Supplementary Information A.

Grass and Gravel:
Investigating Visual Properties Preschool Children and Adults Use When Distinguishing Naturalistic Images

Karola Schlegelmilch[1] and Annie E. Wertz[1]

Author Affiliations:

[1]Max Planck Institute for Human Development, Max Planck Research Group Naturalistic Social Cognition, Lentzeallee 94, 14195, Berlin, Germany.

Web link to data:
https://osf.io/8xy5n/?view_only=6ddced286c31456fae7d20dd86e072e6

A. 1 Wording of Task Instructions

Table A. 1

Translations of the Instructions for the Similarity Sorting Task

German Original	English Translation
Children	
„Hast Du Lust, mit mir zu spielen?"... Experimentator*in holt die Puppe.	*"Would you like to play a game with me?" - the Experimenter picks up the puppet.*
„Erst einmal erzähle ich eine Geschichte - von Uri. Das ist Uri!" Die Puppe wird gezeigt. Sie bewegt sich kurz und winkt. „Uri kann nicht sprechen, aber er kann fliegen! Weißt Du, was Uri für ein Tier ist?" ... „Ja, eine Fledermaus. Und weißt Du, wann Fledermäuse fliegen?" ... „Genau, die fliegen, wenn es dunkel ist. Und dann sehen sie Sterne."	*"First, I'm going to tell you a story about Uri. Here he is!" The puppet is shown to the child. The puppet is moved around and waves. "Uri can't speak, but he can fly! Do you know what sort of animal Uri is?"…"Yes, he's a bat! And do you know when bats normally fly?"…"Exactly, they fly when it's dark. And in the dark, there are lots of stars in the sky."*
"Uri lebt eigentlich in einer Welt, wo es ganz besondere Sterne gibt: Mustersterne! Und die vermisst Uri sehr." Puppe wird hingesetzt. „Mustersterne sind etwas ganz Besonderes: In ihren Strahlen haben sie Bilder, die sich ganz ähnlich sind, weil sie nämlich das gleiche Muster haben. Jeder Strahl hat ein anderes Muster. Und weil Uri selbst keine Mustersterne basteln kann, kannst Du das vielleicht für ihn machen. Ich habe mir ein Spiel ausgedacht, damit es für Dich einfacher ist. Hättest Du Lust, Uri zu helfen?"	*"Now, Uri comes from a world where there are very special kinds of stars: patterned stars! And Uri misses them a lot." The puppet is put back down. "Patterned stars are something special: their rays have little pictures in them, and these pictures look quite similar to one another because they have the same pattern. Every ray has a different pattern. Because Uri can't make a patterned star himself, maybe you can make one for him. Would you like to help Uri?"*

(*Continuing* Table A. 1*)*

German Original	English Translation
„Siehst Du diese Karten? Da sind unterschiedliche Muster drauf. Du kannst mir jetzt zwei Karten zeigen, die sich ähnlich sehen, weil sie ein ähnliches Muster haben, dann legen wir die zwei zusammen, und das ist der Anfang vom ersten Strahl.	*"Do you see these cards? There are different patterns on them. For a start, give me two cards that look similar to one another. We'll put those cards down beside one another, and that will be the start of the first ray."*
Die Kinder fangen an. Die beiden kombinierten Karten werden als Paar beiseitegelegt, und eine neue Karte (oder bei Bedarf auch mehrere neue Karten) füllen die Lücken. Immer wieder sollten - in Sätze eingebaut - die Hinweise kommen: Strahlen mit Bildern, die sich ähnlich sind ... Bilder, die gleiche Muster haben ... ein schöner Musterstern mit langen Strahlen usw. Evtl. der Hinweis: „Es ist gar nicht wichtig, was für Dinge oder Sachen auf den Karten sind, Uri geht es nur um die Muster." Falls nur 2er-Paare gefunden werden wird darauf hingewiesen, dass die Strahlen auch aus mehr Karten bestehen können. Das Kind spielt so lange weiter, bis es nicht mehr möchte oder die Karten verbraucht sind.	*The children begin the game. The first set of combined cards are set down as a pair, and a new card (or if necessary multiple new cards) fill the gaps. Again and again the experimenter should—built into appropriate sentences—repeat the following sorts of hints: "rays with pictures that look similar"…"pictures with the same pattern"…"a nice star with long rays" and so forth.* *Eventually the experimenter may say "it's not at all important what the things in the pictures are—Uri is only interested in the pattern."* *In the case of the child only setting down 2-card groups/rays they will be reminded that they can put more cards onto the rays that are already there. The child plays the game until they no longer want to, or until all the cards are used up.*
„Oh, das sind schöne Muster! Uri, gefallen sie Dir? Freust Du Dich über diesen Musterstern?" Uri fliegt über die Karten und bedankt sich dann durch nicken.	*"Oh, what lovely patterns! Uri, do you like it? Does this star make you happy?"* *Uri flies over the cards and thanks the child by nodding.*

(Continuing Table A. 1*)*

German Original	English Translation
Adults	
„Wir möchten Sie bitten, Karten, die ich Ihnen gleich geben werde, mit Karten zusammen zu legen, denen sie visuell ähnlich erscheinen. Die Gründe, weshalb Sie etwas visuell ähnlich empfinden entscheiden Sie selbst. Dabei ist es nicht wichtig, welche Gegenstände auf den Karten abgebildet sind - die Karten, welche am ähnlichsten aussehen, kommen zusammen in eine Gruppe. Die Größe einer Gruppe ebenso wie die Anzahl der entstandenen Gruppen hängt davon ab, wie viele Karten Sie jeweils als ähnlich empfinden. Da gibt es keine Vorgaben. Haben Sie Fragen dazu?“	*“We ask you to lay down cards—which I will give to you shortly—with other cards on the basis of how visually similar they are.* *What it is for them to be “visually similar” is something you decide for yourself. It’s not at all important what objects are depicted on the cards—if cards strike you as similar, you put them together in a group. The size of these groups and the number of those groups both depend on how similar you find the cards. There are no other requirements.* *Have you any other questions?”*
„Sie können mit Paaren von 2 ähnlichen Karten beginnen, und später noch andere passende Karten dazu legen, so dass die Gruppen größer werden.“ (Es sollten keine Fragen beantwortet werden, welche sich auf die Motive beziehen)	*“You can start with pairs of two similar cards, and later add other matching ones to make the group bigger.”* *(The experimenter cannot answer questions relating to the images' identities)*

Table A. 2

Translations of the Instructions for the Classification Task

German Original	English Translation
Children	
"Nun machen wir noch ein anderes Spiel: Es ist ein Ratespiel. Du überlegst, in welche der 3 Schachteln jede Karte gehört.	*“Now we’re going to play another game: this one is a sorting game. You figure out in which of these three boxes each of these cards belong."*
„In die 1. Schachtel kommt alles wovon Du glaubst, dass es irgendwelche Pflanzen oder Bäume sind. Weißt Du was alles Pflanze ist? *Wenn keine Antwort: „Oh, ich glaube das wirst Du dann sehen“*	*“In the first box goes anything that you think is a plant or tree. Do you know what sort of things plants are? If no answer, say “oh I think you’ll see it soon enough.”*
"In die 2. Schachtel kommen alle Sachen, die von Menschen gemacht wurden. Du kennst sie vielleicht aus Euren Schränken in der Küche oder so, oder siehst sie an Häusern oder auf der Straße. Du eine Idee, was für Sachen das sein können?"	*"In the second box go things that are man-made. You would know them from seeing them in your kitchen, in houses or on the street. Do you have an idea what sorts of things these are?"*
"In die 3. Schachtel kommen Sachen aus der Natur - die man in den Bergen sieht, oder am Meer. Da dürfen aber keine Pflanzen rein, weil die in die erste Schachtel gehören, und auch nichts was von Menschen gemacht wurde, das kommt in die 2. Schachtel. Hast Du eine Idee, was für Sachen da reingehören?"	*" In the third box go natural things – things you might see in the mountains, or by the sea. However, plants can’t go in there, since they belong in the first box, and the same goes for man-made things, which have their place in the second box. Do you have an idea about what kind of things might belong in there?"*
Adults	
"Wir möchten Sie nun bitten, die Karten in diese 3 Schachteln einzuordnen. Dabei trennen Sie bitte die Karten in folgende Kategorien: Vegetation, nicht-lebendige natürliche Elemente, und menschengemachte Dinge ein. Die Karten stellen keine weiteren Kategorien dar."	*" We now ask you to sort these cards into these three boxes. We want you to separate these cards into the following categories: Vegetation, non-living natural things, and manmade objects. There are no further categories for the cards that these."*

(Continuing Table A. 2*)*

German Original	**English Translation**
"In "Vegetation" legen Sie alles was Pflanzen oder Ausschnitte von Pflanzen darstellt In „Natürliche Elemente“ das, was Sie als natürliche Materie erkennen - also alles was man in der natürlichen Umwelt sehen kann, dabei weder Pflanze noch Tier ist und nicht künstlich hergestellt wurde. In „Menschengemachtes“ kommen Alltagsgegenstände und Utensilien - alles, wovon Sie denken es ist nicht von selbst entstanden, sondern wurde von Menschen irgendwie hergestellt. Haben Sie noch Fragen?"	*" In „Vegetation“ go all the cards that depict plants or parts of plants.* *In “Natural Elements” go things that you would know as natural materials – that is, things that you would see in a natural environment, and which are not plants, nor animals, nor manmade objects.* *In “Manmade” go everyday objects and utensils – everything that you would say isn’t naturally occurring, but has instead been produced by people.* *Do you have any questions?*
Es sollten bei Unklarheit der Kategorien keine Beispiele zur Erklärung benutzt werden, sondern Umschreibungen!	*In the case of any confusion about the categories there can’t be any reference to visual examples, only descriptions!*

A. 2 Analysis of Questionnaires on Prior Exposure to the Categories

After the experiments, a questionnaire was given to the adult participants or the caregivers of the children which asked about frequencies of prior exposure to each of the categories: artifacts, natural elements, and vegetation. The questionnaires were developed for the present study. They included questions about exposure to the categories due to activities which the participant him- or herself performed, or exposure to activities a participant passively experienced—for example, when the partner of an adult or a child's parent was involved in the activities. We also assessed general exposure to pictures or picture books, and exposure to more abstract, computer-related activities such as text processing. These more general questions were expected to indicate visual exposure to two-dimensional visual information, which differs from that of the naturalistic environment. For each of the questions, five possible frequencies could be chosen. These were: a) more than 4 times a week; b) 1 - 4 times a week; c) 1 - 4 times a month; d) less than once a month; c) never.

We then averaged the frequencies over active and passive exposures and correlated the averages with performance data of the identification task (i.e., the sensitivity measure dprime) separately for each of the categories depicted in the images. No significant correlations were found between exposure frequencies and sensitivity for a particular category after adjusting *p*-values (Benjamini & Hochberg, 1995). However, when correlating participants overall sensitivity values with frequencies of general exposure to pictures or picture books, we found that more frequent exposure to activities including pictures and picture books led to lower sensitivity for the categories depicted in the study's images in children (Spearman’s $r(223) = -.19$, $p = .02$), but not in adults. Exposure to abstract, computer-related activities was not significantly related to sensitivity in children or adults. One possible explanation for this finding could be that children who spend much time with children's books learn graphical versions of entities, which do not include visual information as it is useful for the perception of photographs. Additionally, the more time a child spends with picture books and pictures, the less outdoor activities this child is exposed to. In contrast to what one might expect, frequent visual exposure to pictures did therefore not lead to an increase in the ability to perceive two-dimensional visual information. Future studies could compare the effect of frequent exposure to graphics designed for children with the effect of frequent outdoor activities on developing perceptual abilities. Moreover, this finding questions the validity of graphical representations of natural entities in categorization studies conducted with young children.

A. 3 Reduction of the Originally Considered Visual Properties

In preparation for the study, we had compiled a more extensive list of visual properties from the literature. However, we reduced the original selection to avoid redundancies (i.e., groups of properties with high inter-correlations, or which are statistically assessed with analogous codes) and to receive a set that covered diverse qualities and referred to substantial visual tasks related to the human environment. In the following, we will describe the selection process.

In order to choose the most meaningful statistical properties for the current project, we started by selecting an excessive set of properties during from the literature on visual scene processing, computational vision, or image-processing. A wide variety of statistical properties is included in this literature (e.g., Clausi, 2002; Gonzalez & Woods, 2018; Materka & Strzelecki, 1998) that sometimes only slightly differ in their algorithms. On the other hand, some image characteristics—such as fractality—are mathematically defined in different ways (e.g., Burton & Moorhead, 1987; Costa et al., 2012; Isherwood et al., 2017), so that the different algorithms refer to different concepts of the related visual phenomena.

Before starting with the main analysis, we reduced the preliminarily selected set of properties by looking at their correlational structure in the 141 images of the pilot study, see Figure A. 1.

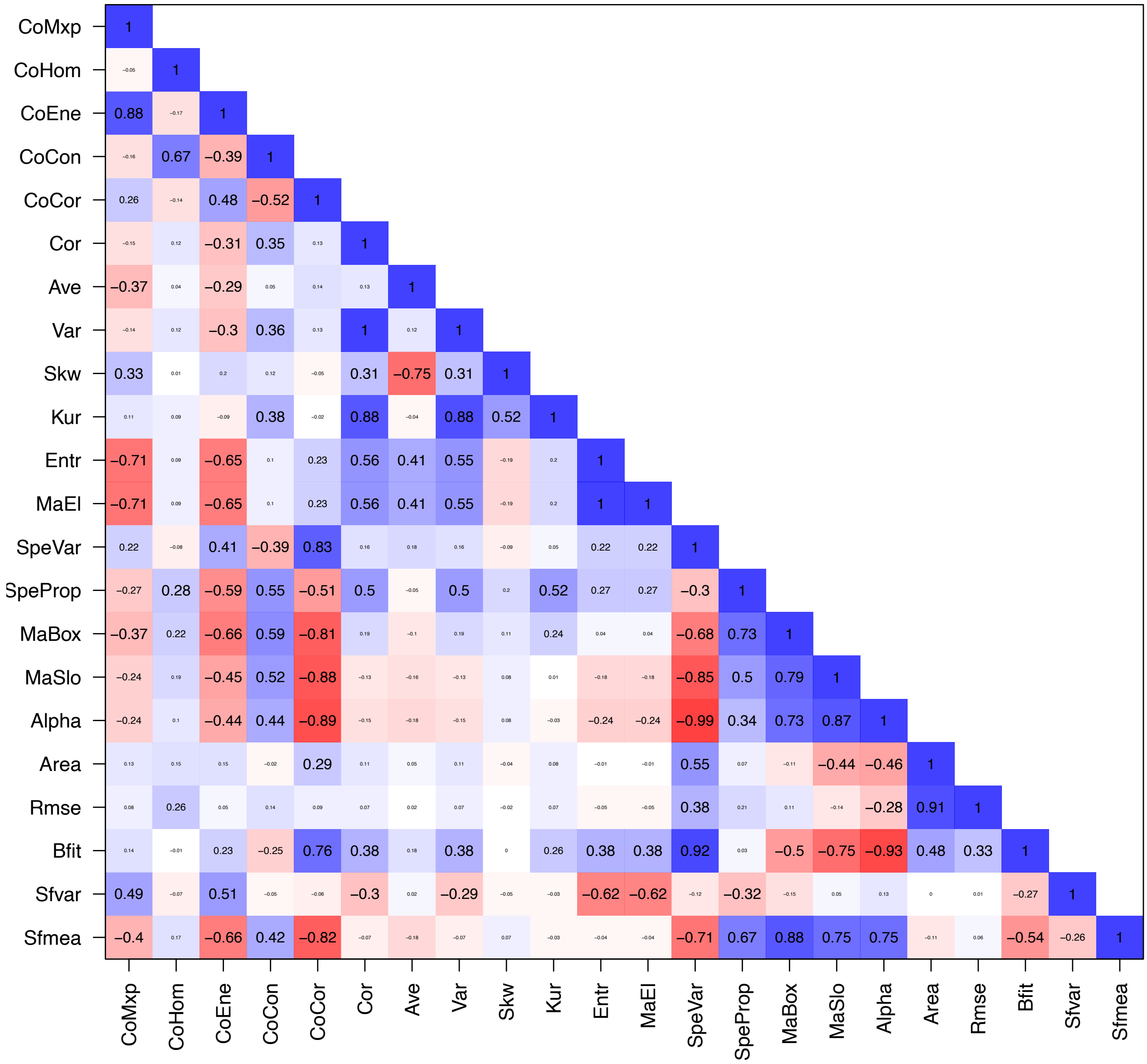

Figure A. 1. Correlation matrix of the preliminary visual properties (pilot study).

Definitions of the preliminary properties:

- Co-occurrence of image pixels (Clausi, 2002; assessed with function from: Gonzalez & Woods, 2018)
  - CoMxp = Max of all co-occurrence probabilities
  - CoHom = Co-occurrence homogeneity
  - CoEne = Mean co-occurrence energy
  - CoCon = Co-occurrence contrast
  - CoCor = Co-occurrence correlation

- Low-level statistics
  - Ave = Mean of the pixel luminance histogram
  - Var = Variance of the pixel luminance histogram

- Skw = Skew of the pixel luminance histogram
- Kur = Kurtosis of the pixel luminance histogram

- Pixel luminance predictability
  - Cor = Correlation of regions of whole image, related to SD
  - Entr = Shannon entropy (assessed with Matlab function of: Gonzalez & Woods, 2018)
  - MaEl = Shannon entropy (assessed with Matlab function of: Mather, 2018)

- Spatial frequency (SF) distribution
  - SpeVar = Variance of the SF spectrum
  - SpeProp = Proportion of low *sf* divided by high SF

- Fractality
  - MaBox = Fractal dimension, differential box count (assessed with Matlab function of: Mather, 2018)
  - MaSlo = Slope of amplitude spectrum (assessed with Matlab function of: Mather, 2018)
  - Alpha = Steepness of SF distribution on a logarithmic scale
  - Area = Deviation from scaling invariance—area under the curve between line fitted by alpha and the true SF distribution
  - Rmse = Root-mean-square deviation from scaling invariance
  - Sfvar = Segmentation-based Fractal Texture Analysis; Variance of 6 segments (assessed with the function sfta of: Costa et al., 2012)
  - Sfmea = Mean of 6 segments of the sfta vector (Costa et al., 2012)

We also summarized the properties to principal components and selected several single properties from the principal components. This was done because we expected a clearer interpretation from the analysis of a single algorithm than from the analysis of a principal component that aggregates several similar algorithms, and because different visual phenomena defined by properties with statistically similar algorithms could be hidden in a single principal component.

Based on these considerations, and by looking at the properties' predictive values for differentiating the true image categories, we decided to include only four statistical properties (i.e., alpha, deviation, skew, CooCor) to keep the risk of overfitting during the analysis low. Note that all rated properties that were originally assessed in the rating procedure (i.e., curvature, depth, gloss, regularity, size, symmetry) remained in the analysis.

A. 4 Correlations between Visual Properties in the Image Set

The visual properties we chose for the current study statistically relate to each other, indicated by correlations between some visual properties in our image set. We still decided not to agglomerate the correlating properties because a) they were chosen for theoretically distinct reasons, and b) even the members of highly correlating property pairs (i.e., alpha-CooCor, or regularity-symmetry) were found to be included very differently in the categorization decisions of the participant groups and led to distinct significance patterns between the categories. This important information would have been obscured by including for example the principal component of property pairs (see main Result section for further discussion on this decision). We present the correlation matrix of the visual properties in Figure A. 2.

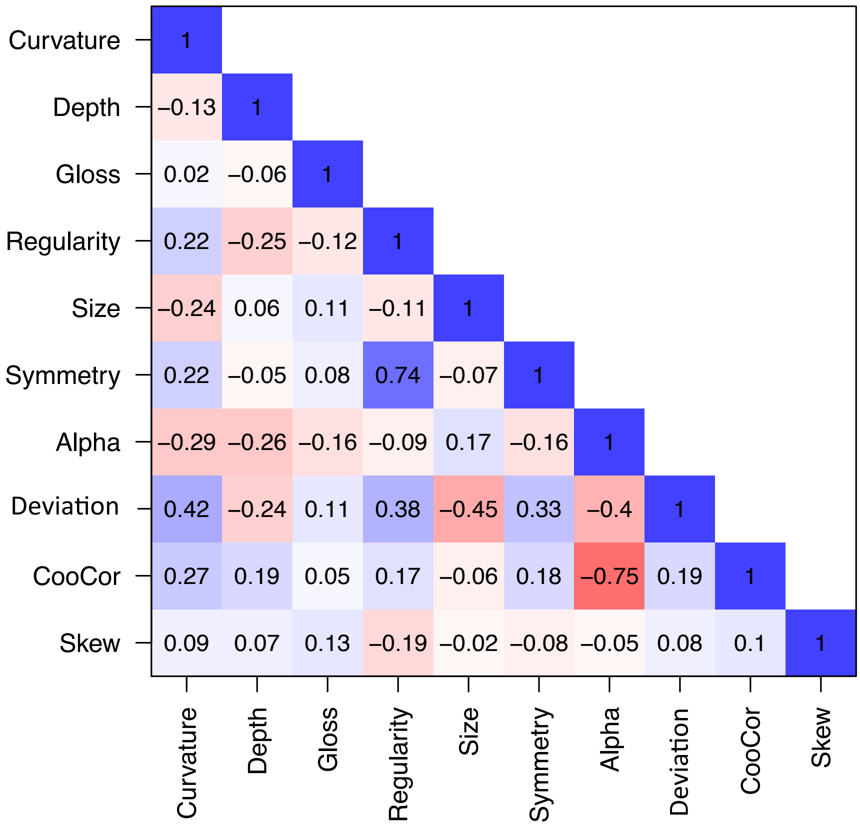

Figure A. 2. Correlation matrix of visual properties included in our image set.

Numbers are Pearson correlation coefficients including the data of 60 images.

A. 5 True Categories in our Image Set Predicted by Visual Properties

In order to determine visual properties which were additionally included in participants' decisions during identification although they did *not* predict categories in our images, or which were *not* included by participants although they *did* predict category membership in our image data, we assessed which visual properties statistically predicted the category membership in our image set.

For each of the three categories, separate GLMs were conducted (R-function glm, R Core Team, 2019) on the visual properties of the 60 images used in our study. The binary dependent variables (DV) indicated if an image depicted the respective category or not (1, 0). We assessed the significance of visual properties by including each visual property individually in a model, resulting in 10 tests, and adjusted *p*-values with the method Benjamini and Hochberg (1995). Coefficients for all visual properties are provided in Supplementary Tables (ST), Table B.1.

A. 6 Classification Task: Separate Results of Children and Adults

A. 6.1 Children's Classification Task Results

Each of the 76 children who participated in the classification task sorted a complete set of 30 cards into artifact, natural element, and vegetation boxes. In sum, they correctly classified $N = 1586$ (69.6%) of a total of $N = 2280$ images. The continuous variable Age, predicted children's proportion of correctly classified images ($F(1) = 10.8$, $p = .001$). We therefore included the covariate age in the GLMMs conducted in the analysis of children's assignment of categories.

Classification Performance Children. A confusion matrix (Table A. 3) shows the structure of responses to each of the presented images. Children most correctly assigned cards in the vegetation category, and least in the natural elements category. We assessed the discriminability index d-prime (*d'*; Wickens, 2002) for each of the true categories. It is calculated by subtracting the proportion of false alarms from the proportion of hits using the R-function *dprime* (R-package psych; Revelle, 2018). Higher values of *d'* indicate a better discriminability of one category from the others (Table A. 3). Analysis of variance of the three categories on *d'* revealed that children's discriminability differed between categories ($F(2, 150) = 20.6$, $p < .001$, $\eta^2 = .07$), in that their ability to discriminate natural elements was lower than for vegetation and artifacts (both adjusted $p < .001$; Post-hoc Tukey's HSD test), while discriminability for vegetation and artifacts did not differ.

Table A. 3

Classification Performance of Children and Adults.

Image category	Assigned category [a]						Decision measures	
	Artifact		Natural element		Vegetation		Discriminability [b]	
	M	SEM	M	SEM	M	SEM	d'	SEM
	Children						Children	
True artifact	**6.88**	**(0.26)**	2.89	(0.23)	1.67	(0.16)	1.70	(0.10)
True natural element	2.22	(0.20)	**6.34**	**(0.24)**	2.39	(0.17)	1.33	(0.09)
True vegetation	1.45	(0.11)	2.04	(0.23)	**7.64**	**(0.21)**	1.78	(0.07)
	Adults						Adults	
True artifact	**9.08**	**(0.76)**	1.19	(0.62)	1.05	(0.21)	2.99	(0.05)
True natural element	1.13	(0.35)	**9.07**	**(0.89)**	1.43	(0.70)	2.63	(0.07)
True vegetation	1.20	(0.42)	1.49	(0.68)	**8.99**	**(1.00)**	2.76	(0.07)

Note. Participants' category assignments (left), and decision measures (right) as functions of true image categories, separate for children (N = 76) and adults (N = 72).

[a] Cells show responses averaged over all participants within the participant group (*M*) and their standard error (*SEM*). Participants viewed 10 images per true category, respectively. Rows correspond to the true image categories, and columns indicate the participants' assignments of the same images to the respective categories. Cells in bold case indicate correct responses (hits), while the remaining cells indicate if an image was assigned to one category, but belonged to another category (false alarms).

[b] Discriminability (*d'*) indicates the participants' ability to discriminate one category from the other two, higher values indicate greater discriminability. Decision measures are averaged over all participants within the participant group.

Visual Properties Predicting Children's Category Assignment. In order to extract visual properties that predicted a child's assignment of an image to one of the three categories, we conducted GLMMs with a binomial error structure. Participants' responses were binarily coded, resulting in three DVs which indicated if an image was assigned to one of the categories or not (1, 0). To account for intra-class correlation, we included participants and images as the units of random intercepts. Visual properties were fixed effects, and age was a covariate in all models. We assessed the impact of visual properties on category assignment by including each visual property individually in a model, resulting in 10 tests. These tests equivalently explored the predictive value of each property. The reason for choosing separate analysis of visual properties were 1) missing prior expectations about their significances, and 2) the fact that interrelations of simultaneously included IVs in a full model might obscure the impact of some predictors (correlation matrix, Figure A. 2). Alternative methods like the agglomeration of properties by PCA might hide their unique contributions, whereas variable selection with specialized methods like lasso regression (Groll & Tutz, 2014) could generate different results if additional properties than those currently selected were included. These alternative types of analysis would not be as appropriate for the exploratory approach taken here. We controlled the false discovery rate in multiple comparisons by adjusting p-values within the 10 tests conducted for each DV using the method of Benjamini and Hochberg (1995).

Visual properties that predicted children's category assignments are specified in Figure A. 3a. Overall, the visual properties depth, symmetry, skew, and deviation significantly predicted children's classification. Children's assignment to artifacts was best predicted for images with high skew, high values of deviation, and low depth. Images with low symmetry had a higher probability to be assigned to natural elements, whereas images with greater depth were more likely to be assigned to vegetation. In particular, pictorial depth cues predicted children's decisions about category assignments. Although alpha was an important predictor of the categories in our image set (Figure 3, main text), it did not contribute significantly to children's assignments. All coefficients are reported in ST, Tables B.2–4.

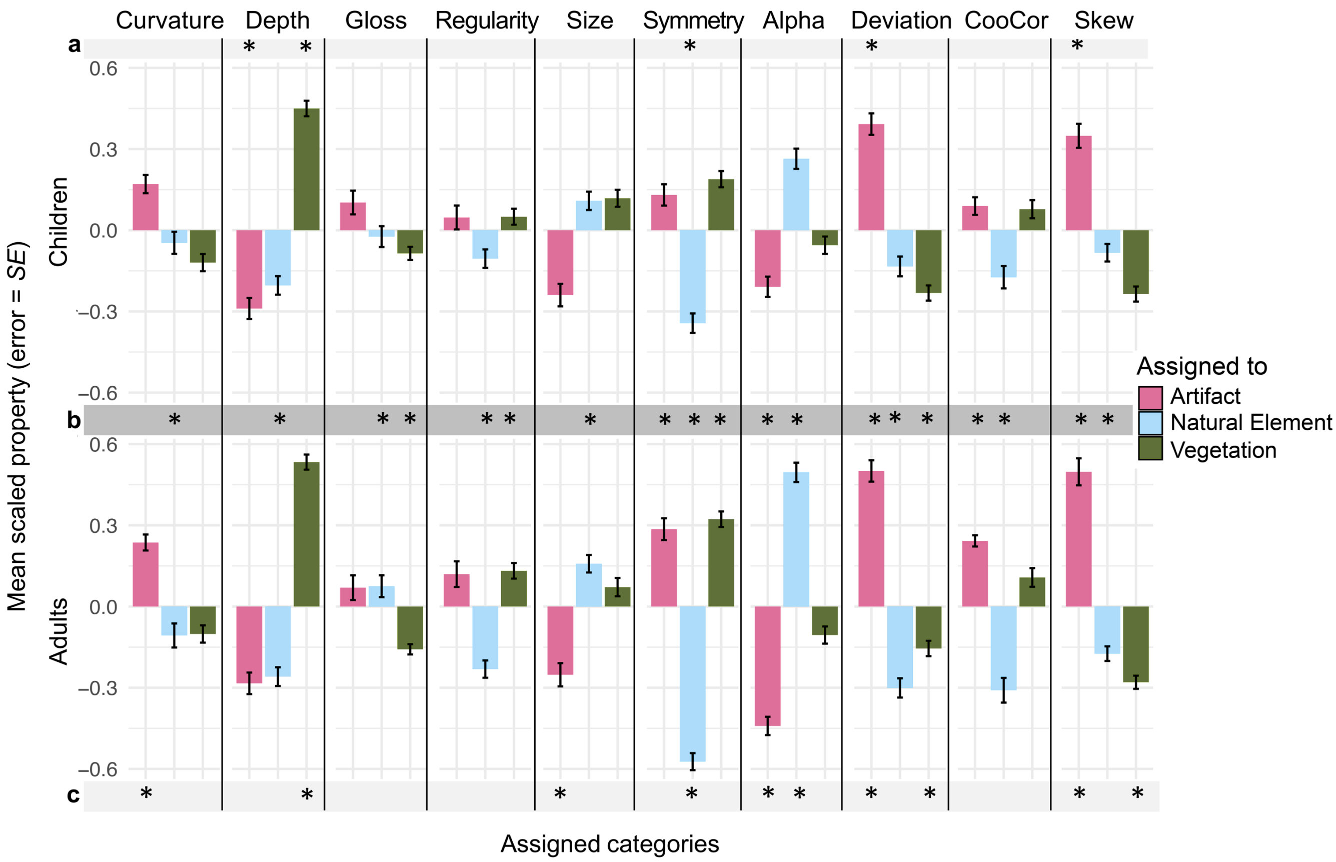

Figure A. 3. Visual properties as functions of assigned categories by children (top) and adults (bottom).

Properties are z-standardized and averaged over all the images that were assigned to a category by each participant, respectively. Asterisks on light-grey band in (a) and (c) indicate significant main effects of a visual property in the GLMMs conducted on the respective assigned category for children (a) and adults (c). Asterisks on dark-grey band in (b) indicate significant interaction terms between participant group and visual property, of the GLMM conducted with the data of both participant groups. All adjusted *p* < .05 (method: Benjamini and Hochberg,1995). Coefficients and *CI* are provided in ST, Tables B.2–4.

A. 6.2 Adults' Classification Task Results

Each of the 72 adult participants sorted a full set of 30 cards into artifact, natural element, and vegetation boxes, which resulted in a total of $N = 2160$ images sorted. Adults correctly classified $N = 1586$ (90.5%) of the sorted images.

Classification Performance Adults. A three level within-subject ANOVA on d' showed that adults' discriminability of the categories differed ($F(2, 142) = 21.4$, $p < .001$, $\eta^2 = .07$), in that discriminability of artifacts was higher than for vegetation and natural elements.

No other contrasts were significant. Inspection of the confusion matrix in Table A. 3 (bottom left) suggests that this effect can mainly be attributed to fewer incorrect assignments to the artifact category.

Visual Properties Predicting Adult's Category Assignment. In adults, category assignments to the images were mainly predicted by the visual properties alpha, deviation, and skew, but also by curvature, depth, size, and symmetry, see Figure A. 3.

Although adults correctly classified 90% of the images, some differences remained when comparing the visual properties predicting adult category assignment with those predicting our true image categories as shown in Figure 4, main text. In particular, curvature and size predicted adults' assignments to the artifact category although these properties did not predict the true category of artifacts in our image set.

A. 7 Similarity-sorting Task: Separate Results of Children and Adults

We first conducted HCAs separately for the children's and adults' similarity matrices, using the R-function hclust (R Core Team, 2019) with the Ward2 agglomerative clustering method (Murtagh & Legendre, 2011; Ward, 1963). HCAs determine a progressive series of more inclusive clusters—starting from unique combinations of image pairs to more general, larger ensembles. The groups' HCA solutions were then related to image characteristics (visual properties, assigned categories) respectively.

A. 7.1 Children's Similarity Sorting

Children sorted between 15 and 30 images into groups (*M* = 29.5, *SE* = .24). In sum, child participants assembled 2153 images. Each child sorted their images into *M* = 9.8 (*SE* = 0.3, range = 4 – 16) groups which each included *M* = 3.2 (*SE* = 0.1, range = 2 – 8) images. This led to 2874 combinations of image pairs that went into the analysis.

Children's HCA. The dendrogram of the children's HCA is shown in Figure A. 4 (left). The scale at the y-axis indicates the distances between clusters which are merged at a certain height. Inter-cluster distances of the children's sorting data ranged from a minimum of .18 between the two most similar images to a maximum of 1.89. In order to assess the predictive value of visual properties on the participants sorting decisions, we added the visual property values of the images to the data indicating the images' cluster membership. The impact of a visual property for each step in the clustering hierarchy was assessed by calculating the proportion of variability between the individual clusters to the total variability of the property, specified by R^2 (frequently termed “explained variance”; for a similar approach see: Friesen et al., 2015). Higher levels of the resulting R^2 values indicate a stronger variation of the visual property between clusters, interpreted as a stronger predictive value of the property on participants' sorting decisions. The predictive value of assigned categories was assessed with the same procedure.

In Figure A. 5 (top row) we plotted the visual properties' R^2 values as a function of the height of the dendrogram. The top left of Figure A. 5 illustrates the development of the impact of visual properties on children's sorting decisions. At the origin of the children's x-axis, each of the 60 images belonged to an individual cluster, resulting in values of $R^2 = 1$ for each of the properties. Visual inspection indicates that the R^2 values of particular visual properties started to vary late with increasing height at about height .4 (56 Clusters). Moreover, visual properties alternated in their predictive strength depending on the number of clusters in which images were organized, and on the corresponding inter-cluster distance. To evaluate the overall predictive value of the visual properties on children's similarity judgments, we included the R^2 values of each step in the agglomeration process (60 to 2 Clusters) in one test. Analysis of variance showed a main effect of visual properties on R^2, $F(9, 522) = 77$, $p < .001$, $\eta^2 = .03$. Post-hoc analyses using Tukey’s HSD indicated that depth predicted children's similarity decisions most strongly, differing from all other properties, all *p* < .005. Skew's predictive value as the second highest, and regularity as the third highest differed from all other properties except each other, all *p* < .005. At the least, gloss and alpha predicted children's similarity sorting (see main text, Figure 7; for all contrasts ST, Table B.6).

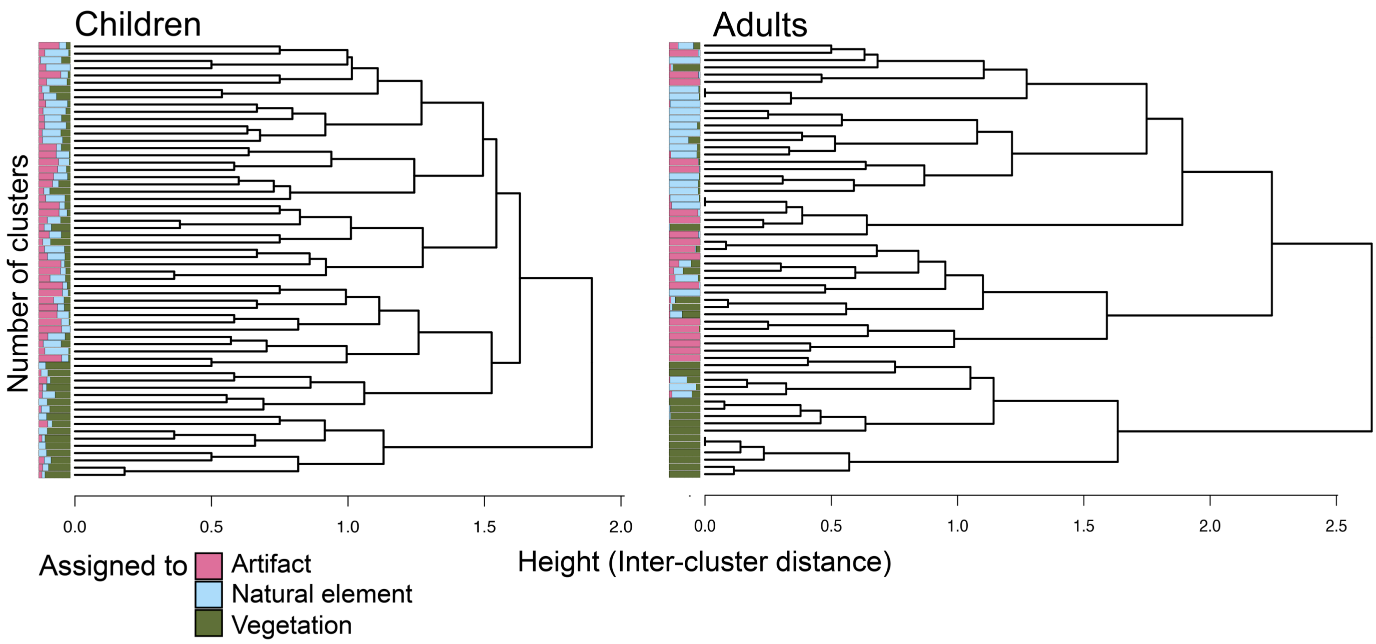

Figure A. 4. Hierarchical clustering results of the sorting task.

Dendrograms representing the structure of image similarities as hierarchical clusters received from the children's (left) and adults' (right) sorting task. For each dendrogram, zero height at the origin of the x-axis was the starting point from which individual images were agglomerated to decreasing numbers of clusters (method Ward2). Colored bars represent the proportion with which individual images were assigned to one of the categories (see main text, Figure 2). The levels of height indicate the dissimilarity of the merged image clusters (inter-cluster distance).

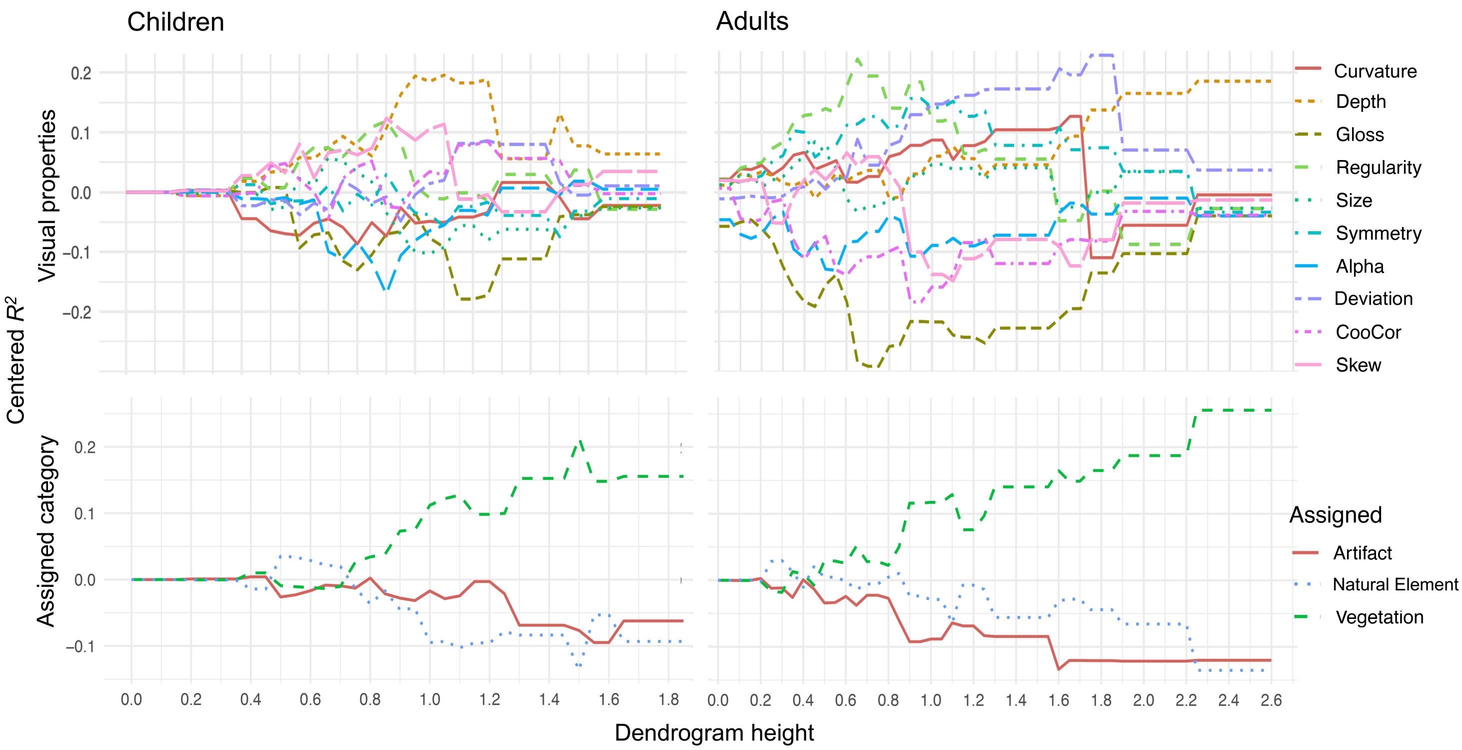

Figure A. 5. Explained variance of image characteristics by participants' sorting, as a function of the number of clusters.

Levels of R^2 for image characteristics as a function of the height of the dendrogram, for children (left) and adults (right). Height (x-axis) indicates the inter-cluster distance of images merged within successive hierarchical clusters. R^2 (y-axis) represents the predictive value of each of the image characteristics on similarity judgments (i.e. explained variance), separately for the 10 visual properties (top row) and for the categories assigned to an image in the classification task (bottom row). R^2 was assessed in steps of height .05. For better comparability of the differences between the properties, R^2 values were centered for each indexed height. A detailed discussion on the predictive value of image characteristics is provided in SI, A. 7.

The predictive value of assigned categories on children's similarity decisions is depicted in Figure A. 5 bottom left. Variability between categories increased late around height .75 (30 clusters). From here onward, the assigned vegetation category explained the largest proportion of variance compared to the other categories, with a value of $R^2 = .37$ at maximum height (2 clusters). An ANOVA with the factor assigned category on R^2 revealed a main effect for assigned categories ($F(2, 116) = 26$, $p < .001$, $\eta^2 = .03$), qualified by high levels of vegetation which differed from the two other categories ($p < .001$), while artifacts did not differ from natural elements within children (Figure A. 5 top right). These results confirm our assumption that children's similarity judgments were also predicted by the categories they perceived in the images.

A. 7.2 Adult's Similarity Sorting

Adults sorted between 26 and 30 images into groups (M = 29.3, SE = .12), in total 2110 images. The remainder of the images were not sorted into groups because no matches were found. Each adult assembled on average 7 (SE = .3, range = 3 – 13) different groups which included on average 4.7 (SE = .2, range = 2 – 10) images. This led to a total of 4592 paired images.

Adults' HCA. The right dendrogram of Figure A. 4 shows the HCA conducted on the adult sorting task data. Inter-cluster distances range between close to zero and the maximum of 2.64. The median of the decrease in the number of clusters lies at height .54. At minimum height (57 Clusters) visual properties already varied in their impact, indicating that 3 pairs of images that were grouped together by adults frequently corresponded in some of their visual properties (Figure A. 5 top right). During the full cluster agglomeration process, varying combinations of visual properties explained the within-cluster variance. Beyond height 1.9 (3 clusters) and until maximum height, depth explained most variance, predicting adults' similarity judgments in a very general way. Analysis of variance of the factor visual property on R^2 values of the whole agglomeration process showed a main effect of visual property, $F(9, 522) = 121$, $p < .001$, $\eta^2 = .08$. Post-hoc Tukey's HSD comparisons indicated that regularity had the highest predictive value for adults' similarity judgments, differing from all other properties except symmetry, with all $p < .001$. Symmetry differed from the remaining properties except deviation, all $p < .05$. Gloss predicted adults' judgments least, differing from the other weaker properties alpha and CooCor, both $p < .001$ (main text, Figure 7; A. 7).

As shown in Figure A. 5 (bottom right), assigned categories started to vary in their predictive value at height .2 (51 clusters) later than visual properties. Vegetation strongly predicted similarity judgments during most of the clustering process until maximum height, where vegetation still explained more than 50% of the variance within the more general clusters. All R^2 values are reported in the online data repository. Analysis of variance on the whole agglomeration process revealed a main effect for assigned categories ($F(2, 116) = 33$, $p < .001$, $\eta^2 = .03$) in adults, qualified by the strongest predictor vegetation which differed from the two other categories ($p < .001$), and by a stronger predictive value of natural elements compared to artifacts, $p = .006$ (main text, Figure 7). These results show that the categories adults perceived in the images relate to their judgments of visual similarity.

A. 8 Participants' Criteria for Determining Similarity

Adults' Questionnaire on their Criteria for Assembling Images

After finishing the sorting task, adult participants received a questionnaire asking about criteria underlying their similarity judgments. Adults were asked a) to provide terms which described image similarity within their image groups in general (question 1), and b) to choose two to three of their assembled image groups and describe image similarity within each of these groups (question 2a-c).

We coded the answers by noting if they included terms which belonged to one of four qualities, defined by the variables:

- Appearance–descriptions of pattern, shape, or grey tone.
- Entity–labels of depicted objects.
- Haptic–adjectives which describe experiences with the depicted objects.
- Paraphrases, which include labels of entities (e.g. "leave-like", "rock-pattern").

We then calculated the proportion in which each quality contributed to similarity decisions. This was done by dividing the total of cases in which the particular quality was mentioned by the total of all mentioned qualities (Table A. 4). We only included answers to question 2 because some of the general criteria were difficult to understand, and because answers to question 1 included terms which were repeated in the examples of question 2.

Table A. 4

Qualities related to similarity judgments

Quality	Frequency	Proportion [a]
Children [b]		
Appearance	53	0.62
Entity	33	0.38
Adults [c]		
Appearance	152	0.55
Entity	76	0.27
Haptic	17	0.06
Paraphrase	33	0.12

[a] Calculated by dividing the frequency of the quality by the total of qualities in the participant group.
[b] Children's spontaneous comments on the similarity between images were video-recorded or noted.
[c] Obtained by a questionnaire asking adult participants about criteria underlying their similarity judgments.

Children's Comments during Similarity Sorting

In order to assess children's criteria on what they perceived as similar, we had video-recorded children during the sorting task. If caregivers did not agree to video-recording, we had taken notes of children's spontaneous comments on the images during the sorting task. We then coded all comments by first separating them according to the context in which they occurred, and the intention we assumed behind the comment (e.g. describing similarity between images versus naming an object or describing an impression independent of similarity to another image). In the analysis of the comments, we only included those that referred to the similarity between images. Because the quality of the terms children used were more difficult to categorize than those of adults, we only assessed the qualities: a) appearance, and b) entity because they were of great interest for our analysis. The assessment of criteria determining similarity indicates that children and adults attended to visual appearance as well as the depicted entities during the sorting task.

Adults' self-reported criteria underlying similarity judgments included only 55% of terms which related to visual properties, whereas the remaining terms related to the identity of or experiences with the depicted entities. In children, visual properties were mentioned in 62% of the comments referring to visual similarity, **Fehler! Verweisquelle konnte nicht gefunden werden.**. The equivalence of semantic and property-related perception is additionally supported by a detailed comparison of the R^2 values referring to visual properties and categories. In western cultures, photographic representations are understood to include at least two levels of information, which either relate to the image object itself (the sorting card), or to the entity which it represents. Children become acquainted with this cultural habit from an early age (DeLoache, 2011; Liben, 2003). Independent of age, it might have been difficult to ignore the referent of the image but exclusively attend to its visual properties. Additionally, higher-order properties are frequently associated with experiences of physical states (e.g., soft, fluid, grainy). These experiences might also be closely entwined with particular categories, making differentiation within them difficult.

A. 9 Can Variance of Visual Properties be Separated from Variance of Assigned Categories in the Sorting Task?

We assessed and presented R^2 values of visual properties and of assigned categories–both are predicting the participant's similarity decisions Figure A. 5 Figure A. 6). One can argue that assigned categories and visual properties are not independent of each other, and that it is not clear whether participants attended to the category of an image, or the visual properties which are predicting the category. If categories were primarily attended to, then the R^2 values of the visual properties should develop in patterns which are congruent to those provided by the R^2 values of the categories. We therefore evaluated the relationship between assigned categories and visual properties by visual inspection and did not find a clear relationship. For example, children predominantly relied on depth, skew and deviation in their assignment of artifacts and vegetation. In the sorting task, these properties had elevated impact on similarity perception in accordance with higher values of assigned vegetation and artifacts above natural elements (height 1.1 to 1.2). This gives the impression that visual properties illustrate the impact of assigned categories. However, around height 1.4 (6 clusters), when vegetation is elevated high above the other properties, predictors of assigned vegetation only play a secondary role, while alpha–which predicted natural elements–increased its impact. Concerning the general sequence, we found that symmetry which was predicting children's assignment to natural elements and vegetation only played a minor role in their similarity decisions. In contrast, CooCor which was *not* found to predict category assignment, explained a moderate to high proportion of variance during the children's clustering hierarchy, compared to the other properties. This inspection gives the impression that assigned categories and visual properties do not play an exclusive role for similarity judgments, but were attended in parallel.

Visual inspection of the relationship between assigned categories and visual properties in the adult sorting task did not reveal a clear overlap. Recall that depth, skew and deviation had been found to predict adults' assignment to vegetation. Between height 1.7 and 2.6 (5 to 2 clusters), depth and deviation were elevated in parallel to assigned vegetation. However, skew generally had a minor impact on similarity perception in adults. Moreover, regularity, which did not reach significance in the adults' identification task, was one of the properties with the highest amount of explained variance value in the sorting task (main text, Figure 7). As with children, these examples show that the impact of particular visual properties cannot be fully attributed to participants' inclusion of assigned categories.

Nevertheless, we cannot exclude that relationships exist which cannot be observed in this way. Alternative explanations of the partial overlap are discussed in the main discussion section.

A. 10 How True Categories Predict Similarity Judgments

In Post-hoc analysis we additionally assessed how variance of the similarity-sorting was explained by the true categories.

In Figure A. 6 the true categories' R^2 values are plotted as a function of the height of the dendrogram, to illustrate the development of true categories' predictive value on children's and adults' sorting decisions.

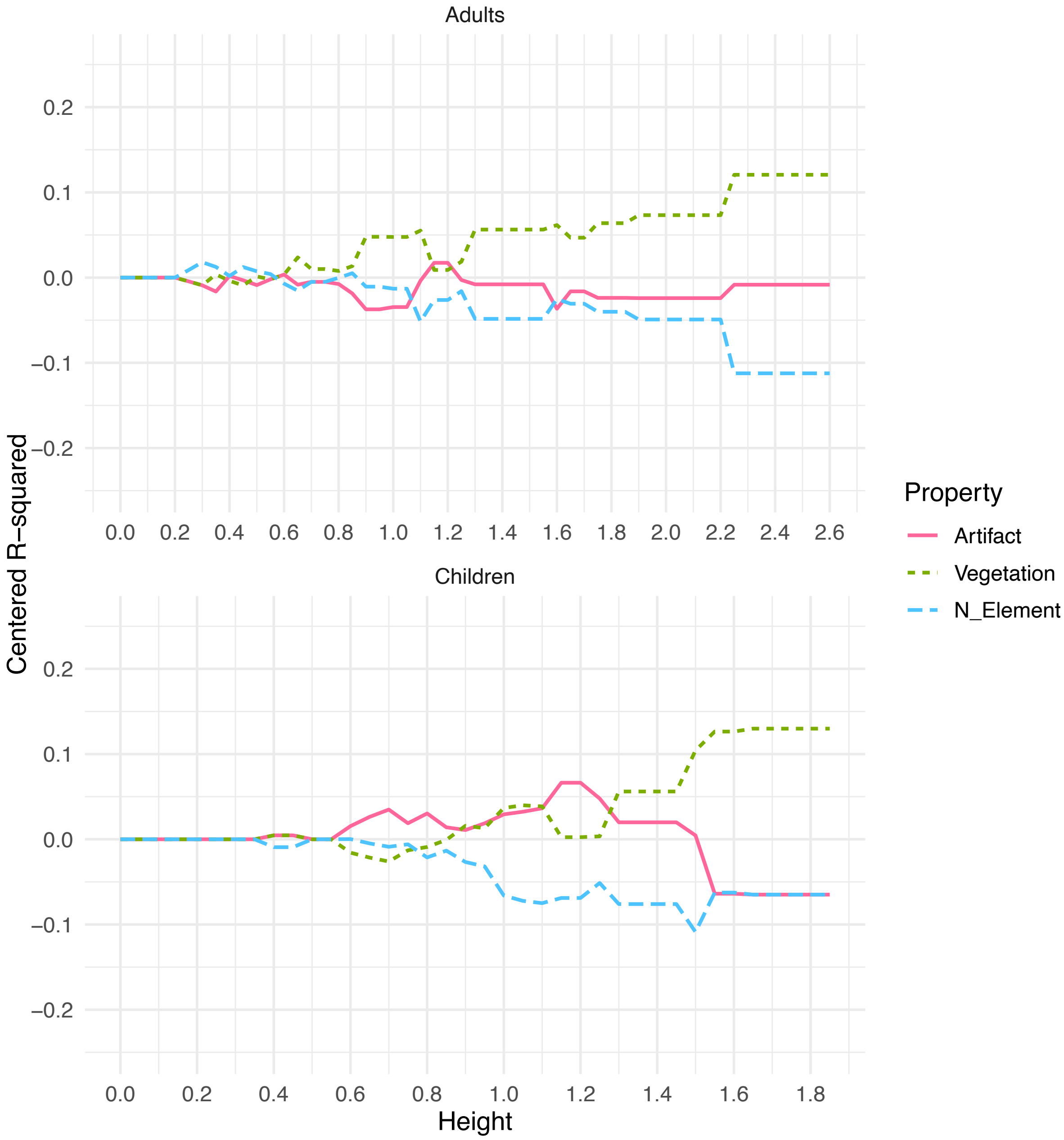

Figure A. 6. Explained variance of the images' true categories by participants' sorting.

Levels of R^2 for true categories as a function of the height of the dendrogram, for adults (top) and children (bottom). Height (x-axis) indicates the inter-cluster distance of images merged within successive hierarchical clusters. R^2 (y-axis) represents the impact of each of the image characteristics on similarity judgments (i.e. explained variance), separately for the true image categories in the classification task.
R^2 was assessed for each step in the clustering hierarchy by calculating the proportion of variability between the individual clusters to the total variability of the image property (here: the true category; for a similar approach see: Friesen et al., 2015). Higher levels of the resulting R^2 values indicate a stronger variation of the image property between clusters,

interpreted as a stronger predictive value of the property on participants' sorting decisions. For the figure, R^2 was assessed in steps of height .05, and R^2 values were centered for each indexed height.

We compared the R^2 values of the HCA's agglomeration process (60 to 2 Clusters) within children and adults, respectively. In children, analysis of variance on R^2 within the factor true category revealed a main effect for true category ($F(2, 116) = 22$, $p < .001$, $\eta^2 = .017$), qualified by a lower predictive value of true natural elements which differed from the two other categories ($p < .001$), while true artifacts did not differ from true vegetation (Figure A. 7, bottom).

In adults, the ANOVA on R^2 showed a main effect for true category ($F(2, 116) = 14$, $p < .001$, $\eta^2 = .005$), qualified by a higher predictive value of vegetation than the two other categories ($p < .001$), while true artifacts did not differ from natural elements (Figure A. 7, top).

We also compared the predictive value of true categories on similarity judgments between the participant groups in a 3 × 2 ANOVA on R^2 values. A main effect for true category ($F(2, 232) = 24$, $p < .001$, $\eta^2 = .007$) was qualified by the true natural elements' lower predictive value compared to true artifacts and true vegetation, which did not differ. There was no main effect for participant group ($F(1, 116) = 1.4$, $p = .24$), indicating that children and adults similarity judgments were similarly predicted by the true categories. However, a significant interaction between participant group and true category was revealed ($F(2, 232) = 15$, $p < .001$, $\eta^2 = .004$) qualified by lower predictive values of true natural elements and true vegetation in children compared to adults, respectively, Figure A. 7.

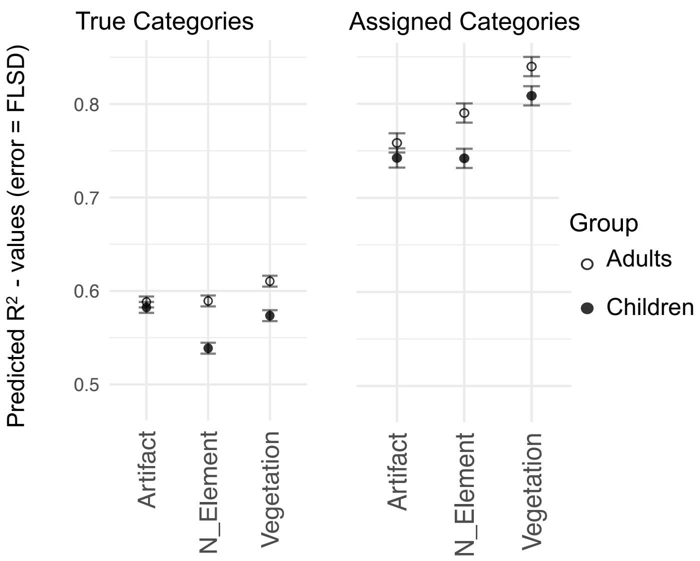

Figure A. 7. Comparison of R^2 values of the similarity-sorting data as functions of true and assigned categories, and participant group.

Predicted means of R^2 values in the ANOVAs of True categories × Group (left), and of Assigned categories × Group (right). R^2 values (i.e., explained variance) were obtained from the HCA on similarity judgments at each step in the clustering process (60 to 2 clusters).

Note that for the *assigned* categories, vegetation had a stronger predictive value in children and adults than artifacts and naturel elements, which did not differ, and the interaction term between participant group and assigned category did not reach significance (Figure A. 6 and main results section). Thus, the true categories predicted the participants' similarity judgments with a different pattern than the assigned categories. Because of the ceiling effect in the adults' classification (90.5% correct classifications, compared to 69.6% correct classifications in children), the difference in the predictive value of between true and assigned categories for similarity judgments can be mainly attributed to children's assumptions about category membership. Moreover, the predicted values of true categories on similarity judgments are generally lower than those of assigned categories (Figure A. 7), suggesting that an image's assumed category membership related more strongly to judgments of visual similarity than the actual image category.

A. 11 Additional Analysis: The Effect of Children's Age on the Inclusion of Visual Properties during Identification

During the analysis of the identification task, we compared which visual properties predicted the assignment of categories in children and adults. We had included the covariate age in the analysis of the children's identification task, because of its significant impact on a child's general performance in this task (i.e., Spearman correlation of correctly identified images; $r(226) = .21, p < .001$). The findings indicated that some visual properties which relied on detailed visual information were not included in an adult-like way by children. To receive better understanding of this finding, we decided to analyze the relationship between the inclusion of a visual property and the age of a child. This analysis could show, if younger children were including less visually-detailed information in their decisions than older children and support our interpretation. We ran additional GLMMs including the continuous variables age, visual property, and the interaction term Age × Visual Property. We ran separate models for each of the assigned categories and each particular visual property (further information about procedures and software are provided in the main result section).

After adjusting *p*-values (Benjamini & Hochberg, 1995), we found significant interactions between age and visual property for the visual properties regularity, size, and symmetry on both of the assigned categories artifacts and natural elements, and, moreover, for the visual properties curvature, deviation and CooCor on artifacts (all $p < .05$). No properties led to an interaction with age in assigned vegetation. The directions of the effects are shown in Figure A. 8.

These results indicate that with regard to images of artifacts and natural elements, preschool children's inclusion of some visual properties drawn upon during categorization changed with age. Moreover, depth and skew, which were the strongest predictors of children's similarity judgments and did not differ between adults and children in the sorting task also do not show differences between younger and older children during identification.

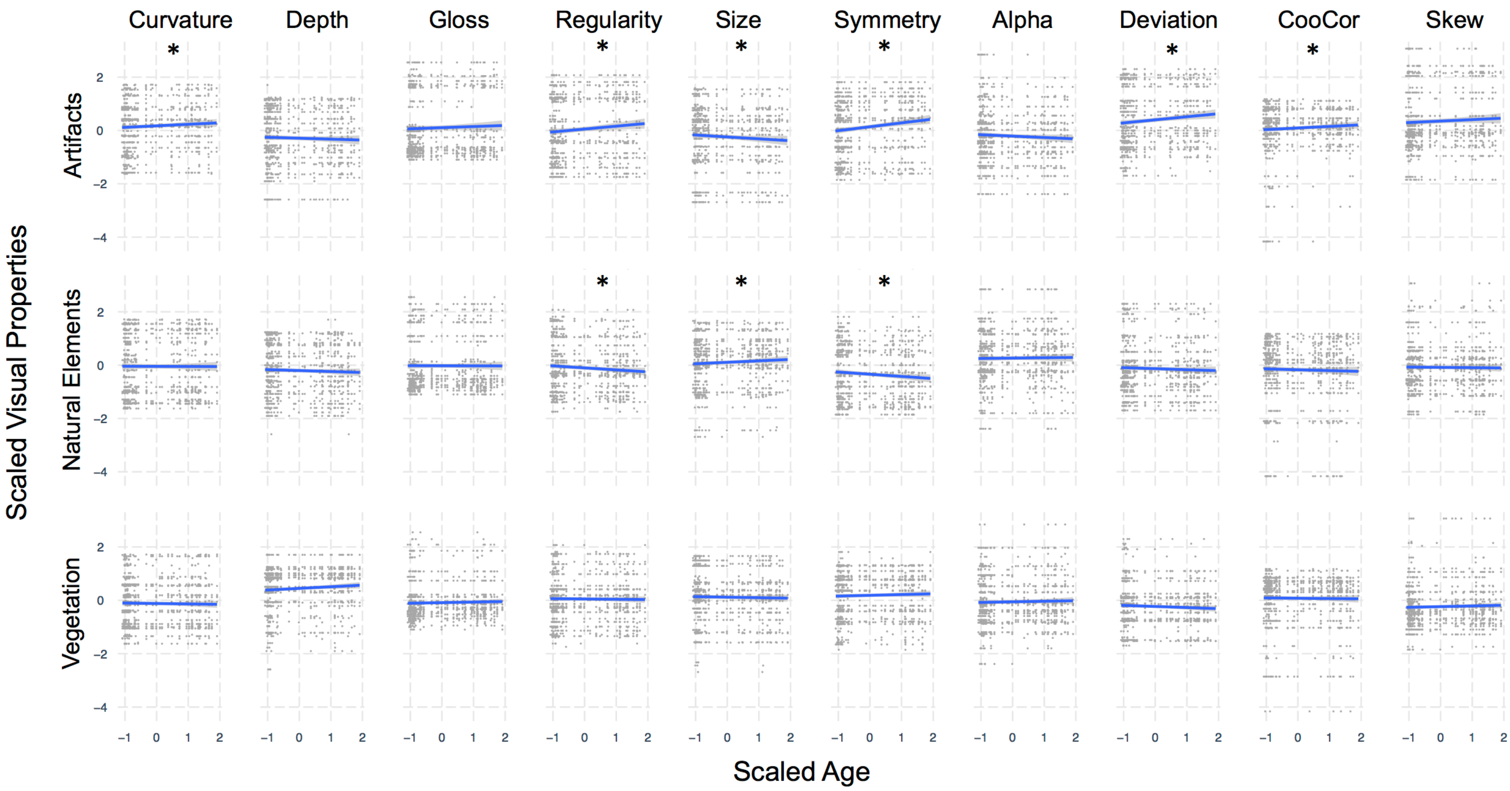

Figure A. 8. Visual Properties as Function of Children's Age and Assigned Category

Note. Error bars are *SE*.
* adjusted *p* < .05 (method: Benjamini and Hochberg, 1995).

A. 12 Property Variance Within Categories

It might be argued that the images showing vegetation were perhaps more similar within their category than the images of the other categories. Consequently, this would have led to an increase in assembled vegetation images. Although we had analyzed the categories to which the images were assigned instead of the true categories, this argument could still apply to the relatively large number of correctly classified images in children and adults. We therefore aimed to statistically evaluate whether the variance of the current set of properties differed between the categories by conducting an analysis of variance on the combined visual-property values in the image data. The ANOVA did not reveal a main effect for the factor category: $F(2, 8) = 1.1$, $p = .37$, *n.s*; Means (*SD*) of artifacts = .09 (1.1), natural elements = -

.14 (1), vegetation = .05 (.8). This analysis shows that statistically, there were no differences in the variance of the visual properties between the categories

A. 13 Data Link

The data underlying the statistical analysis of this study is accessible under the link https://osf.io/8xy5n/?view_only=6ddced286c31456fae7d20dd86e072e6.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300.

Burton, G. J., & Moorhead, I. R. (1987). Color and spatial structure in natural scenes. *Applied Optics*, *26*(1), 157–170.

Clausi, D. A. (2002). An analysis of co-occurrence texture statistics as a function of grey level quantization. *Canadian Journal of Remote Sensing*, *28*(1), 45–62.

Costa, A. F., Humpire-Mamani, G., & Traina, A. J. M. (2012). An efficient algorithm for fractal analysis of textures. *Graphics, Patterns and Images (SIBGRAPI), 2012 25th SIBGRAPI Conference On*, 39–46.

DeLoache, J. S. (2011). Early development of the understanding and use of symbolic artifacts. In U. C. Goswami (Ed.), *The Wiley-Blackwell handbook of childhood cognitive development* (Vol. 2, pp. 312–336). Blackwell Publishers Ltd.

Friesen, M. C., Shortreed, S. M., Wheeler, D. C., Burstyn, I., Vermeulen, R., Pronk, A., Colt, J. S., Baris, D., Karagas, M. R., & Schwenn, M. (2015). Using hierarchical cluster models to systematically identify groups of jobs with similar occupational questionnaire response patterns to assist rule-based expert exposure assessment in population-based studies. *Annals Of Occupational Hygiene*, *59*(4), 455–466.

Gonzalez, R. C., & Woods, R. E. (2018). *Digital image processing*. Pearson.

Groll, A., & Tutz, G. (2014). Variable selection for generalized linear mixed models by L 1-penalized estimation. *Statistics and Computing*, *24*(2), 137–154.

Isherwood, Z. J., Schira, M. M., & Spehar, B. (2017). The tuning of human visual cortex to variations in the 1/f α amplitude spectra and fractal properties of synthetic noise images. *NeuroImage*, *146*, 642–657. https://doi.org/10.1016/j.neuroimage.2016.10.013

Liben, L. S. (2003). Beyond point and shoot: Children's developing understanding of photographs as spatial and expressive representations. In R. V. Kail (Ed.), *Advances in child development and behavior* (Vol. 31, pp. 1–43). Elsevier Science.

Materka, A., & Strzelecki, M. (1998). Texture analysis methods–a review. *Technical University of Lodz, Institute of Electronics, COST B11 Report, Brussels*, *10*(1.97), 4968.

Mather, G. (2018). *Visual Image Statistics in the History of Western Art*. https://doi.org/10.1163/22134913-20181092

Murtagh, F., & Legendre, P. (2011). Ward's hierarchical clustering method: Clustering criterion and agglomerative algorithm. *ArXiv Preprint ArXiv:1111.6285*.

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Revelle, W. (2018). *psych: Procedures for Psychological, Psychometric, and Personality Research* (1.8.10) [R]. Northwestern University. https://CRAN.R-project.org/package=psych

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*(301), 236–244.

Wickens, T. D. (2002). *Elementary signal detection theory*. Oxford University Press, USA.