent Pearson's *r*. *Values in square brackets* indicate the 95% confidence interval for each correlation. * indicates $p < .05$. ** indicates $p < .01$

abstractness. Additionally, we established norms for stimuli features such as animacy, complexity, familiarity, favorableness, and memorability, as well as modal names and indices of naming agreement, which have been assessed in previous research with meaningful stimuli (e.g., Brodeur et al., 2010). We complemented the norms with external validation measures, obtained by computational methods to estimate verbalizability/abstractness and complexity of the fractals and by empirical data in a separate memory experiment.

Low verbalizability and high abstractness of the fractals database was confirmed in several analyses. First, we observed the expected high ratings for abstractness and low ratings for verbalizability, indicating that the provided database includes abstract stimuli which are difficult to verbalize overall. Second, indices of naming agreement revealed generally very low naming agreement for the fractals, which was remarkably lower than naming agreement for meaningful stimuli in the BOSS database (Brodeur et al., 2010). We observed medium correlations between naming agreement and abstractness as well as low to medium correlations between naming agreement and verbalizability because fractals with higher naming agreement were rated higher in verbalizability and lower in abstractness. The participants were thus able to differentiate between different degrees of abstractness and verbalizability, which were also reflected in naming agreement. These values now facilitate

the determination of the exact degree of abstractness and verbalization for each stimulus.

Classifier probability values from the DNN AlexNet (cf. King et al., 2019) served for external validation of the abstractness and verbalizability ratings. Mean classifier probability was significantly lower for the fractals than for meaningful stimuli (i.e., photographs of objects and textures), indicating that the DNN could label meaningful stimuli with higher probability compared to the fractals in the present database. Comparing the labels given by the DNN with the modal names of the human sample revealed virtually no overlap. An explanation is that the labels available to AlexNet are only a small subset of the labels available to humans. However, classifier probability was related with the indices for naming agreement from the labeling questionnaires, indicating that fractals classified with lower probability also received lower naming agreement. Although the effect sizes were small, this relation supports the use of the DNN classifier probability as external validation measure for verbalizability and abstractness. The external validation with the DNN probability value was confirmed for both features, although numerically better reflecting verbalizability (approaching a medium effect, *rho* = .26) than abstractness (low effect, *rho* = − 0.10). This is not surprising, as the goal of the DNN is to name (i.e., to verbalize) the content of the images, rather than to rate the abstractness of image contents. Taken together, the present database provides reliable information about the stimulus features abstractness and verbalizability of individual fractals. This is especially important when testing with abstract stimuli, as previous research revealed that participants tend to induce meaning into meaningless stimuli, such as minimalistic geometric shapes (cf. Lupyan, 2012; Voss et al., 2010).

Furthermore, the average ratings for complexity, as for abstractness, were high compared to the other features. Thus, the fractals in the database were generally perceived as complex visual stimuli. Complexity ratings were externally validated by means of zip ratios, as complex files can be compressed to a lower level than less complex files (cf. Casali et al., 2013; Sarasso et al., 2014). Note that lower zip ratios indicate that a file was compressed to a higher extend. Visual inspection revealed that fractals with lower zip ratios generally consisted of larger contiguous areas of the surface in the same shade, a visual structure that might be perceived as less complex. Complementary, fractals with higher zip ratios generally had more fine-grade perceptual structures with smaller contiguous areas of the surface in the same shade, a visual structure that might be perceived as more complex. Accordingly, the zip ratio was positively correlated with complexity, approaching a medium effect (*r* = .25), indicating that fractals with higher complexity ratings indeed consist of a more complex structure. Information about complexity of individual visual stimuli is of importance in experimental research, as it influences for instance

early perception (e.g., Bradley et al., 2007) and also higher cognitive performance such as memory (e.g., Eng et al., 2005; Murphy & Hutchinson, 1982).

Ratings of memorability revealed medium average ratings across all fractals, with high variability between fractals. The ratings for memorability were externally validated with actual performance in a memory task. For this task, 60 fractals were randomly selected from the present database. Correlational analyses revealed medium-sized relations ($r = .36$) between the norms for the perceived memorability and the actual recognition-memory performance. These results confirm validity of the ratings for memorability. Thus, the present database provides norms to manipulate or control abstract stimuli with respect to their memorability. This specifically allows to control for task-difficulty in memory tasks based on stimulus selection and might be interesting in research comparing memory performance between different age groups (e.g., Brehmer et al., 2012) or clinical samples (e.g., Whittington et al., 2006).

Ratings for the stimulus feature familiarity in the presented database were generally low. This observation is in line with our assumption that the semantic or conceptual content of the fractals in the current database is very restricted. Information about familiarity of visual stimuli is essential for different levels of cognitive processing (cf. Gernsbacher, 1984; Kamas & Reder, 1994): previous research revealed faster responses to familiar than to unfamiliar stimuli in visual search tasks (e.g., Wang et al., 1994 or in naming tasks (e.g., Alario et al., 2004; Ralph et al., 1998). With regards to memory, repeatedly presented stimuli are supposed to be perceived as more familiar and enhance short-term memory performance in some studies (e.g., Mayer et al., 2011; Xie & Zhang, 2017), but not in others (e.g., Chen et al., 2006). In our study, familiarity was not related to performance (*d*-prime) in a recognition-memory task. Hence, our database of standardized fractals might serve as a useful tool for an approach to solve the reported inconsistencies in this field.

It is noteworthy that we observed correlations beyond our specific predictions (i.e., unspecific correlations, cf. Fig. 5). A potential reason for these unspecific correlations is that the fractals are generally hard to verbalize, whereas some of the fractals' configurational content consists of salient shapes, which are easier to verbalize. Thus, verbalizability might also influence the ratings of other features. For instance, if the configurational content of a fractal consists of a salient shape, e.g., spiral, its verbalizability rating will be higher and in conjunction also its other ratings such as memorability, familiarity or even complexity, compared to fractals with less salient configurational content. In line with this assumption, naming agreement, representing actual verbalizability, was at least weakly related to all feature ratings. However, there are also alternative explanations for the observed unspecific correlations.

Classifier probability of AlexNet revealed unspecific correlations with all ratings except animacy. The medium correlation with complexity implies that classifier probability of AlexNet was lower for fractals with high complexity ratings. This might be in line with the fact that the DNN processes increasingly complex image information in every layer, resulting in image classification in the last layer (e.g., Jozwik et al., 2017; Wen et al., 2018). Moreover, low- to medium-sized correlations revealed that the DNN was more confident with fractals rated as more familiar, favorable, or memorable. Together with the relation between DNN classifier probability and indices of naming agreement from the labeling questionnaire, these correlations could be interpreted as support for the assumption of AlexNet as a model for human object recognition. The zip ratio revealed low unspecific correlations with abstractness, animacy and familiarity. Crucially, images can be zipped to a lesser degree when they contain less structured or more irregular information. Fractals could be zipped to a higher degree when abstractness was rated higher, thus certain regularities or repetitive patterns in image structures might be perceived as more abstract. In contrast, fractals that could be zipped to a lesser degree, were associated with higher animacy or familiarity ratings. Higher familiarity ratings for more irregular stimuli might thus reflect our experience with a highly irregular perceptual environment in everyday life. Memory performance showed one unspecific correlation with animacy. This medium-sized effect indicated worse memory performance for fractals with higher animacy ratings. This finding is surprising, since it contradicts the animacy effect usually observed in memory for meaningful stimuli, but also for nonwords (for a review see Nairne et al., 2017). Notably, the alternative explanations for the unspecific effects are post hoc explanations and, hence, must be confirmed in future studies. However, they are broadly consistent with the literature.

Complementing the norms, we provide the modal name for each fractal. The modal name reflects the label given with highest percentage for a fractal, i.e., the name for which highest consensus between participants was achieved. In total, we obtained 178 different modal names. The most frequent modal names were snail, flower, spiral, water, star, waves, stair, tunnel, forest, rose, each given to at least ten fractals. As the fractals contain semantic content to a very low degree, the modal name rather reflects the configurational content of the fractals, such as salient shapes. Therefore, we assume that fractals with equal modal names point to similar configurational content and thus group in categories of visual similarity. Fractals with similar modal names are less distinct than fractals with different modal names. The modal names hence point to underlying categorical structures in the database and allow to create stimulus sets with distinct or similar fractals.

With the present study, we provide a standardized database of abstract and complex stimuli with valid norms for relevant stimulus features such as abstractness, animacy, complexity, familiarity, favorableness, memorability, and verbalizability as well as indices of naming agreement. This is essential in research investigating, for instance, visual processing as independently as possible from verbal influences. In research presenting abstract stimuli as novel stimuli, information about the stimulus feature of familiarity is crucial. For memory research, the stimulus feature of perceived memorability could be used to control for task difficulty. Moreover, the modal names of the fractals point to salient configurational content such as shape and equal modal names for multiple fractals can thus be used as indicators for underlying categorical structures of the database. Ultimately, our standardized database offers wide-ranging possibilities for conducting experimental research with abstract stimuli under controlled stimulus-based conditions.

Author contributions Rebecca Ovalle-Fresa: conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, visualization, writing - original draft, writing - review & editing. Sarah Valérie di Pietro: software, investigation, writing - review & editing. Thomas Peter Reber: analyses, writing - review & editing. Eleonora Balbi: software, investigation. Nicolas Rothen: conceptualization, funding acquisition, project administration, supervision, methodology, writing - review & editing.

References

Abdelmounaime, S., & Dong-Chen, H. (2013). New Brodatz-Based Image Databases for Grayscale Color and Multiband Texture Analysis. *ISRN Machine Vision*, *2013*, 1–14. https://doi.org/10.1155/2013/876386

Alario, F. X., Ferrand, L., Laganaro, M., New, B., Frauenfelder, U. H., & Segui, J. (2004). Predictors of picture naming speed. *Behavior Research Methods, Instruments, and Computers*, *36*(1), 140–155. https://doi.org/10.3758/BF03195559

Bellhouse-King, M. W., & Standing, L. G. (2007). Recognition memory for concrete, regular abstract, and diverse abstract pictures. *Perceptual and Motor Skills*, *104*, 758–762. https://doi.org/10.2466/pms.104.3.758-762

Bentley, W. A., & Humphreys, W. J. (1962). Snow crystals. Dover.

Bradley, M. M., Hamby, S., Löw, A., & Lang, P. J. (2007). Brain potentials in perception: Picture complexity and emotional arousal. *Psychophysiology*, *44*(3), 364–373. https://doi.org/10.1111/j.1469-8986.2007.00520.x

Brehmer, Y., Westerberg, H., & Bäckman, L. (2012). Working-memory training in younger and older adults: Training gains, transfer, and maintenance. *Frontiers in Human Neuroscience*, *6*(63), 1–7. https://doi.org/10.3389/fnhum.2012.00063

Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The bank of standardized stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PLoS ONE*, *5*(5), e10773. https://doi.org/10.1371/journal.pone.0010773

Brodeur, M. B., Guérard, K., & Bouras, M. (2014). Bank of Standardized Stimuli (BOSS) phase II: 930 new normative photos. *PLoS ONE*, *9*(9). https://doi.org/10.1371/journal.pone.0106953

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., … Massimini, M. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Science Translational Medicine*, *5*(198). https://doi.org/10.1126/scitranslmed.3006294

Chen, D., Yee Eng, H., & Jiang, Y. (2006). Visual working memory for trained and novel polygons. *Visual Cognition*, *14*(1), 37–54. https://doi.org/10.1080/13506280544000282

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (second). Lawrence Erlbaum Associates.

Csárdi, G., Podgórski, K., & Geldreich, R. (2020). zip: Cross-Platform "zip" Compression. Retrieved at 2021-09-21 from https://cran.r-project.org/package=zip%0A

Eng, H. Y., Chen, D., & Jiang, Y. (2005). Visual working memory for simple and complex visual stimuli. *Psychonomic Bulletin and Review*, *12*(6), 1127–1133. https://doi.org/10.3758/BF03206454

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, *113*(2), 256–281. https://doi.org/10.1037/0096-3445.113.2.256

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012,* 759–765.

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019). THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images (preprint). *BioRxiv*. https://doi.org/10.1101/545954

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2021). lab.js: A free, open, online study builder. Behavior Research Methods. https://doi.org/10.3758/s13428-019-01283-5

Horikawa, T., & Kamitani, Y. (2017a). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, *8*, 1–15. https://doi.org/10.1038/ncomms15037

Horikawa, T., & Kamitani, Y. (2017b). Hierarchical neural representation of dreamed objects revealed by brain decoding with deep neural network features. *Frontiers in Computational Neuroscience*, *11*(4), 1–11. https://doi.org/10.3389/fncom.2017.00004

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in Psychology*, *8*, 1726. https://doi.org/10.3389/fpsyg.2017.01726

Kamas, E. N., & Reder, L. M. (1994). The role of familiarity in cognitive processing. In R. F. Lorch & E. J. O'Brien (Eds.), *Sources of Coherence in Reading* (pp. 177–202). Erlbaum.

King, M. L., Groen, I. I. A., Steel, A., Kravitz, D. J., & Baker, C. I. (2019). Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic

images. *NeuroImage*, *197*, 368–382. https://doi.org/10.1016/j.neuroimage.2019.04.079

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 1097–1105.

Lempel, A., & Ziv, J. (1976). On the complexity of finite sequences. *IEEE Transactions on Information Theory*, *22*(12), 75–81. https://doi.org/10.1109/TIT.1976.1055501

Leynes, Andrew, P., Batterman, A., & Abrimian, A. (2019). Expectations alter recognition and event-related potentials (ERPs). *Brain and Cognition*, *135*, 103573. https://doi.org/10.1016/j.bandc.2019.05.011

Lupyan, G. (2012). Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in Psychology*, *3*, 54. https://doi.org/10.3389/fpsyg.2012.00054

Maguire, E., Valentine, E. R., Wilding, J. M., & Kapur, N. (2003). Routes to remembering: the brains behind superior memory. *Nature Neuroscience*, *6*(1), 90–95. https://doi.org/10.1038/nn988

Mandelbrot, B. B. (1983). Fractals and the geometry of nature. W.H. Freeman and Co.

Mayer, J. S., Kim, J., & Park, S. (2011). Enhancing visual working memory encoding: The role of target novelty. *Visual Cognition*, *19*(7), 863–885. https://doi.org/10.1080/13506285.2011.594459

Moreno-Martínez, F. J., & Montoro, P. R. (2012). An ecological alternative to Snodgrass & Vanderwart: 360 high quality colour images with norms for seven psycholinguistic variables. *PLoS ONE*, *7*(5), e37527. https://doi.org/10.1371/journal.pone.0037527

Murphy, G. L., & Hutchinson, J. W. (1982). Memory for forms: Common memory formats for verbal and visual stimulus presentations. *Memory & Cognition*, *10*(1), 54–61. https://doi.org/10.3758/BF03197625

Nairne, J. S., VanArsdall, J. E., & Cogdill, M. (2017). Remembering the living: episodic memory is tuned to animacy. *Current Directions in Psychological Science*, *26*(1), 22–27. https://doi.org/10.1177/0963721416667711

Nishimoto, T., Ueda, T., Miyawaki, K., Une, Y., & Takahashi, M. (2010). A normative set of 98 pairs of nonsensical pictures (droodles). *Behavior Research Methods*, *42*(3), 685–691. https://doi.org/10.3758/BRM.42.3.685

Nishiyama, M., & Kawaguchi, J. (2014). Visual long-term memory and change blindness: Different effects of pre- and post-change information on one-shot change detection using meaningless geometric objects. *Consciousness and Cognition*, *30*, 105–117. https://doi.org/10.1016/j.concog.2014.09.001

Ovalle-Fresa, R., Uslu, A. S., & Rothen, N. (2021). Levels of Processing Affect Perceptual Features in Visual Associative Memory. *Psychological Science*, *32*(2), 267–279. https://doi.org/10.1177/0956797620965519

Parsons, S. (2021). Splithalf: Robust estimates of split half reliability. *Journal of Open Source Software,* *6*(60), 3041. https://doi.org/10.21105/joss.03041

Ralph, M. A. L., Graham, K. S., Ellis, A. W., & Hodges, J. R. (1998). Naming in semantic dementia - What matters? *Neuropsychologia*, *36*(8), 775–784. https://doi.org/10.1016/S0028-3932(97)00169-3

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: a meta-analytic review. *Psychological Bulletin*, *137*(1), 131–148. https://doi.org/10.1037/a0021705

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Sarasso, S., Rosanova, M., Casali, A. G., Casarotto, S., Fecchio, M., Boly, M., … Massimini, M. (2014). Quantifying cortical EEG responses to TMS in (Un)consciousness. *Clinical EEG and Neuroscience*, *45*(1), 40–49. https://doi.org/10.1177/1550059413513723

Schooler, J. W. (2002). Verbalization produces a transfer inappropriate processing shift. *Applied Cognitive Psychology*, *16*(8), 989–997. https://doi.org/10.1002/acp.930

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Simpson, E. H. (1949). Measurement of Diversity. *Nature*, *688*(1943), 688. https://doi.org/10.1038/163688a0

Smith, W., & Dror, I. E. (2001). The role of meaning and familiarity in mental transformations. *Psychonomic Bulletin and Review*, *8*(4), 732–741. https://doi.org/10.3758/BF03196211

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34–50. https://doi.org/10.1037/0096-3445.117.1.34

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 174–215. https://doi.org/10.1037/0278-7393.6.2.174

Voss, J. L., Schendan, H. E., & Paller, K. A. (2010). Finding meaning in novel geometric shapes influences electrophysiological correlates of repetition and dissociates perceptual and conceptual priming. *NeuroImage*, *49*(3), 2879–2889. https://doi.org/10.1016/j.neuroimage.2009.09.012

Voss, J. L., Lucas, H. D., & Paller, K. A. (2012). More than a feeling: Pervasive influences of memory without awareness of retrieval. *Cognitive Neuroscience*, *3*(3–4), 193–207. https://doi.org/10.1080/17588928.2012.674935

Wang, Q., Cavanagh, P., & Green, M. (1994). Familiarity and pop-out in visual search. *Perception & Psychophysics*, *56*(5), 495–500. https://doi.org/10.3758/BF03206946

Ward, J., Hovard, P., Jones, A., & Rothen, N. (2013). Enhanced recognition memory in grapheme-color synaesthesia for different categories of visual stimuli. *Frontiers in Psychology*, *4*, 762. https://doi.org/10.3389/fpsyg.2013.00762

Wen, H., Shi, J., Zhang, Y., Lu, K. H., Cao, J., & Liu, Z. (2018). Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral Cortex*, *28*(12), 4136–4160. https://doi.org/10.1093/cercor/bhx268

Whittington, C., Podd, J., & Stewart-Williams, S. (2006). Memory deficits in Parkinson's disease. *Journal of Clinical and Experimental Neuropsychology*, *28*(5), 738–754. https://doi.org/10.1080/13803390590954236

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

Xie, W., & Zhang, W. (2017). Familiarity increases the number of remembered Pokémon in visual short-term memory. *Memory and Cognition*, *45*(4), 677–689. https://doi.org/10.3758/s13421-016-0679-7

Ziv, J., & Lempel, A. (1977). A Universal Algorithm for Sequential Data Compression. IEEE Transactions on Information Theory, 23(3), 337–343. https://doi.org/10.1109/TIT.1977.1055714

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.