Children's use of visual summary statistics for material categorization

Benjamin Balas

Department of Psychology and Center for Visual and Cognitive Neuroscience, North Dakota State University, Fargo, ND, USA

Although adults' ability to recognize materials from complex natural images has been well characterized, we still know very little about the development of material perception. When do children exhibit adult-like abilities to categorize materials? What visual features do they use to do so as a function of age and material category? In the present study, we attempted to address both of these issues in two experiments that we administered to school-age children (5–10 years old) and adults. In both tasks, we asked our participants to categorize natural materials (metal, stone, water, and wood) using original images of these materials as well as synthetic images made with the Portilla–Simoncelli algorithm. By including synthetic images in our stimulus set, we were able to assess both how material categorization develops during childhood and how visual summary statistics are recruited for material perception across age groups. We observed that when asked to provide category labels for individual images (Experiment 1), young children were disproportionately bad at categorizing some materials after they were synthesized, suggesting material-specific changes in information use over the course of development. However, when asked to match real and synthetic images according to material category without labeling (Experiment 2), these effects were weakened. We conclude that while children have adult-like abilities to encode and compare images based on summary statistics, the mapping between summary statistics and category labels undergoes prolonged development during childhood.

Introduction

Adult observers are capable of recognizing images of natural materials quickly and accurately (Fleming, 2013). Relative to different types of object recognition, for example, material categorization from natural images is as fast as basic-level object categorization (Wiebel, Valsecchi, & Gegenfurtner, 2013). Indeed, even when image presentation times are as short as 40 ms, adults are able to achieve impressive levels of accuracy (~80% correct), even though material categorization appears to require computations beyond simple analyses of color, lightness, or texture properties (Sharan, Rosenholtz, & Adelson, 2014). Adults can maintain a broad range of material classes (Fleming, Wiebel & Gegenfurtner, 2013) and are able to assign natural images to these classes efficiently. This behavioral profile of material categorization performance is matched by equally impressive neural responses. BOLD responses elicited by images of natural materials reflect both image-based representations of materials and higher-level representations of material appearance that reflect perceptual similarity (Hiramatsu, Goda, & Komatsu, 2011). The application of multivoxel classification techniques to functional magnetic resonance imaging (fMRI) data has further revealed multiple areas that carry diagnostic information regarding material category, including early visual areas (Baumgartner & Gegenfurtner, 2016) and higher-order cortical loci like the parahippocampal gyrus (Jacobs, Baumgartner, & Gegenfurtner, 2014). Similar classification techniques applied to event-related potentials (ERP) data demonstrate that the timing of diagnostic neural responses is fairly fast: After approximately 140 ms, ERP responses elicited during a Go/No-Go judgment are sufficient to distinguish between natural material categories regardless of task demands (Wiebel, Valsecchi, & Gegenfurtner, 2014). Overall, material categorization during adulthood is for the most part fast and accurate, considered both in terms of behavioral performance and neural indices of high-level processing (though this does vary somewhat across studies; see Sharan, Rosenholtz, & Adelson, 2014). Moreover, both domains of research have revealed that there are low-level properties (color, in particular) that carry useful information for material categorization, but that there are also higher order features that contribute to performance as well (Fleming, 2013). The utility of low-level features versus high-level representations of appearance differs according to the particular material judgment being considered, but it seems

Citation: Balas, B. (2017). Children's use of visual summary statistics for material categorization. *Journal of Vision*, *17*(12):22, 1–11, doi:10.1167/17.12.22.

doi: 10.1167/17.12.22 Received April 2, 2017; published October 31, 2017 ISSN 1534-7362

fair to say that there are useful tools for categorizing and discriminating between materials across multiple levels of representation in the visual system.

Although adults' abilities to categorize natural materials have been well-characterized, there is a remarkable paucity of data describing how material categorization develops. Indeed, the majority of developmental research relevant to visual material categorization either focuses on the perception of surface properties like gloss, roughness, and other aspects of surface reflectance, or is instead focused on characterizing children's sensitivity to texture properties in abstract images. For example, infants are capable of distinguishing between yellow and gold surfaces (Yang, Kanazawa, & Yamaguchi, 2013), which suggests that specularities can be used early in visual development to discriminate between matte appearance and shiny materials. More general sensitivity to surface gloss is also evident in infancy (Yang, Otsuka, Kanazawa, Yamaguchi, & Motoyoshi, 2011), further demonstrating that reflectance properties are available to infants in the first year of life. These abilities obviously reflect some extant ability to categorize and or discriminate natural materials, but do not necessarily provide a great deal of insight into how a broad range of material classes (e.g., water, wood, plastic, etc.) are categorized or distinguished. In particular, some surface properties like gloss (Sharan, Motoyoshi, Nishida, & Adelson, 2008; Weibel, Toscana, & Gegenfurtner, 2016) and roughness (Padilla, Drbohlav, Green, Spence, & Chantier, 2008) can be estimated reasonably well using low-level image statistics, which could mean that young infants can use low-level visual features to solve some material categorization problems but lack higher level representations of materials. Without more extensive study of infant performance across a broader range of tasks (and with an emphasis on material categorization rather than the recovery of reflectance properties), it is difficult to say (though see Yang et al., 2015 for results concerning infants' performance with synthetic textures similar to the ones used here). Older children's abilities to recognize natural materials are even less specified in the literature, though there is evidence suggesting that texture perception may change during school-age years, which may in turn have consequences for material categorization. Texture segmentation abilities as well as search performance that depends on texture cues changes during middle childhood, for example (Sirteneau & Rieth, 1992), though this has only been demonstrated with abstract textures made of discrete, structured elements. In this same age range (9 to 10 years old), children's sensitivity to power-law coefficients in fractal noise also becomes more adult-like (Ellemberg, Hansen, & Johnson, 2012), which further suggests that there is important development of texture processing mechanisms during middle childhood. Again, however, these results were obtained from artificial stimuli depicting fractal noise with different power spectra (e.g., pink noise vs. white noise), making it difficult to generalize from these results to a more ecologically valid setting. Still, these results, imply that material categorization may develop gradually during childhood. Further, in other domains, children exhibit failures to integrate visual information across space that may point to more general developmental processes that also constrain material perception. For example, Kovacs, Kozma, Feher, and Benedek (2009) demonstrated that young children have profound difficulties interpreting two-tone images of complex objects, in some cases even after being shown a grayscale version of the same image. This inability to integrate structural information and achieve visual closure in two-tone images may also reflect limits on pooling operations like those that have been hypothesized to support texture processing, which in turn supports material perception. Regarding material perception itself, what is largely unknown is how children's development unfolds in the context of natural images, and also how visual features at different levels of complexity contribute to material categorization as a function of age.

Our goal in the current study was therefore to examine how material categorization develops during childhood, with an emphasis on children's and adults' abilities to use specific visual features for identifying materials in natural images. Specifically, we chose to work with children between the ages of 5 and 10 years old, largely because texture processing appears to be undergoing development during this age range and also because other recognition tasks (e.g., face recognition, see Pascalis et al., 2011) also undergo development during this span in terms of what visual information is being used for categorization and discrimination tasks. The simplest questions we can ask via these participants are (1) What is the developmental time course of material categorization during childhood?, and (2) Does performance with different material categories develop on different trajectories? Besides these basic questions about how material perception develops, we also wanted to understand how the use of specific visual features may also change during childhood. To investigate this aspect of visual development, we tested child and adult participants using both original images of natural materials and synthetic images of the same materials created using the Portilla–Simoncelli texture synthesis model (Portilla & Simoncelli, 2000). This model has been increasingly used as a tool for investigating human vision in multiple ways, and here we apply it to our images with the goal of determining how "summary statistics" for texture appearance support material categorization across development.

The synthetic images we created using the algorithm closely match the original images in terms of the aggregated texture statistics that define the P–S model, but lack higher-order features that capture extended features and global image structure. By comparing performance with real images and synthetic images, we are thus able to determine how the lack of these more complex features affects children and adults as a function of material category. Should the removal of these features incur a substantial performance cost, this would suggest that material categorization depends critically on higher-order features. If instead the removal of these features does not disrupt performance much, the alternative account is that material categorization is largely accomplished via summary statistics and does not require additional information. Whether or not this reliance on summary statistics versus high-order features changes with age or with material category is the key question we hoped to address.

In two experiments, we asked our participants to make material judgments using natural images of water, wood, metal, and stone. In Experiment 1, we used a simple four-alternative forced choice (4AFC) categorization task to describe children's and adults' abilities to recognize natural and synthetic materials. In Experiment 2, we carried out a material discrimination task with the same images in order to evaluate the extent to which the need to map materials to labels constrained children's abilities to process material images across categories and appearance conditions. In both experiments, we find evidence that material categorization does develop during childhood, both in terms of information use and possibly also in a material-specific fashion.

Experiment 1

In our first experiment, we examined children's ability to use summary statistics for material categorization relative to adults. Participants in this task completed a 4AFC categorization task using natural and synthetic images of real-world materials.

Methods

Participants

We recruited 40 children [5–7 year-olds (n = 20) and 8–10 year-olds (n = 20)] as well as 20 adults (aged 18–25) to take part in this experiment. Children were recruited from the greater Fargo–Moorhead community and adults were recruited from the North Dakota State University (NDSU) Undergraduate Psychology study pool. All participants reported either normal or corrected-to-normal vision. Prior to participation in the experiment, we obtained written informed consent from a legal guardian (child participants) or from the participants (adult participants). Children older than seven years provided written assent to participate as well. Consent procedures were consistent with the principles stated in the Declaration of Helsinki, as approved by the NDSU Institutional Review Board.

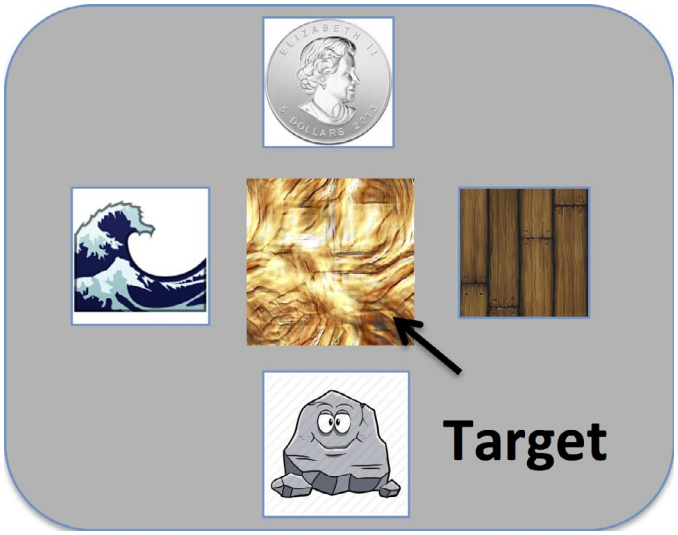

Figure 1. A schematic view of the material categorization task used in Experiment 1. Participants were presented with either a real or synthetic image in the center of the screen and chose the response image associated with the category they wished to label it with.

Stimuli

We selected a total of 64 full-color images from the Flickr Materials Database (Sharan, Rosenholtz, & Adelson, 2009) drawn from the Metal, Stone, Water, and Wood categories. Within each category, we selected 16 original images per category to serve as the basis for our stimulus set (Figure 1), choosing images to promote appearance variability within categories (diverse colors, orientations, etc.).

We cropped each original image to a size of 512 by 512 pixels and created synthetic versions of each stimulus using the default parameters of the Portilla–Simoncelli texture synthesis algorithm (Liang, Simoncelli & Lei, 2000; Portilla & Simoncelli, 2000). Briefly, this model adjusts an initial noise image until it matches the texture statistics of the target image. The texture statistics used to characterize target images are comprised of a number of correlations measured between wavelet coefficients at different positions, orientations, and spatial scales. See Balas, Nakano, and Rosenholtz (2009) for a more thorough discussion of the model.

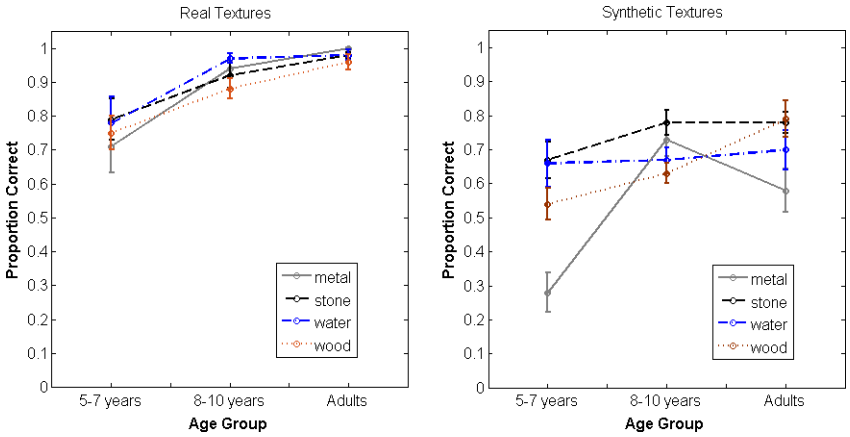

Figure 2. Average performance (proportion correct) across all age groups and material categories for real (left) and synthetic (right) textures in Experiment 1. Error bars represent ± *SEM*.

Procedure

Participants from all age groups completed a 4AFC material categorization judgment using the real and synthetic images described already. On each trial, we presented participants with a single image depicting either a real image or a synthetic image from one of the four material categories we selected. Participants were asked to categorize the target image according to material category by touching one of the four cartoon response images arranged around the target image (Figure 1). Participants had unlimited time to respond to each image and did not receive feedback regarding their responses. Participants' eye movements and head position were neither constrained nor monitored during the task (free viewing).

Participants completed the task in a darkened room using an 800-by-600 Elo touchscreen display (ELO Touch Solutions, Inc., Milpitas, CA) and were seated at a comfortable reaching distance. Although this does mean stimulus size varied somewhat across our age groups, target images subtended approximately 5°–7° of visual angle. Stimulus images were presented in a pseudorandomized order for each participant and participants saw each image only once for a total of 128 trials in the experimental session. All stimulus presentation and response collection routines were implemented using the Psychophysics Toolbox Version 3 library for MATLAB (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997).

Results

We calculated the proportion of correctly labeled images in each category and analyzed these values using a mixed-design 2x3x4 analysis of variance (ANOVA) with texture appearance (real or synthetic) and material (metal, stone, water, and wood) as within-subjects factors and age group (5–7 years old, 8–10 years old and adults) as a between-subjects factor. In Figure 2 we display average accuracy as a function of all three of these factors.

This analysis revealed main effects of texture appearance, $F(1, 56) = 324.24$, $p < 0.001$, partial $\eta^2 = 0.85$; material category, $F(3, 168) = 5.55$, $p = 0.001$, partial $\eta^2 = 0.09$; and age group, $F(2, 56) = 15.85$, $p < 0.001$, partial $\eta^2 = 0.36$. The main effect of texture appearance was the result of higher performance for real textures [M = 0.89, 95% CI = [(0.85–0.92)] than synthetic textures [M = 0.65, 95% CI = (0.62–0.68)]. The main effect of material category was the result of significant differences between performance with metal images (M = 0.71) relative to stone images [M = 0.82, 95% CI of the difference = (0.05–0.18)] and performance with metal relative to water images [M = 0.76, 95% CI of the difference = (0.004–0.18)]. Finally, the main effect of age group was the result of significantly lower accuracy in the 5- to 7-year-old age group [M = 0.65, 95% CI = (0.59–0.70)] relative to both the 8- to 10-year-old age group [M = 0.82, 95% CI = (0.76–0.87)] and adults [M = 0.85, 95% CI = (0.79–0.90)].

These main effects were qualified by significant two-way interactions between texture appearance and material category, $F(3, 168) = 11.32$, $p < 0.001$, partial $\eta^2 = 0.17$, and between material category and age group, $F(6, 168) = 2.91$, $p = 0.010$, partial $\eta^2 = 0.09$. These were further qualified by a significant three-way interaction between all three factors, $F(6, 168) = 4.48$, $p < 0.001$, partial $\eta^2 = 0.14$). To examine the nature of this interaction, we carried out separate two-way ANOVAs for real and synthetic images so we could determine how material category and participant age affected performance differently when texture appearance was natural or artificial. For real images, the only significant effect we observed was a main effect of age group ($F(2, 65) = 13.3$, $p < 0.001$, partial $\eta^2 = 0.32$). Neither the main effect of material ($p = 0.40$) nor the

interaction between material and age group ($p = 0.64$) reached significance. For synthetic images, however, we observed main effects of both material category, $F(3, 168) = 9.67$, $p < 0.001$, partial $\eta^2 = 0.15$), and age group, $F(2, 56) = 14.2$, $p < 0.001$, partial $\eta^2 = 0.97$, and a significant interaction between these factors, $F(6, 168) = 4.44$, $p < 0.001$, partial $\eta^2 = 0.14$). Post-hoc t tests revealed that this interaction appears to be driven by significant differences between young children's performance and adult's performance with both metal and wood images, whereas other material categories yielded no significant age effects.

Discussion

Briefly, these results demonstrate that there are aspects of material categorization that develop during childhood. That children get better with age is, of course, not surprising and certainly reflects more general cognitive development to some extent. What is much more interesting is that the way information is used for material categorization changes with age. Specifically, although neither material category nor synthetic appearance does much to impact performance with real textures in an age-dependent way, these factors do affect performance with synthetic images differently in our youngest observers. The ability to use summary statistics for material categorization appears to develop at different rates for different material categories, suggesting material-specific development of representations for material perception.

We will expand upon the consequences of these results in the General discussion and offer some ideas as to what we think they may mean. For now, however, we raise an important issue that Experiment 1 does not allow us to address. Poor performance with synthetic images in this task could imply a few different things and we would like to try and distinguish between these possibilities as much as we can. One way to explain what we have seen here is to say that children may be unable to measure summary statistics well in all of the material categories we have used here, so they do not have as much useful information as adults to use for labeling synthetic images. Alternatively, they may be measuring summary statistics just as well as adults in all cases, but lack the necessary mapping between summary measurements and material categories. In Experiment 2, we attempt to distinguish between these possibilities by asking another group of children and adults to complete a material matching task using real and synthetic images. By removing the need to assign explicit labels to materials, we remove the requirement that participants maintain any sort of mapping between summary statistics and material categories. If this is the critical aspect of material perception that is developing in childhood, the critical three-way interaction we have observed in Experiment 1 should disappear. On the other hand, if young children are not able to measure summary statistics as effectively as adults, performance in this task should look much like performance in Experiment 1.

Experiment 2

In our second experiment, we examined material categorization abilities using a discrimination task. By removing the need to ask participants to label individual images according to category, we hoped to determine whether children's abilities to categorize real and synthetic materials were limited by their ability to assign the right name to a candidate image, or by their ability to distinguish materials based on visual content.

Participants

As in Experiment 1, we recruited a total of 40 children [5- to 7-year-olds (n = 20) and 8- to 10-year-olds (n = 20) and 20 adults to take part in this task. None of these individuals had taken part in Experiment 1, and all participants reported normal or corrected-to-normal vision. All procedures for obtaining informed consent were the same as described in Experiment 1.

Stimuli

We used the same stimulus images as described previously in Experiment 1.

Procedure

Participants in this experiment were asked to make a match-to-sample judgment based on material characteristics. On each trial, we simultaneously presented a *sample image* at the top of the screen and two *test images* at the bottom of the screen (Figure 3). One of the test images was a different exemplar from the same material category as the sample, and the other was an exemplar of a different material. All of the images presented on each trial were either real textures or synthetic textures. Participants chose which test image they believed matched the sample according to material category by touching the test image and had unlimited time to make their response.

We presented participants with a total of 200 trials (25 trials per condition) presented in a pseudorandomized order. All stimulus parameters and testing procedures were identical to the methods described in Experiment 1.

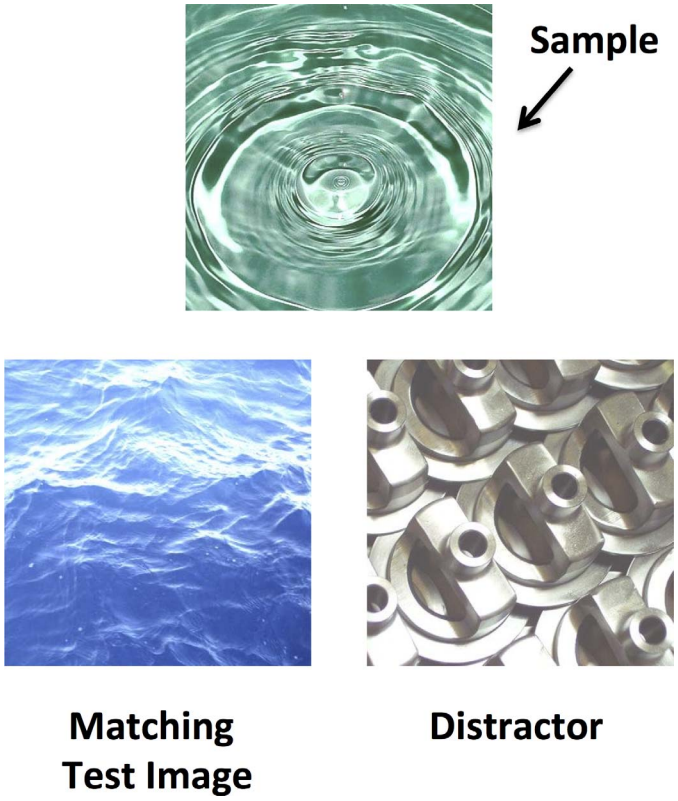

Figure 3. A schematic view of the material discrimination task used in Experiment 2. Participants were presented with a sample image depicting a real or synthetic texture and were asked to choose the test image that depicted the same material. All three images were either real or synthetic.

Results

As in Experiment 1, we calculated the proportion of correct trials in each material category and analyzed these values using a mixed-design 2x3x4 ANOVA with texture appearance (real or synthetic) and material (metal, stone, water, and wood) as within-subjects factors and age group (5- to 7-year-olds, 8- to 10-year-olds, and adults) as a between-subjects factor. In Figure 4 we display average accuracy as a function of all three of these factors.

This analysis revealed significant main effects of texture appearance, $F(1, 55) = 309.7$, $p < 0.001$, partial $\eta^2 = 0.85$, and material category, $F(3, 165) = 7.59$, $p < 0.001$, partial $\eta^2 = 0.12$, as well as a main effect of age group, $F(2, 55) = 7.22$, $p = 0.002$, partial $\eta^2 = 0.21$. The effect of texture appearance was the result of better accuracy for real images [M = 0.88, 95% CI = (0.85–0.90)] than synthetic images [M = 0.67, 95% CI = (0.65–0.69)], and the effect of material category was the result of better performance with images of water [M = 0.82, 95% CI = (0.80–0.84)] than all three of the other categories. Finally, the main effect of age group was the result of significantly lower performance in 5- to 7-year-olds [M = 0.72, 95% CI = (0.68–0.76)] than both 8- to 10-year-olds [M = 0.80, 95% CI = (0.76–0.84)] and adults [M = 0.81, 95% CI = (0.77–0.84)].

These main effects were qualified by significant two-way interactions between texture appearance and age group, $F(2, 55) = 3.35$, $p = 0.042$, partial $\eta^2 = 0.0.11$, and between texture appearance and material, $F(3, 165) = 16.0$, $p < 0.001$, partial $\eta^2 = 0.23$. The three-way interaction between all of our factors did not reach significance. The interaction between texture appearance and age group was driven by significant differences in performance across age groups in the real images condition that were not evident in the synthetic images condition. Specifically, 5- to 7-year olds [M = 0.80, 95% CI = (0.76–0.85)] performed more poorly than 8- to 10-year-olds [M = 0.92, 95% CI = (0.87–0.96)] and adults [M = 0.92, 95% CI = (0.88–0.96)] when real images were used, but performance across all three groups did not differ when synthetic images were used [5- to 7-year-olds, M = 0.64, (0.60–0.67); 8- to 10-year-olds, M = 0.68, (0.64–0.72); adults, M = 0.70, (0.66–0.73)].

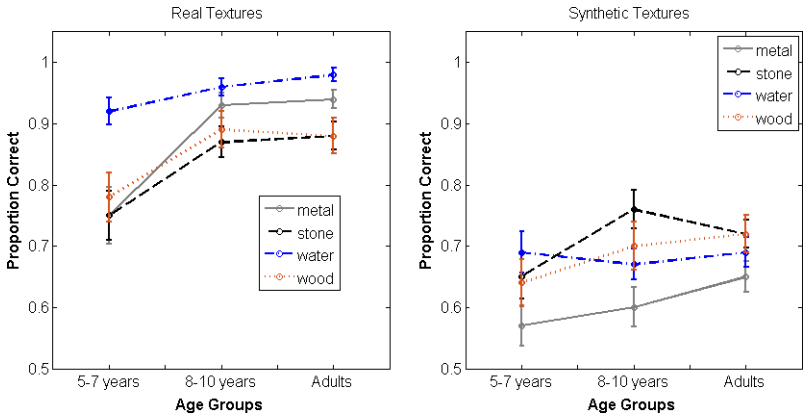

Figure 4. Average performance (proportion correct) across all age groups and material categories for real (left) and synthetic (right) textures in Experiment 2. Error bars represent ± *SEM*.

The interaction between texture appearance and material category was the result of different effects of material category for real and synthetic textures. For real textures, performance with water images [M = 0.95, 95% CI = (0.93–0.97)] was significantly higher than all three of the other conditions [Metal, M = 0.88, (0.84–0.91); Stone, M = 0.83, (0.80–0.87); Wood, M = 0.85, (0.81–0.89)]. For synthetic images, performance with metal images [M = 0.61, 95% CI = (0.57–0.64)] was significantly lower than performance with stone [M = 0.71, 95% CI = (0.67–0.74)] and water images [M = 0.68, 95% CI = (0.65–0.71)].

General discussion

Our results from both experiments confirm prior reports that the summary statistics comprising the Portilla–Simoncelli model do not support the synthesis of true metamers for natural materials. That is, removing higher-order structures via texture synthesis leads to consistent performance decrements for material categorization (Experiment 1) and discrimination (Experiment 2), similar to the outcomes reported in a range of previous studies concerned with different aspects of texture perception and other types of "statistical" vision. For example, Wallis, Bethe, and Wichmann (2016) recently demonstrated that P–S textures were not true metamers (Freeman & Simoncelli, 2011) of natural scenes in a peripheral discrimination task and Balas, Conlin, and Shipman (2016) reported similar results in a peripheral material categorization task. Other features of peripheral encoding that are apparently not adequately captured by the P–S algorithm include the perception of numerosity in peripherally presented dot arrays (Balas, 2016) and the appearance of natural textures of various types (Balas, 2006). Thus, despite important advances in characterizing visual crowding (Balas et al., 2009) and visual search (Rosenholtz et al., 2012) in terms of summary-statistic representations, these results provide additional evidence that the particular texture code used by the P–S algorithm also has important shortcomings in settings where it may seem to offer good expressive power. By itself, this is not particularly surprising to see given that material images in both tasks were presented at the fovea in this task (not in peripheral vision like many of the studies described already), and so we would expect observers to be capable of measuring features beyond those available in P–S summary statistics, as is evident in foveal tasks requiring invariant texture recognition (Balas & Conlin, 2015b) and ERP results demonstrating sensitivity to higher-order structure at early visual components (Balas & Conlin, 2015a) Nonetheless it is an important demonstration that material images contain diagnostic information in higher-order visual features and that observers use this information for categorization. In particular, contextualizing our results this way is useful because whereas studies examining the limits of true metamerism for the P–S representation of textures (or other representations such as those obtained from deep neural nets) usually emphasize discrimination at the level of an image, the representational vocabulary of such a feature set can also be evaluated at the level of a category even if the image level representation is not high-fidelity. That is, even though we know from multiple studies that the P–S algorithm will not in general yield a synthetic image that is indistinguishable from the original when viewed foveally, which does not imply that the synthetic image does not contain sufficient information for a range of category judgments. Thus, though there was no reason to expect strong metamerism in these experiments, our results further demonstrate the limits of this representation for carrying information that reliably signals material categories.

Critically, the results of Experiment 1 demonstrate that the cost incurred by synthesizing natural materials varies as a function of age and material. Specifically, the three-way interaction we observed between these three factors supports the hypothesis that children's use of higher-order visual features for material categorization develops in a material specific way. Although performance with real textures exhibited straightforward improvement with age across materials, material categorization was disproportionately affected by synthetic appearance in our youngest age groups such that categorization of some images was at or near chance levels. For our purposes, it is not so important that texture synthesis affects some materials differently—we would anticipate that the P–S feature vocabulary would be intrinsically more useful for some kinds of images than others—but it is interesting to see that this impact differs with age, reflecting properties of the developing appearance code for material categories. Also, the disproportionate cost of synthetic appearance in young childhood is consistent with developmental outcomes observed in other domains, including face recognition. Just as there are multiple information channels available for material categorization (operationalized here as summary statistics and an unspecified set of higher-level features not explicitly matched by the P–S algorithm), face recognition has frequently been evaluated when independent information channels defined by spatial frequency (Ruiz-Soler & Beltran, 2006) or orientation energy (Dakin & Watt, 2009) are available or not. In some cases, young children's performance with suboptimal features (e.g. vertical orientation sub-bands for faces) exhibits the same kind of disproportionately low values as we have observed

here (Balas, Huynh, Saville, & Schmidt, 2015; Balas et al., 2017). This may reflect a more general principle underlying the development of representations for visual recognition: Children may quickly develop representations that depend on the "best" features for recognition (e.g., midrange spatial frequency bands and horizontal orientation energy for faces, higher-level correlations for material categorization), but only incorporate other features more slowly, leading to profound costs when more highly diagnostic features are unavailable to younger observers. We note that we cannot make strong statements about what those higher-order features are, however. The P–S algorithm works by explicitly matching a set of correlations between wavelet coefficients, but the space of candidate features that are *not* matched by the algorithm is too vast to be meaningfully commented on based on our results. Although we could cite specific candidate features that we expect may not be matched following application of the P–S synthesis routines (e.g., joint statistics describing co-occurrence of wavelet orientations and positions across spatial scales), we cannot provide any evidence that any of these candidates are features children or adults are definitely sensitive to and using for categorization. Thus, our results provide a sort of existence proof for useful information beyond the P–S descriptors without offering insights into what that information is.

The results of Experiment 2 also require that we include an important caveat when we consider the account described already. Specifically, the lack of a three-way interaction between age group, material category, and real/synthetic appearance in this task suggests that removing the need to explicitly categorize images according to material changes the nature of developmental effects. In fact, although we observed profound differences in performance across our synthetic image conditions in Experiment 1, performance with these images in Experiment 2 is flat across age groups. We suggest that this means children's ability to match summary statistics across different images is more or less mature in young childhood, but that the ability to map summary-statistic measurements to material categories is the critical aspect of visual recognition that is changing during this period. That is, our data imply that a lack of robust mapping between summary statistics and category labels at a young age contributes to the poorer performance exhibited by young children in material categorization tasks. To our knowledge, this aspect of visual recognition has not been systematically examined in other contexts, so it's possible that this is also a more general feature of development. Regardless, the present study demonstrates that while children's ability to use summary statistics for material categorization develops in a material-specific fashion during childhood, their ability to match material images according to summary statistics is stable. This latter outcome is important to interpreting the first, both because it suggests a key dissociation of distinct recognition tasks, but also because it implies that the failure to successfully categorize materials in Experiment 1 is probably not due to some failure to encode summary statistics effectively at young ages. Were this the case (that young children just were not measuring enough relevant summary statistics to categorize materials well), we would expect that this difference in encoding would lead to measurable differences in distinguishing between images based on their summary statistics. Also, while Experiment 1 and Experiment 2 did differ in terms of response format (4AFC vs. two-alternative forced choice or 2AFC), our data cannot be easily explained in terms of any main effect of task difficulty associated with that difference. Critically, the pattern of interaction effects (and their absence) across tasks does not lend itself easily to an explanation based on any main effect of task difficulty, especially given the absence of any clear ceiling or floor effects. Although comparing children's performance across tasks can be challenging given the varying demands placed on children's perceptual and cognitive abilities as a function of different testing paradigms, the current results are not easily accounted for solely by task effects. We conclude therefore, that children probably measure summary statistics (at least, *these* summary statistics) about as well as adults, but do not use them as effectively to assign material categories to images.

There are many interesting questions regarding other aspects of material categorization that our current results do not allow us to speak to. Some of these concern simple extensions of our approach to include other candidate summary statistics (Briand, Vacher, Galerne, & Rabin, 2014; Efros & Freeman, 2001; Heeger & Bergen, 1995) and a wider variety of material categories—how general are the effects we observed here? More interesting, however, is the potential for further exploring the nature of the developmental effects we observed in Experiment 1 and how they reflect the diagnosticity of summary statistics for material categorization across different materials. Specifically, we have suggested that young children may lack a representational vocabulary that includes summary statistics for some materials, depending critically on how diagnostic higher-order features are relative to summary features: If summary features are not very diagnostic, it will take longer to develop adult-like competence with them. Testing this conjecture is complicated, however, by the fact that the diagnosticity of summary statistics for material categorization will depend on both how reliably a given set of summary statistics signals a material category and how well original images belonging to some category tend to be

well-rendered by a candidate synthesis model. The former property can be measured using adult observers by asking them to carry out a task much like our Experiment 1, but the second property will require a task more like what Wallis et al. (2016) carried out in peripheral vision. They key idea is that summary statistics might seem more diagnostic either because they support a synthetic image that does not differ much from the original *or* because even a poor synthetic image carries sufficient information about appearance to disambiguate the material depicted in the parent image. Measuring both of these properties for a wide range of material categories and characterizing developmental effects relative to these characteristics would help provide important insights into how feature diagnosticity determines the content of representations for visual recognition as a function of age. In particular, our participants were only presented with images drawn from four material categories, which obviously is a meaningful limitation of the present results. Although there are practical limitations on how long testing sessions involving child participants can be, continued efforts to examine material perception developmentally would benefit from the inclusion of a broader set of material categories.

Another important limitation of the current study is that we have not considered the role that simple global features like color distributions may play in material categorization as a function of age. By synthesizing material images (and removing/disrupting a range of higher-order feature correlations), we may also be leading children to rely more heavily on low-level features like hue and saturation. We took no special measures in this study to manipulate or control color appearance across categories, opting instead to maximize color variability across images within a category such that color would not be especially useful for material categorization. An advantage of this approach is that the current results have reasonable ecological validity: In natural settings, materials' color variability is neither controlled nor manipulated and our data thus likely reflect how performance may unfold in real-world environments. An obvious drawback, however, is that differences in color variability as a function of material category may underlie our results to some extent. For example, young children in Experiment 1 were particularly bad at categorizing synthetic metal images accurately. Could that be the result of higher color variability for this category? Critically, the answer to that specific question is no. Because the P–S algorithm offers robust color histogram matching, the disproportionately poor performance with synthetic metal images cannot be explained by color variability within this category. Were this the case, we should observe the same level of performance for the real metal textures, which are instead categorized far more accurately by young children. Although we must be careful drawing strong inferences about how aspects of natural material appearance like color variability may drive performance across different material categories, we can be confident that these properties are not the basis of the interactions we observed with synthetic texture appearance. Nonetheless, by either manipulating color availability explicitly or examining confusion matrices in more detail, we could potentially also examine how developing representations for material perception recruit low-level features for recognition and discrimination. We also encourage readers to examine our full stimulus set, which is available at the following link: https://dl.dropboxusercontent.com/u/4961099/balas_jov_kidmat_images.zip, so that they may assess for themselves how various image properties may vary across the material categories we chose to use here, and the images we selected as members of that category.

Finally, we also note that by using texture synthesis as a tool for "lesioning" natural images such that some classes of visual features are excluded from some of our stimulus conditions, we are able to maintain local correlational structure while disrupting aspects of global appearance. The other half of this manipulation could be an important means of characterizing what information children are and are not using at different developmental stages, but how do we disrupt local structure while preserving aspects of global appearance? One possible means of doing so that we have explored in previous work with adults (Balas, 2012) and infants (Balas & Woods, 2014) is contrast negation. Contrast-negated images preserve the global layout of edges (and spatial layout more broadly) while disrupting the local polarity of contrast relationships. If young children are relying heavily on higher-order correlations for some material categorization judgments (as in Experiment 1) we would predict that this manipulation should have *less* of an impact on these categories than on others, even with the profound change in image appearance that accompanies negation. Were this the case, it would be important supporting evidence in favor of our conjecture regarding the way representations of appearance are changing with development.

Overall, our results demonstrate that the way children use summary statistics for material categorization changes between the ages of 5 and 10 years. Critically, these effects do not appear to reflect broad inabilities to measure summary statistics or distinguish between different sets of summary features. Instead, reliance on summary features vs. higher-order measurements varies by material. Understanding how the diagnosticity of summary statistics across material categories (or other properties of these categories) predicts these relationships is a key question for future

work and may yield insight into how rich representations for recognition develop more generally.

Keywords: visual development, material perception, summary statistics

Acknowledgments

This study was supported by NSF grant BCS-1727427 awarded to BB. Special thanks to all the families who volunteered to participate in both experiments, and also to Dan Gu for technical support.

Commercial relationships: none.
Corresponding author: Benjamin Balas.
Email: benjamin.balas@ndsu.edu.
Address: Department of Psychology and Center for Visual and Cognitive Neuroscience, North Dakota State University, Fargo, ND, USA.

References

Balas, B. (2006). Texture synthesis and perception: Using computational models to study texture representations in the human visual system. *Vision Research*, *46*, 299–309.

Balas, B. (2012). Contrast-negation and texture synthesis differentially disrupt texture discrimination. *Frontiers in Perception Science*, *3*, 515.

Balas, B. (2016). Seeing number using texture: How summary statistics account for reduced numerosity in the visual periphery. *Attention, Perception & Psychophysics*, *78*, 2313–2319.

Balas, B., Auen, A., Saville, A., & Schmidt, J. (2017). Body emotion recognition disproportionately depends on vertical orientations during childhood. *International Journal of Behavioral Development*, E-pub ahead of print, doi:10.1177/0165025417690267.

Balas, B., & Conlin, C. (2015a). The visual N1 is sensitive to deviations from natural texture appearance. *PLoS One*, *10*(9), e0136471.

Balas, B. & Conlin, C. (2015b) Invariant texture perception is harder with synthetic textures: Implications for models of texture processing. *Vision Research*, *115*, 271–279.

Balas, B., Conlin, C., & Shipman, D. (2016). Summary-statistics and material categorization in the visual periphery. *Transactions on Applied Perception*, *14*, 2, 8.

Balas, B., Huynh, C., Saville, A., & Schmidt, J. (2015). Orientation biases for facial emotion recognition in early childhood and adulthood. *Journal of Experimental Child Psychology*, *140*, 71–83.

Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision*, *9*(12):13, 1–18, doi:10.1167/9.12.13. [PubMed] [Article]

Balas, B., & Woods, R. (2014). Infant preference for natural texture statistics is modulated by contrast polarity, *Infancy*, *19*, 262–280.

Baumgartner, E. & Gegenfurtner, K.R. (2016). Image statistics and the representation of material properties in the visual cortex. *Frontiers in Psychology*, *7*, 1185, doi: 10.3389/fpsyg.2016.01185.

Brainard D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Briand, T., Vacher, J., Galerne, B., & Rabin, J. (2014). The Heeger-Bergen pyramid-based texture synthesis algorithm. *Image Processing Online*, *4*, 276–299.

Dakin, S., & Watt, R. J. (2009). Biological "bar codes" in human faces. *Journal of Vision*, *9*(4):2, 1–10, doi: 10.1167/9.4.2. [PubMed] [Article]

Efros, A. A., & Freeman, W. T. (2001). Image quilting for texture synthesis and transfer. *Proceedings of the 28th annual conference on computer graphics and interactive techniques* (pp. 341–346). New York: ACM.

Ellemberg, D., Hansen, C., & Johnson, A. (2012). The developing visual system is not optimally sensitive to the spatial statistics of natural images. *Vision Research*, *67*, 1–7.

Fleming, R. W. (2013). Visual perception of materials and their properties. *Vision Research*, *94*, 62–75.

Fleming, R. W., Wiebel, C. M., & Gegenfurtner, K. R. (2013). Perceptual qualities and material classes. *Journal of Vision, 13(9):8*, 1–12, doi:10.1167/13.9.8. [PubMed] [Article]

Freeman J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience*, *14*, 1195–1201.

Heeger, D. J., & Bergen, J. R. (1995). Pyramid-based texture analysis/synthesis. *Proceedings of the 22nd Annual Conference on Computer Graphics & Interactive Techniques*, *30*, 229–238.

Hiramatsu, C., Goda, N. & Komatsu, H. (2011). Transformation from image-based to perceptual representation of materials along the human ventral visual pathway. *NeuroImage*, *57*, 482–494.

Jacobs, R. H. A. H., Baumgartner, E., & Gegenfurtner, K. R. (2014). The representation of material categories in the brain. *Frontiers in Psychology*, *5*, 146.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception*, *36*, 1.

Kovacs, I., Kozma, P., Feher, A., & Benedek, G. (1999). Late maturation of visual spatial integration in humans, *Proceedings of the National Academy of Sciences*, *96*, 12204–12209.

Liang, Y., Simoncelli, E. P., & Lei, Z. (2000). Color channels decorrelation by ICA transformation in the wavelet domain for color texture analysis and synthesis. Presented at IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head, SC, USA.

Padilla, S., Drbohlav, O., Green, P. R., Spence, A., & Chantier, M. (2008). Perceived roughness of 1/f beta noise surfaces. *Vision Research*, *48*, 1791–1797.

Pascalis, O., de Vivies, X. D. M., Anzures, G., Quinn, P. C., Slater, A. M., Tanaka, J. W., & Lee, K. (2011). Development of face processing. *Wiley Interdisciplinary Reviews of Cognitive Science*, *2*, 666–675.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies, *Spatial Vision*, *10*, 437–442.

Portilla, J., & Simoncelli, E. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, *40*, 49–71.

Rosenholtz, R., Huang, J., Raj, A., Balas, B., & Ile, L. (2012). A summary-statistic representation in peripheral vision explains visual search. *Journal of Vision*, *12*(4):14, 1–17, doi:10.1167/12.4.14. [PubMed] [Article]

Ruiz-Soler, M., & Beltran, F. S. (2006). Face perception: An integrative review of the role of spatial frequencies. *Psychological Research*, *70*, 273–292.

Sharan, L., Li, Y., Motoyoshi, I., Nishida, S., & Adelson, E.H. (2008). Image statistics for surface reflectance perception. *Journal of the Optical Society of America, A*, *25*, 846–865.

Sharan, L., Rosenholtz, R., Adelson, E. H. (2009). Material perception: What can you see in a brief glance? *Journal of Vision*, *9*(8): 784, doi:10.1167/9.8.784. [Abstract]

Sharan, L., Rosenholtz, R. & Adelson, E. H. (2014). Accuracy and speed of material categorization in real-world images. *Journal of Vision*, *14*(9):12, 1–24, doi:10.1167/14.9.12. [PubMed] [Article]

Sireteanu, R., & Rieth, C. (1992). Texture segregation in infants and children. *Behavioural Brain Research*, *49*, 133–139.

Wallis, T. S. A., Bethge, M., & Wichmann, F. A. (2016). Testing models of peripheral encoding using metamerism in an oddity paradigm. *Journal of Vision*, *16*(2):4, 1–30, doi:10.1167/16.2.4. [PubMed] [Article]

Wiebel, C., Toscani, M., & Gegenfurtner, K. R. (2015). Statistical correlates of perceived gloss in natural images. *Vision Research*, *115B*, 175–187.

Wiebel, C., Valsecchi, M. & Gegenfurtner, K. R. (2013). The speed and accuracy of material recognition in natural image. *Attention, Perception, & Psychophysics*, *75*, 954–966.

Wiebel, C. B., Valsecchi, M., & Gegenfurtner, K. R. (2014). Early differential processing of material images: Evidence from ERP classification. *Journal of Vision*, *14*(7):10, 1–13, doi:10.1167/14.7.10. [PubMed] [Article]

Yang, J., Kanazawa, S., & Yamaguchi, M. K. (2013). Can infants tell the difference between gold and yellow? *PLoS One*, *8*(6), e67535.

Yang, J., Kanazawa, S., Yamaguchi, M. K., & Motoyoshi, I. (2015). Pre-constancy vision in infants. *Current Biology*, *25*, 3209–3212.

Yang, J., Otsuka, Y., Kanazawa, S., Yamaguchi, M. K., & Motoyoshi, I. (2011). Perception of surface glossiness by infants aged 5 to 8 months. *Perception*, *40*, 1491–1502.