bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Material category determined by specular reflection structure mediates the processing of image features for perceived gloss

Authors:

Alexandra C. Schmid[1], Pascal Barla[2], Katja Doerschner[1,3]

[1]Justus Liebig University Giessen, Germany

[2]INRIA, University of Bordeaux, France

[3]Bilkent University, Turkey

Correspondence:
Alexandra.Schmid@psychol.uni-giessen.de (A.C.S.);
pascal.barla@inria.fr (P.B.);
Katja.Doerschner@psychol.uni-giessen.de (K.D.)

31.12.2019

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ABSTRACT

There is a growing body of work investigating the visual perception of material properties like gloss, yet practically nothing is known about how the brain recognises different material classes like plastic, pearl, satin, and steel, nor the precise relationship between material properties like gloss and perceived material class. We report a series of experiments that show that parametrically changing reflectance parameters leads to qualitative changes in material appearance beyond those expected by the reflectance function used. We measure visual (image) features that predict these changes in appearance, and causally manipulate these features to confirm their role in perceptual categorisation. Furthermore, our results suggest that the same visual features underlie both material recognition and surface gloss perception. However, the predictiveness of each feature to perceived gloss changes with material category, suggesting that the pockets of feature space occupied by different material classes affect the processing of those very features when estimating surface glossiness. Our results do not support a traditional feedforward view that assumes that material perception proceeds from low-level image measurements, to mid-level estimates of surface properties, to high-level material classes, nor the idea that material properties like gloss and material class are simultaneously “read out” from visual gloss features. Instead, we suggest that the perception and neural processing of material properties like surface gloss should be considered in the context of material recognition.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

INTRODUCTION

Our visual world is made up of light that has been reflected, refracted, or transmitted by surfaces. From this light we can tell whether a surface is light or dark, shiny or dull, translucent or opaque, made from platinum, plastic, or pearl (Figure 1A). This ability of the human visual system is not trivial as the structure, spectral content, and amount of light reaching our eyes depends on surface reflectance and transmittance properties, 3D surface geometry, the position and intensity of light sources, and observer viewpoint. An important but unresolved challenge in visual neuroscience is determining how the brain disentangles these conflated contributing factors to the retinal image to recognise materials and their intrinsic properties.

There is a growing body of work investigating the visual perception of material properties like colour, lightness, transparency, translucency, and gloss (for reviews see Anderson, 2011; Foster, 2011; Chadwick & Kentridge, 2015; Fleming 2014, 2017), yet practically nothing is known about how the brain recognises different material *classes* like plastic, pearl, satin, and steel. A likely reason for this is that studying properties like colour and gloss seems more tractable than discovering the necessary and sufficient conditions for recognising the many material classes in our environment. For example, previous research has discovered a limited set of conditions that trigger the perception of a glossy versus matte surface, involving the intensity, shape, position, and orientation of specular highlights (bright reflections; Beck & Prazdny, 1981; Blake & Bülthoff, 1990; Wendt, Faul, & Mausfeld, 2008; Todd, Norman, & Mingola, 2004; Anderson & Kim, 2009; Kim, Marlow, & Anderson, 2011; Marlow, Kim, & Anderson, 2011), and lowlights (dark reflections; Kim, Marlow, & Anderson, 2012) with respect to diffuse shading. In contrast, there is nearly no work trying to ascertain the conditions for perceiving different material classes (Todd & Norman, 2018, 2019; Tamura, Higashi, & Nakauchi, 2018). The challenge is that the perceptual space of materials is unspecified; there are many different optical "appearances" that can look like steel (think polished, scratched, rusted), or plastic (smooth and glossy or rough and dull), which renders this a seemingly intractable problem. Furthermore, if a traditional feedforward view is assumed (Komatsu & Goda, 2018) in which "material perception proceeds from low-level image measurements, to mid-level estimates of surface properties, to high-level material classes" (Fleming, 2017), then it would make sense to first study how the brain computes mid-level properties like gloss from images. In this framework, material class is thought to be subsequently computed from the high-dimensional feature space defined by these component mid-level properties (Fleming, 2017).

An alternative view suggests that the visual system does not try to estimate physical surface properties like diffuse reflectance (seen as body colour/pigment) and specular reflectance (seen as

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

gloss) per se., but rather re-defines mid-level features as statistical variations in image structure from which both material properties (like gloss) and material classes (like plastic) can be "read out" simultaneously (Fleming, 2014, 2017; Fleming & Storrs, 2019). For example, different surface reflectance properties affect the amount and way in which incident light is scattered by a surface, often causing differences in the "appearance" or structure of specular highlights (i.e. their size, contrast, sharpness, and colour; Figure 1B). Such specular highlight features have been shown to predict the perception and, importantly, misperception of surface gloss (Marlow, Kim, & Anderson, 2012). Figure 1B shows that the appearance of specular reflections can also cause each surface to resemble different materials like "shiny chrome", "dull plastic", and "pearlescent coating". That is, the structure of specular reflections seems to simultaneously affect the (quantitative) perceived gloss level (shiny or dull) and the *quality* – or material category – of the surface, and it is possible that the same features underlie both processes. Yet, it is currently unknown the extent to which the appearance of specular reflections is diagnostic of material category and, furthermore, what the precise relationship between material properties like gloss and perceived material class is (Schmid & Doerschner, 2019).

We think a third possibility exists that describes the relation between material categories and mid-level image features for gloss: namely that the brain's sensitivity to specular reflections has developed primarily for the successful identification of and discrimination between material classes, and that our ability to report a surface's "level of gloss" is a perceptual outcome of this.

Here, we report a series of experiments that support this last view. We show that parametrically changing reflectance parameters leads to qualitative changes in material appearance beyond those expected by the reflectance function used. We measure visual (image) features that predict these changes in appearance, and causally manipulate these features to confirm their role in perceptual categorisation. Furthermore, our results suggest that the same visual features underlie both material recognition and surface gloss perception. However, the predictiveness of each feature to perceived gloss changes with material category, suggesting that the pockets of feature space occupied by different material classes affect the processing of those very features when estimating surface glossiness. Taken together, our results suggest that perceived gloss should not be treated as a unitary concept (c.f. Motoyoshi et al., 2007), that the dimensionality of gloss perception is much larger than previously reported (c.f. Pellacini, Fererda, & Greenberg, 2000; Vangorp, Barla, & Fleming, 2017), and that the perception and neural processing of material properties like surface gloss should in fact be considered in the context of material recognition (c.f. Nishio, Goda, & Komatsu, 2012). Our results also have implications for the object recognition literature (Grill-Spector & Weiner, 2014; Bracci, Ritchie, & Op de Beeck, 2017; Long, Yu, & Konkle, 2018) by suggesting that

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the specific photogeometric constraints on image structure that trigger our perception of surface gloss play an important role in visual categorisation.

Figure 1. The appearance of specular reflections is determined by the way surfaces scatter light. **A.** Most objects that we see every day are made from materials that produce characteristic specular reflections. **B.** The same sphere is rendered with different surface reflectance properties, demonstrating how the appearance of specular reflections can determine both perceived glossiness and material class. Properties of the diffuse shading and specular component affect various aspects of image structure (e.g. the size, contrast, sharpness, and colour of specular highlights), and this image structure causes each surface to resemble different materials like “shiny chrome”, “dull plastic”, and “pearlescent coating”.

RESULTS

If the visual system is sensitive to specular reflections for material recognition beyond whether a surface is shiny or matte (Marlow et al., 2011), then materials should be recognisable from their specular reflection structure and we should be able to identify features of this structure that predict and alter material category. To test this, we developed a large stimulus set of computer-rendered complex glossy objects (Figure 2) and generated material category terms for each stimulus (Figures 3 and 4) through a free-naming and multiple-alternate-forced-choice task (Experiments 1 and 2, respectively). Dimension reduction via a factor analysis revealed thirteen “emergent categories” for our stimulus set (Figure 5). We used linear discriminant analysis (LDA) to predict these emergent categories from a set of measured image features (also referred to as “visual

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

features" or "features") that describe the appearance of specular reflections (Figures 6 and 7A-B). We additionally predicted differences in (original) category membership between the stimuli from differences in visual features using a representational similarity analysis (RSA; Figure 8). We then causally manipulated image features to transform perceived category in predicted ways (Experiment 3, Figure 9).

A new set of participants rated perceived gloss of the stimuli (Experiment 4) to test whether gloss (as a unitary concept) is used to determine material category (as in the *feedforward approach*), whether the same image features are used to simultaneously "read out" both material category and surface gloss level (as in the *simultaneous approach*), or whether perceived gloss is mediated by material recognition. We examined the distribution of gloss ratings within each emergent material category and found that categories were not associated with a particular gloss level, nor did emergent category membership of the stimuli correlate with their perceived glossiness (Figure 7A). Furthermore, RSA showed that differences in (original) category membership between stimuli were not well predicted by differences in gloss level even when colour was taken into account (Figure 8C), providing evidence against the *feedforward approach*. Another RSA revealed that differences in gloss level between the stimuli were predicted by differences in the measured visual features (Figure 8C); however, the weighting of the features to gloss level differed across category (Figure 7C and D), providing evidence against the *simultaneous approach* and instead suggesting that processing of image features for perceived gloss is mediated by material recognition. We report these results in detail below.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 2. Computer-rendered stimuli in the experiments were complex 3D shapes embedded in natural illumination fields. **A.** The visual effects of manipulating different rendering parameters. **B.** A sample of the stimuli used in the multi-AFC and gloss rating experiments. Only reflectance parameters were manipulated and yet we found that these variations yielded many different perceived material classes. Here the reader can get a sense of the range of material appearances that participants saw. Stimuli marked with black spots are defined by very rough, anisotropic specular reflections, which gives the effect of elongated, low-clarity specular reflections that, despite this low clarity, have rapidly changing (sharp/high contrast) boundaries between

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

highlights and lowlights relative to isotropic surfaces. These stimuli are often classified as fabrics (also see Figure 5C).

Specular reflection appearance yields a diverse range of material classes

We developed a stimulus set to test whether glossy materials are recognisable from their specular reflection structure. Figure 2A shows the same dragon model rendered in the same natural illumination field with different surface reflectance properties. We manipulated base colour (lightness and saturation of the diffuse component) in addition to five specular reflection parameters (specular strength, specular tint, specular roughness, anisotropy, and anisotropic rotation) to control the appearance of specular reflections with respect to diffuse shading, resulting in 270 stimuli. In a free-naming task (Experiment 1), participants viewed the stimuli one at a time and judged what material each object was made from with no restrictions. This allowed us to create unbiased participant-generated categories for the stimulus set. After processing for duplicates and similar terminology, 209 terms were used to describe the materials, 121 of which were *category terms* (i.e. nouns like porcelain, gold, plastic). There were 17 *non-category terms* that included descriptions of fillings (e.g. filled with sand), coatings (e.g. glazed, varnished, painted, covered), finishes (e.g. polished, sanded), and states (e.g. liquid, wet, melting). Participants also used a total of 71 *other adjectives* to describe the stimuli (e.g. dark, yellow, glossy, matte). These other adjectives were excluded from analyses altogether but can be found in Supplementary Figure 1. Figure 3A and B plot the number of times each term was used by each participant, for the 64 category and 10 non-category terms that were used by more than one participant. Importantly, the use of each term was distributed quite well among participants, i.e. category labels did not come from the same few participants. Figure 3C supports this by showing that each subject used many category terms (range = 18-45, median = 24).

These results demonstrate that changing the appearance of specular reflections yields an extremely diverse range of perceived materials. In the real world, materials like porcelain, pearl, soap, wax, velvet, metal, and many others produce image structure that is caused by the extent and way in which they disperse, refract, and transmit light in addition to pigment variations and mesoscale details like fibres in fabrics, or scratches in anisotropic metals. All these sources of image structure could potentially provide the visual system with information about material category. However, participants reported seeing these materials when surfaces were only defined by (uniform) diffuse and specular reflectance properties. Thus, the human visual system appears to be

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

highly sensitive to the image structure produced by specular reflections for material classification, even for complex materials.

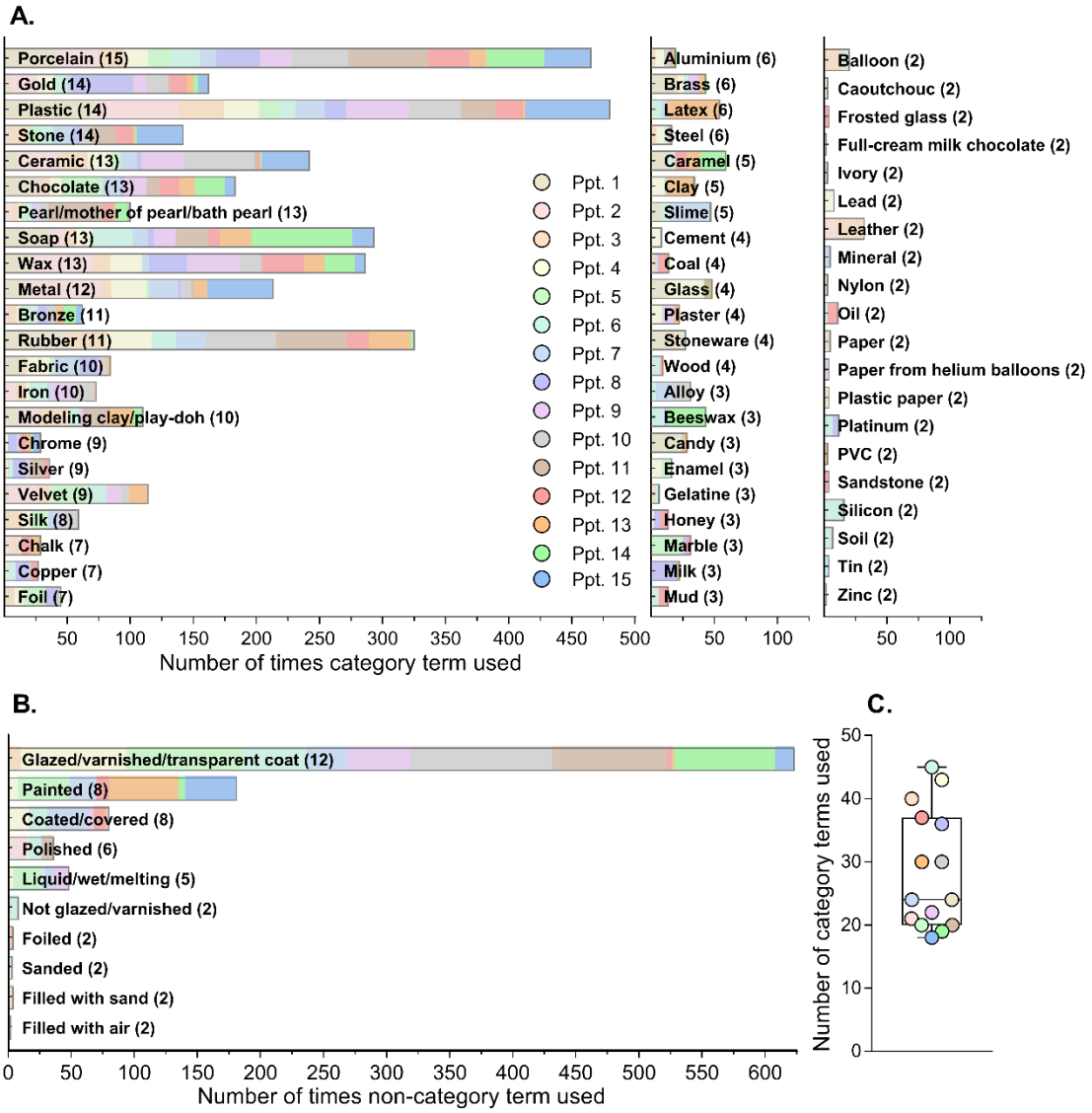

Figure 3. Results of the free-naming experiment. **A.** Number of times each category term was used. **B.** Number of times each non-category term was used (excluding *other adjectives*; see main text). In (A) and (B) the bars correspond to the total number of times a term was used in the experiment. The numbers in brackets correspond to the number of participants (out of 15) that used that term. Only terms that were used by at least two participants are shown. Not shown: the 57 category terms and 7 non-category terms that were not used by more than one participant, and the 71 *other adjectives*. **C.** Number of category terms used in the experiment by each participant.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The dimensionality of specular reflection appearance is large

Given the ability of specular reflections to transform perceived material, we aimed to identify which aspects of specular reflection image structure underlie material appearance. For this purpose, semantic labels are only relevant to the extent that they capture qualitative perceptual differences in visual material appearance. For example, dark brown stimuli with medium-clarity specular reflections might be labelled as “melted chocolate” or “mud” but would be qualitatively visually equivalent to one another (they would be hard to discriminate without the help of cues from other senses like smell, taste, or touch). Such stimuli would qualitatively differ (visually) from light yellow stimuli with low-clarity, dim reflections, which might be labelled as “wax” or “soap”. Therefore, before we could identify the image features that determine material category, we first sought to reveal the underlying perceptual space of visual material classes for our stimulus set, with the following steps.

First, we collected the 29 category terms that at least one third of participants used in the free-naming task (i.e. at least 5 out of 15; see Figure 3A). Some terms that were visually or semantically similar were merged to create a reduced set of category labels, guided by correlations between the categories (Supplementary Figure 2). We separated categories where particular coatings (glazed), finishes (polished), states (liquid/melting) applied. The reduced set of 18 category terms is shown in Figure 4.

Second, a separate set of participants completed a multiple-alternate-forced-choice task (multi-AFC task; Experiment 2) where they were asked to choose the material category that best applied to each stimulus (Figure 4A). For this experiment we extended our stimulus set to 924 stimuli that included a larger range of reflectance properties, shapes, and lighting environments (Figure 2B). Participants provided confidence ratings (converted to a score between 1-3), which allowed them to indicate their satisfaction with the category options presented. The confidence ratings for each category for each stimulus were summed across all participants, providing a distribution of category responses for each stimulus (*category profiles*) and a distribution of stimuli for each category (*stimulus profiles*; Figure 4B). For many stimuli, more than one category term applied; for example, stimulus 1 was almost equally classified as “glazed ceramic” and “covered in wet paint”. Participants potentially chose different categories for the same stimulus due to redundancies in terminology and/or having different decision boundaries due to imperfect category membership (e.g. “looks a bit like plastic but also a bit like rubber”). Indeed, we found that the stimulus profiles for some of the categories were correlated with one another (Supplementary Figure 3).

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A.

B.

	Stimulus 1	Stimulus 2	Stimulus 3	Stimulus 4	Stimulus 5	Stimulus 6	Stimulus 7	Stimulus 8	Stimulus 9	Stimulus 10	Stimulus 11	Stimulus 12	Stimulus 13	Stimulus 14	Stimulus 15	...
Vevlet / silk / fabric												3		3		
Glazed porcelain	3			3												
Glazed ceramic	21			17				18			3		3		15	
Unglazed porcelain / chalk / stone			9		3	3	9							3		
Modelling clay / play-doh		9	15		15		5			3						
Latex / rubber	3	9	20	21	24	12	21		3		3	9	3	9		
Wax / soap			3	6	5	9	9		3	3	3	3	3			
Pearl / mother-of-pearl / bath pearl																
Plastic	2	20	6	8	3	27	6		6	4	6					
Brass / bronze / copper								3	3	6			3			
Gold / golden metal																
Unpolished steel / iron / chrome		6			3											
Polished steel / iron / chrome							3									
Silver / aluminium																
Foil (incl. coloured)					3					3						
Chocolate (solid, incl. white)		5		3					36	30	27	44	41	43		
Covered in melted chocolate / caramel		3						36	9	6	18		3		24	
Covered in wet paint / slime	30	6	6		3	6	6	3		3			3		21	

Category profile

Stimulus profile

Figure 4. Multi-AFC experiment (Experiment 2) with the reduced set of 18 category terms. **A.** A screenshot of an example trial. Observers were presented one stimulus at a time and were asked to choose the category that best applied to each stimulus. If they were unhappy with their choice, they could change the confidence rating at the bottom right of the screen. At the bottom centre of the screen was a trial counter indicating the current trial and total number of trials in that block. Note that the categories were presented in German in the actual experiment (see Supplemental methods for translations). **B.** Sum of confidence ratings from the multi-AFC experiment for each stimulus and each category for the first 15 stimuli (out of 924). *Category profiles* are the

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

distribution of category responses (confidence interval sums) for each stimulus. *Stimulus profiles* are the distribution of stimuli (confidence interval sums) for each category. Stimulus profiles were used for the factor analysis (Figure 5), and category profiles were used for the representational similarity analysis (Figure 8).

Third, the stimulus profiles were subjected to a factor analysis to find the latent perceptual category space for our stimulus set. We found that 12 factors were needed to explain at least 80% of the shared variance between stimuli (Figure 5A). Retaining 10 or 8 factors would explain approximately only 70% and 60% of the common variance in category profiles, respectively. Figure 5C shows 13 emergent categories (with example stimuli) from the reduced category space after factor analysis (twelve plus one dimension that emerged from the negative loadings). These emergent categories were interpreted based on the categories that loaded most strongly onto each factor (see Figure 5B). Note that the emergent category labels are arbitrary and are only included for convenience. Stimuli were assigned to one of the 13 emergent categories based on the factor onto which it most strongly loaded.

These results demonstrate that the perceptual dimensionality of gloss objects (defined by a diffuse and specular component) is larger than the small number of “gloss” dimensions that have previously been explored (e.g. Pellacini et al., 2000; Vangorp et al., 2017; Hunter, 1937). Note that we did not sample the whole perceptual space of glossy objects defined by a diffuse and specular component, and this dimensionality is likely to expand with larger sampling (e.g. we only sampled stimuli with one base colour – yellow – with different levels of colour-saturation).

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

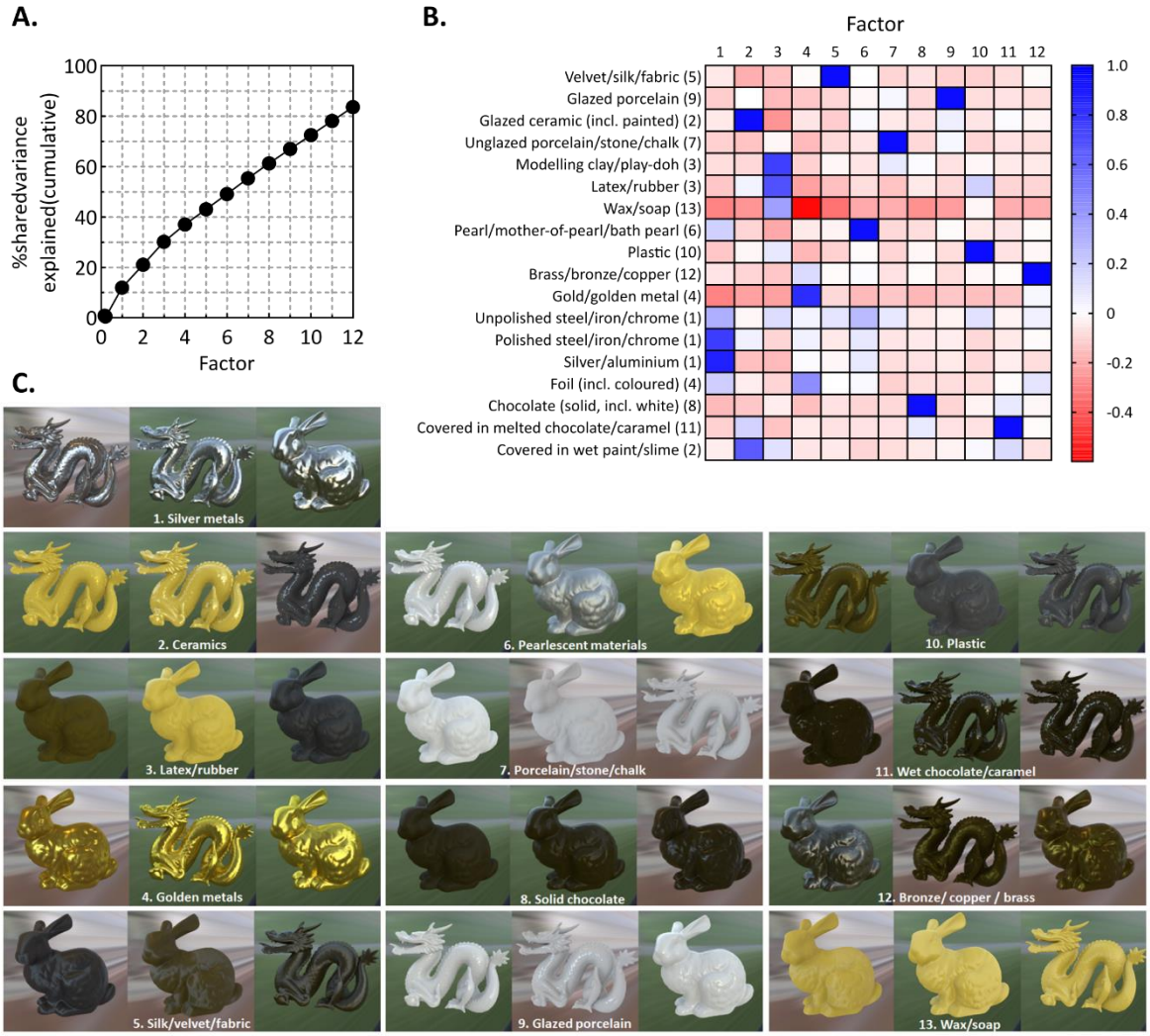

Figure 5. The results of the factor analysis performed on the multi-AFC data (Experiment 2). **A.** Cumulative shared variance explained by each factor with the 12-factor solution. 12 factors were needed to explain at least 80% of the shared variance between stimulus profiles. **B.** Heat plot of the factor loadings for each category, with blue and red cells showing positive and negative loadings for each category, respectively. The numbers in brackets correspond to the emergent category that most related to each of the original category terms. **C.** The 13 emergent categories from the reduced category space after factor analysis, with example stimuli from each emergent category. Note that the 13th category results from the negative loadings onto Factor 4 in (B).

Specular reflection image features predict perceived material category

If the structure of specular reflections is important for determining material class, then we should be able to identify specular reflection characteristics that predict the emergent categories that were revealed by the factor analysis in the previous section. To this end, we developed

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

measures of six *visual gloss features* – *coverage*, *contrast*, *sharpness*, *highlight saturation*, *lowlight saturation*, and *lowlight value* – that describe the appearance of specular reflections. These features are based on generative constraints on how light interacts with surfaces that have a diffuse and glossy component (Marlow & Anderson, 2013). Three of these features – *coverage*, *contrast*, and *sharpness* – have previously been found to predict observers' judgments of surface glossiness (Marlow et al., 2012; Marlow & Anderson, 2013). Since we additionally varied parameters to do with the colour of the diffuse and specular components, we included three colour-related features – *highlight saturation*, *lowlight saturation*, and *lowlight value* – that differ between our stimuli and could potentially be used in material categorisation. Details of our measures can be found in the Analyses section and Figure 6.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

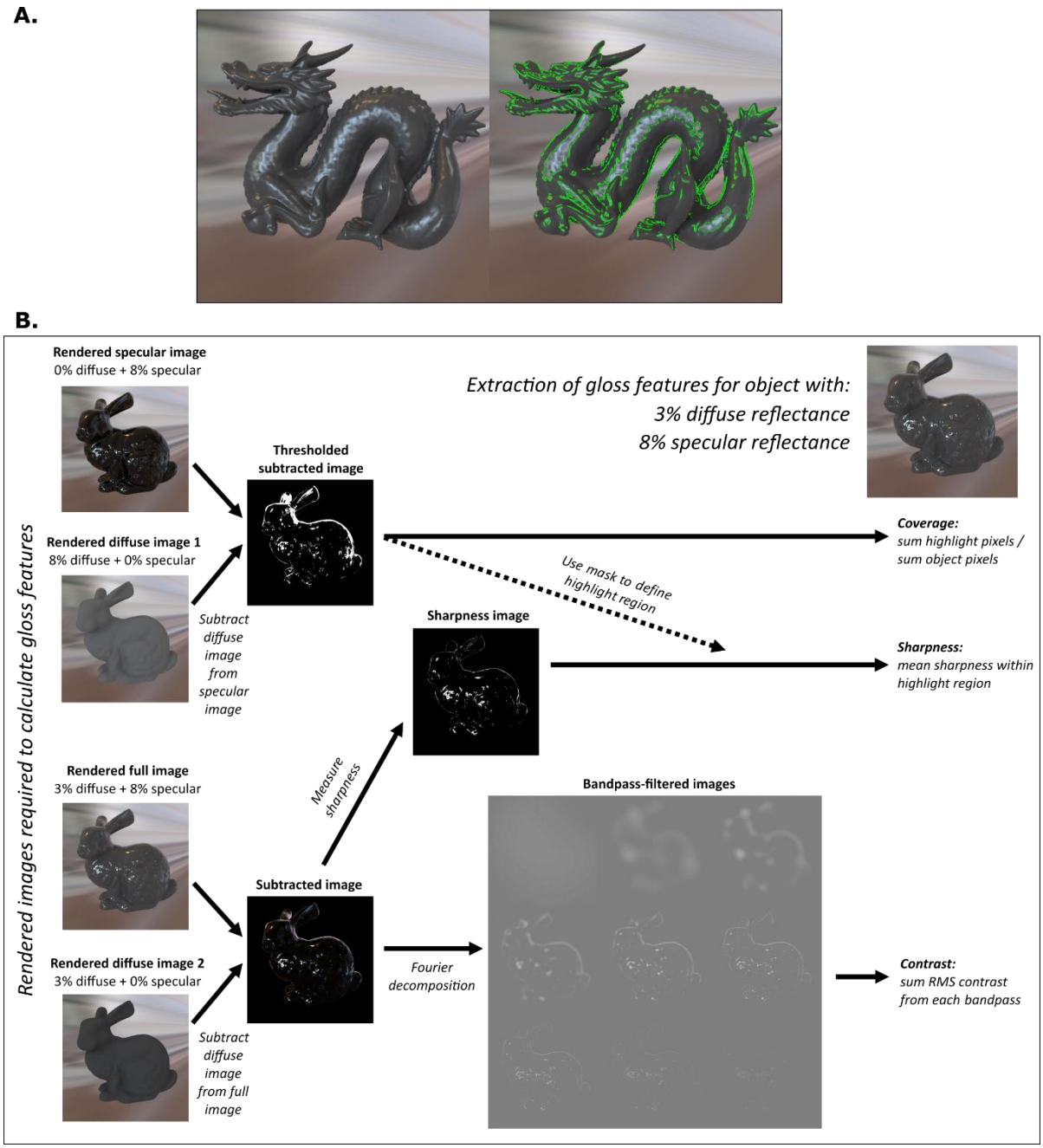

Figure 6. Calculation of visual gloss features. To help with segmenting the image structure caused by specular reflections and diffuse shading, we rendered additional images that isolated specular and diffuse components. **A.** We extracted both specular highlights ("bright" reflections) and specular reflections, which include both specular highlights and lowlights (i.e. both bright and dim reflections; Kim et al., 2012). Left: original rendered image. Right: results of specular highlight segmentation. The green outline shows the boundary between highlights and lowlights. **B.** Image-based gloss feature measurements. *Coverage* was defined as the proportion of object pixels that were calculated to be specular highlights (excluding lowlights), based on a threshold above the maximum diffuse shading; *contrast* was the sum of root-mean-squared (RMS) contrast of extracted specular reflections at different spatial frequency bandpasses; *sharpness* of extracted specular reflections was calculated for each pixel within the highlight regions using a measure of local phase coherence (Hassen, Wang,

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

& Salama, 2013), then these values were averaged; *highlight saturation* and *lowlight saturation* were calculated as the average colour saturation of pixels within the highlight region, and outside of the highlight region (which we call the lowlight region), respectively; *lowlight value* was calculated as the average value (or lightness) of pixels in the lowlight region.

Figure 7A plots the distribution of measured visual features (coloured violin plots) for each emergent category. To provide a fair comparison between emergent categories, these plots include the 36 stimuli that loaded most strongly onto each emergent category. This is because 36 was the least number of stimuli within a category (glazed porcelain). In subsequent analyses we will distinguish "top 36 stimuli" from "all stimuli" within an emergent category. Visualised in this way (Figure 7A), the image features provide "material signatures" for each emergent category. Data from the top 36 stimuli were subjected to a linear discriminant analysis (LDA) that predicted the emergent material category of each stimulus from the measured visual features. Figure 7B (left) plots the accuracy of the model for each emergent category. Overall accuracy was 68%, which is well above chance (7.7%, red dotted line) and demonstrates that visual features can predict human material categorisation for our stimulus set. A cross-validation test showed that the model generalises across shape and lighting conditions (Figure 7B, right); that is, if the classifier is trained on one shape and tested on the other, or trained on one lighting environment and tested on the other, the model's performance is relatively stable.

Interestingly, some categories were classified better than others. The emergent categories with the highest classification accuracies (metals, chocolate, porcelain, and wax/soap) are those with the most distinct material signatures (i.e. the most distinct set of visual features). That is, there are at least a few features that seem to "characterise" that material (e.g. silvers must have uncoloured highlights and lowlights; wax must be light and coloured with low-contrast reflections). The emergent categories with the lowest classification accuracies (glazed ceramic, latex, pearlescent materials, plastic, and fabrics) are less specifically defined by the features. One possible reason for this is that there are many types of plastics, for example, and different specific (nonlinear) combinations of features define these subtypes – something that LDA does not capture. A second possibility is that the emergent categories are not truly perceptually orthogonal, and the stimuli would be better represented as a smooth continuation from one category dimension to another (that is, stimulus categories still overlap after dimension reduction). To allow for this, below we test whether differences in category profile (Figure 8A) can be predicted by differences in the measured visual features (Figure 8B) with a representational similarity analysis (RSA; Kriegeskorte, Mur, & Bandettini, 2008). This also allows us to include all stimuli in the analysis rather than just the top 36

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in each emergent category. Finally, a third possibility is that our features do not fully capture the differences in appearance between the emergent categories. We test this possibility in the next section by attempting to transform one category into another through causal manipulation of the visual features.

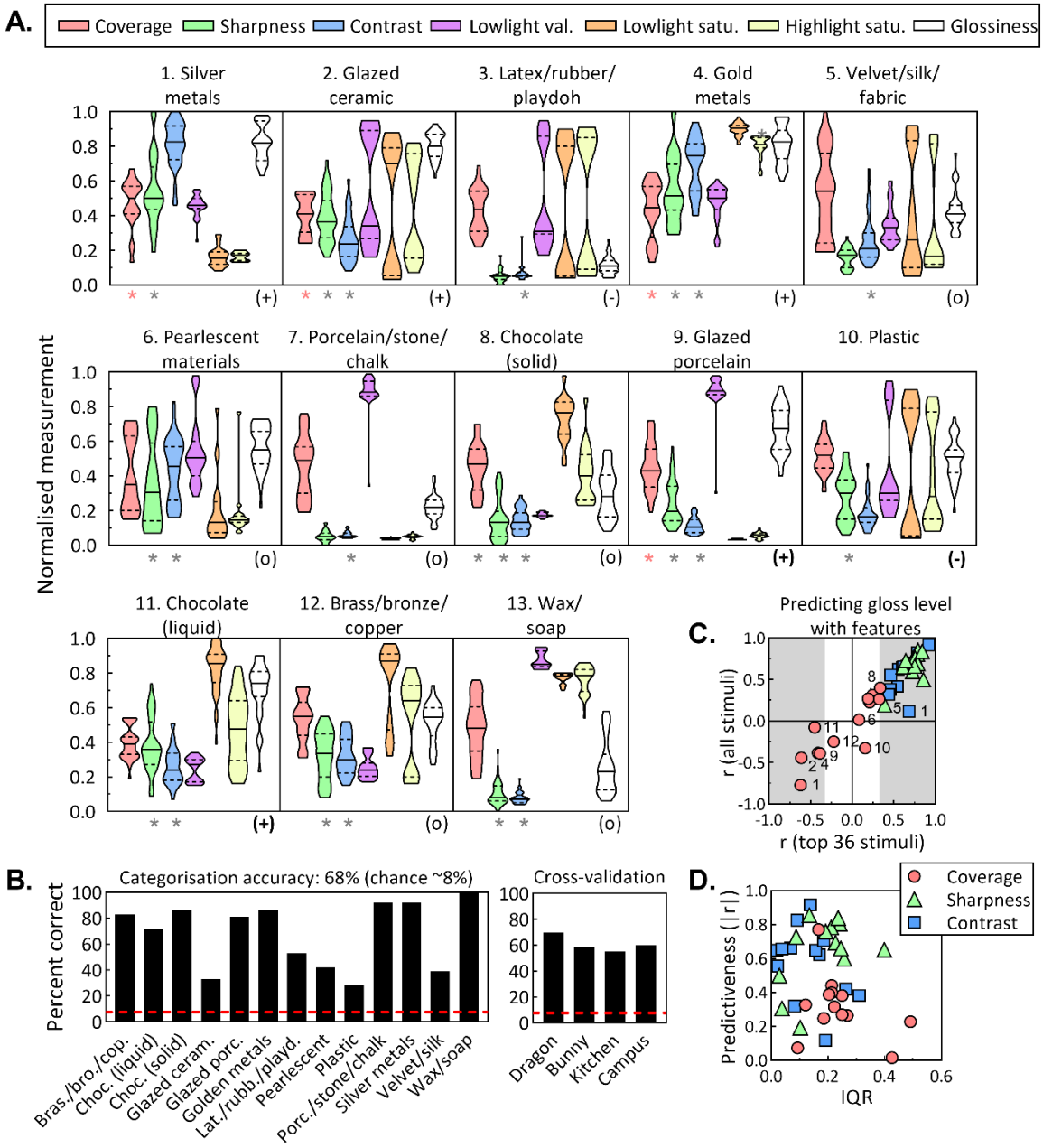

Figure 7. The relationship between measured visual features, perceived emergent category (Experiment 2) and perceived gloss level (Experiment 4). **A.** The violin plots show the distribution of measured image features (coloured plots) and distribution of gloss level (white plots) for each emergent category. The symbols in parentheses below these plots show whether stimulus factor scores correlate with perceived gloss level

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

positively (+), negatively (-) or not at all (o), for the top 36 stimuli (bolded symbols), or all stimuli (unbolded symbols) in each emergent category. **B.** Classifier accuracy for each emergent category from the linear discriminant analysis. The red dotted line indicates chance level. **C.** Pearson correlations (r) between visual features (coverage, sharpness, and contrast) and perceived gloss level, for each emergent category (some of which are labelled with a number that corresponds to the plots in [A]). Correlations between gloss ratings and visual features are shown for the top 36 stimuli (x-axis) and for all stimuli (y-axis) in an emergent category. The grey zone shows significant correlations, which are also indicated with a star below the plots in (A) (negative correlations are shown in red). **D.** There is no relationship between predictiveness (|r|) of each feature to perceived gloss, and variability of each feature within an emergent category, defined by the interquartile range (IQR).

For the RSA, dissimilarity scores were calculated between all stimuli in terms of category dissimilarity and visual feature dissimilarity. Category dissimilarity scores were calculated as one minus the spearman correlation coefficient for each pair of category profiles (Figure 8A). Feature dissimilarity scores were calculated as the absolute difference between each measured gloss feature. Representational dissimilarity matrices (RDMs) are shown in Figure 8A and B, where each point shows how dissimilar each stimulus is to each other stimulus in terms of their category profile (Figure 8A), and each measured gloss feature (Figure 8B), with yellow pairs being more dissimilar. Dissimilarity scores in the upper triangles of each RDM were vectorized and subjected to a multiple linear regression, with visual feature dissimilarities as the predictors and category profile dissimilarity as the outcome. The model significantly predicts category profile differences, $R^2 = 0.32$, $F(6,426419) = 32797$, $p < 0.001$, demonstrating that changes in the category profiles (Figure 8A) are reflected by systematic changes in the measured gloss features. This is slightly better than a model that predicts category profile differences with the physical rendering parameters that were used to create the stimuli, $R^2 = 0.27$, $F(8,426417) = 20010$, $p < 0.001$. This suggests that our measured visual features capture the variance in category profiles caused by changes in the physical rendering parameters, and the slight improvement in predictive value for visual features is likely because they additionally take into account differences in image structure caused by shape and lighting differences.

To see the contribution of each visual feature to perceived category differences, Figure 8C (left) shows normalised beta weights (*beta*) and squared structure coefficients (rs^2) for each predictor. Due to multi-collinearity among the predictors (Supplementary Figure 4) and the potential for suppressor variables to influence the beta weights (Ziglari, 2017), it is useful to look at the squared structure coefficients when interpreting the predictive utility of each gloss feature.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Structure coefficients, calculated as the zero-order correlation (r) of each predictor divided by the multiple correlation (R), provide information about the contribution of each gloss feature to the total explainable variance. Contrast contributes most to explaining differences in perceived material category ($rs^2 = 0.78$), followed by sharpness ($rs^2 = 0.31$) and lowlight saturation ($rs^2 = 0.26$), which contribute about equally. There is a lower contribution of highlight saturation ($rs^2 = 0.12$) and lowlight value ($rs^2 = 0.05$), with coverage contributing least to the explainable variance ($rs^2 = 0.02$). Note that despite its low contribution in our stimulus set, we do not rule out coverage as an important factor in material categorisation generally. Since coverage did not vary much for our stimuli (range = 7% to 54%), to better determine its contribution we would have to test stimuli with a wider range of coverage values via different shape and/or light field manipulations (Marlow et al., 2012; Marlow & Anderson, 2013).

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

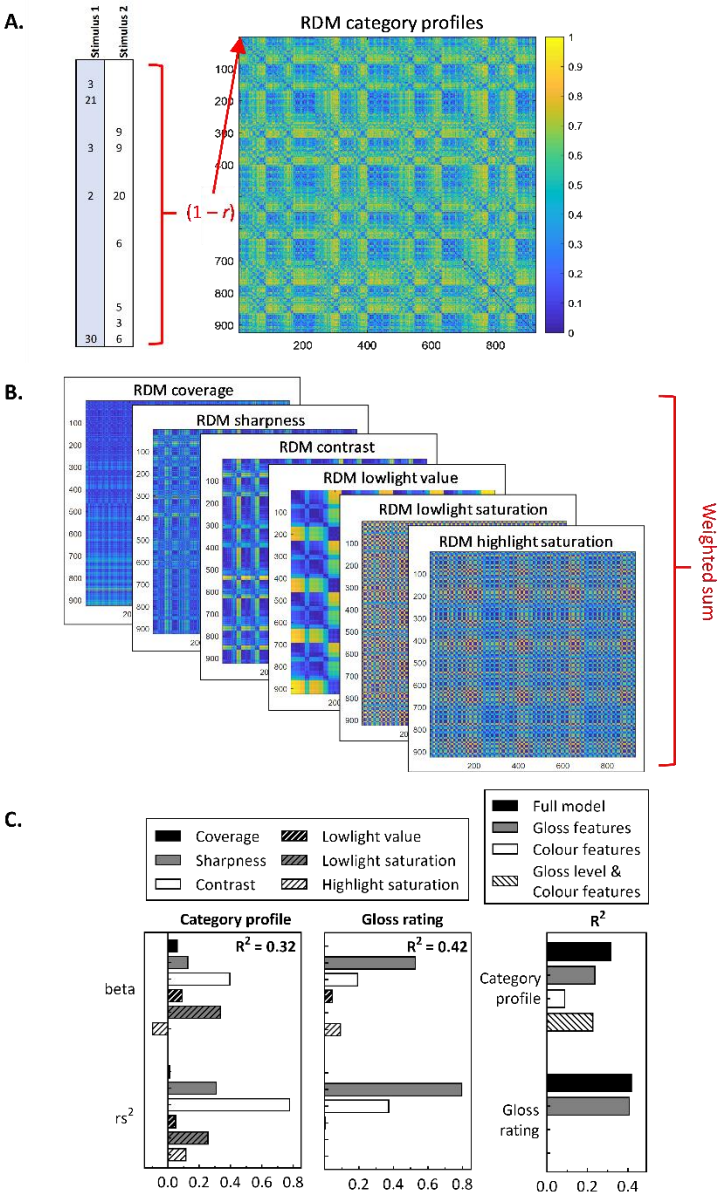

Figure 8. Results of the representational similarity analysis (RSA) to predict differences in category profile (Experiment 2) or differences in gloss ratings (Experiment 4) from visual features. **A** Representational dissimilarity matrix (RDM) for category profiles. Each point in the matrix shows the dissimilarity between two stimuli, calculated as one minus the Spearman correlation (*r*) of their category profiles. **B.** RDMs for each visual feature, which are linearly combined (weighted sum) in a multiple linear regression to predict the category profile RDM in (B) in Experiment 2 (multi-AFC task), and gloss-level RDM in Experiment 4 (gloss rating task). **C.** Normalised beta weights and squared structure coefficients (rs^2) from the linear regressions in Experiment 2 (multi-AFC task; left) and Experiment 4 (gloss rating task; middle), in addition to total explained variance (R^2) of each regression (right). All RDMs are normalised between zero and one.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Manipulating specular reflection image features transforms perceived material category

The LDA and linear regression with RSA in the previous section showed that linear combinations of specular reflection features do not fully capture participant's categorisation behaviour (68% classification accuracy with the LDA; 32% explained variance with the RSA). One possibility for this is that we missed important visual features that differed between the stimuli that are important for categorisation. Another possibility is that non-linear combinations of visual features trigger material category percepts. We test these possibilities by attempting to transform one category into another through causal manipulation of the six visual features. This is also a better test of whether the measured image features are causally responsible for the perceived category shifts, or whether they merely correlate with other changes in image structure that are important for categorisation but that we did not measure.

We attempted to transform the "glazed ceramic" emergent category into each of the remaining emergent categories, for each shape and light field (see Supplementary Figure 5 for examples). Figure 9A illustrates these manipulations with example stimuli. Our approach consists of a sequence of filters that rely on additional rendering outputs besides the original rendered stimuli: 1. A diffuse component image is used to obtain a specular component by subtraction (from the original stimulus); the filters are then applied independently on each component (specular & diffuse), which are summed in the end to yield the final manipulated image. 2. An image of surface normal orientation is used to control the blurring of the specular component in a shape-dependent way. Our filters affect the intensity and colour of specular or diffuse components (see Methods for further details). We created two sets of stimuli to test the success of our feature manipulations: a first set that contained "pure" feature manipulations (pure manipulations condition) using only linear multiplications of intensity values. We call these pure manipulations because they closely corresponded to the measures of visual features (which we verified; see Supplementary Figure 6). We empirically noticed that some transformations (e.g. those from glazed ceramic into silver metals or into fabrics) did not have the intended appearance after feature manipulation (Supplementary Figure 5). We hypothesised that for these stimuli particular types of contrast are important that cannot be achieved through a simple linear multiplication of intensity values. For metals, high contrast that covers the entire surface may be important (including "lowlight" regions). Fabrics (in our stimulus set) seem to be exclusively defined by very rough, anisotropic specular reflections, which cause a high degree of specular image blur in one direction. This gives the effect of elongated, low-clarity specular reflections that, despite this low clarity, have rapidly changing (sharp/high contrast) boundaries between highlights and lowlights relative to an isotropic surface (see stimuli marked with black spots in Figure 2B and the fabric category in Figure 5C). This is a different kind of

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

specular sharpness than seen in say ceramics, where there is a lot of structure *within* the highlights (i.e. the whole highlight is sharp, not just the contour, such that the reflected environment is discernible). We therefore included some additional manipulations to create a second set of stimuli (extra manipulations condition) involving non-linear manipulations of pixel intensities, including gamma corrections and, specifically for fabrics, a non-monotonic remapping of specular highlights (see Supplementary Figures 7 and 8 for details). The reasoning for the gamma correction was to boost the contrast of the specular lowlights for metals, and the non-monotonic remapping of highlights for fabrics created two aspects of image structure – elongated highlights and increased contrast between low-clarity highlights and lowlights.

In Experiment 3, a new set of participants performed a multi-AFC task identical to Experiment 2, but with the image-transformed stimuli. The stimulus profiles were used to calculate factor scores in the factor space from Experiment 2, to discover the emergent categories of these new stimuli. The average factor scores for each stimulus are plotted in Figure 9B and C show that observers agree with our informal observations. In the pure manipulations condition, silver metals and fabric were seen as glazed ceramic and plastic, respectively. In the extra manipulations condition, the categories that observers perceive better align with the intended categories.

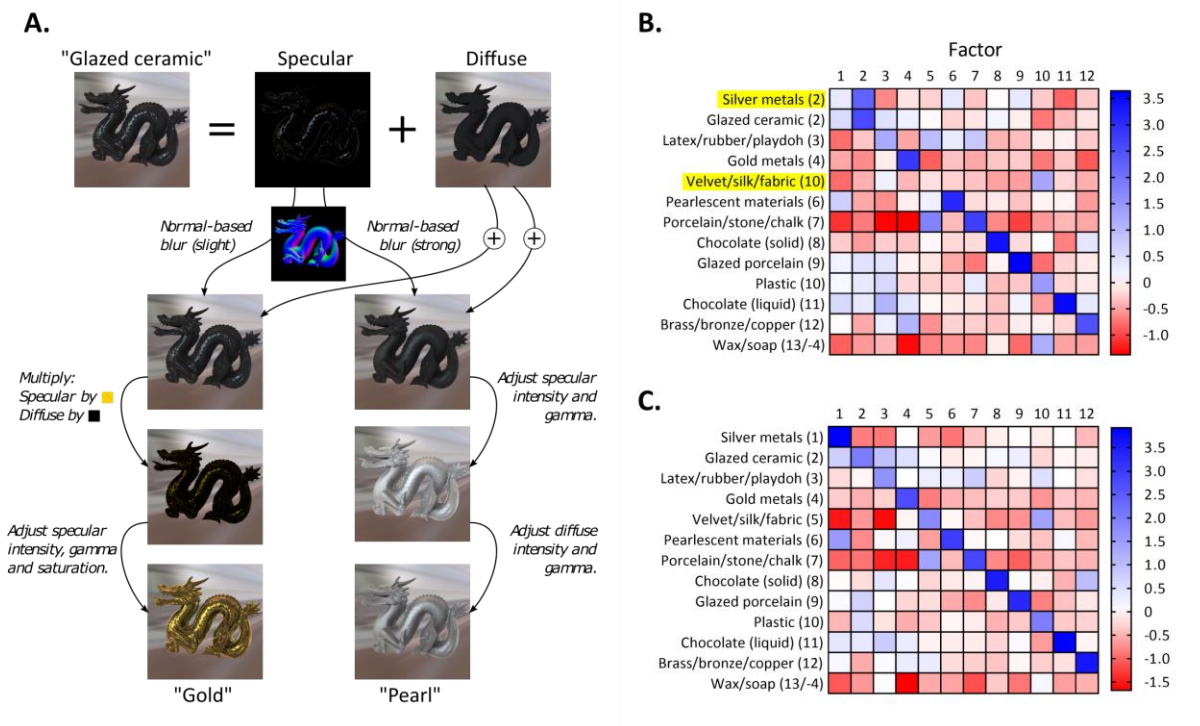

Figure 9. Perceived material categories resulting from manipulated visual features (Experiment 3). **A.** The manipulations were performed on specular and diffuse images separately before being recombined; as in Figure 6, the specular component is obtained by subtraction of the diffuse component from the full rendering. The sharpness of reflections is first modified by blurring pixels of the specular component that have sufficiently

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

similar normals (hence it requires a normal map). The subsequent filters adjust the colour and intensity of each component. For instance, for materials like gold (left column), the specular component is first multiplied by a colour (the diffuse component is discarded, or multiplied by 0); then its intensity and saturation are adjusted. For materials like pearl (right column), the intensity of the diffuse component is also adjusted. See the Methods section for a detailed description of each filter, along with the special filter used for velvet, based on a non-monotonic remapping of the specular component. **B.** Heat maps showing average factor scores of stimuli from each intended category after pure feature manipulations. The highlighted ones show intended categories that were misclassified. **C.** Heat maps showing average factor scores of stimuli from each intended category after extra feature manipulations (with gamma and fabric manipulations). With the addition of gamma correction for silver and the non-monotonic remapping of specular reflection intensity for fabric, stimuli are now perceived as the intended category. The labels in (B) and (C) refer to the intended category after transformation, and the numbers in brackets next to the category labels in (B) and (C) indicate the emergent category (factor) onto which each intended category most strongly positively loaded (except for wax/soap, where we took the strongest negative loading; see Figure 5). Note that a stimulus taken from the "glazed ceramic" emergent category was transformed into each category.

Material categories are not associated with a gloss level

Experiments 1-3 show that varying specular reflection image features transforms perceived material class. However, thus far the path to material categorisation is unclear. One possibility is that specular reflection features determine perceived gloss level (Marlow et al., 2010), which in turn combines with other estimated mid-level properties (such as colour/pigment) to determine material category (as in the *feedforward approach*). A second possibility is that the same specular reflection features are used to simultaneously "read out" both material category and surface gloss level (as in the *simultaneous approach*). A third possibility is that image features used to estimate surface gloss are mediated by material recognition.

If gloss (as a unitary concept) is used to help determine category, as is suggested by the feedforward approach, then the emergent categories from Experiment 2 should be associated with a particular gloss level. In Experiment 4, a separate group of participants rated the perceived glossiness of each of the 924 stimuli from Experiment 2. Inter-subject agreement for gloss ratings was high (median r = 0.70). Yet the white violin plots in Figure 7A show a wide distribution of gloss settings for the top 36 stimuli in each emergent category. Although some categories have a narrower range of gloss settings than others, stimuli from very visually distinct categories like glazed ceramics and (gold and silver) metals have completely overlapping distributions of gloss settings.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Note that this distribution would become even wider if we had plotted gloss ratings for all stimuli (i.e. not just the top 36).

It is possible that the high variability of gloss settings is caused by imperfect category membership, or that colour discriminates categories with similar gloss distributions. If the former is true, then perceived glossiness should correlate with the loading of stimuli onto category dimensions (i.e. factor scores). The bolded symbols in parentheses below the violin plots in Figure 7A show that stimulus factor scores correlate with perceived gloss level positively (+) or negatively (-) for only three of the emergent categories (glazed porcelain, plastic, and liquid chocolate). If all stimuli are considered (i.e. not just the top 36), then only about half of the emergent categories show a relationship between factor scores and perceived gloss level (unbolded symbols), with the other half showing no relationship at all (o).

RSA was used to determine the extent to which changes in category profile are reflected by systematic changes in perceived gloss and colour features (highlight saturation, lowlight saturation, and lowlight value). Colour feature RDMS are the same as those used in the previous section. A gloss-level RDM was created from gloss dissimilarity scores for each pair of stimuli (calculated as the absolute difference between gloss ratings). Dissimilarity scores in the upper triangles of each RDM were vectorized and subjected to a multiple linear regression, with gloss level and colour feature dissimilarities as the predictors and category profile dissimilarity as the outcome. We find that such a model performs worse (R^2 = 0.23) than the model with all six visual features as predictors (R^2 = 0.32; Figure 8C, right). Taken together, these results suggest that gloss as a unitary concept does not account for category membership and is unlikely to be a component dimension of material category. In the next section we test whether category and gloss are simultaneously computed from visual features (*simultaneous approach*).

Specular reflection image features predict both material categorisation and perceived gloss level

If the same visual features are used to simultaneously compute both material category and surface gloss level, as is suggested by the *simultaneous approach*, then the same visual features should predict both material categorisation and perceived gloss level. We used RSA to determine the extent to which changes in perceived gloss level is reflected by systematic changes in the measured gloss features. Dissimilarity scores in the upper triangles of each RDM were vectorized and subjected to a multiple linear regression, with gloss feature dissimilarities as the predictors and gloss-level dissimilarity as the outcome. The model predicts gloss level differences just as well as category

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

profile differences, $R^2 = 0.42$, $F(6,426419) = 50635$, $p < 0.001$. Interestingly, the relative contribution of features differs when predicting gloss level versus category profile. The structure coefficients in Figure 8C (middle) show that coverage and the colour features (lowlight value, lowlight saturation and highlight saturation) capture practically none of the explainable variance in gloss-level dissimilarity ($rs^2 = 0.002$, 0.009, 0.001, and 0.005, respectively). The fact that colour features play basically no role in perceived gloss level is backed up by comparing the performance of the full model to models containing only gloss features (coverage, sharpness, and contrast) and only colour features (lowlight value, lowlight saturation, highlight saturation) as predictors. Compared to the full model, which explained 42% of the variance in perceived gloss level differences, a model with the original gloss features performs nearly the same, $R^2 = 0.41$, $F(3,426422) = 98678$, $p < 0.001$, whereas a model with only colour features performs extremely poorly, $R^2 = 0.005$, $F(3,426422) = 674$, $p < 0.001$ (Figure 8C, right). Furthermore, sharpness contributes most to explaining differences in perceived gloss level ($rs^2 = 0.80$), followed by contrast ($rs^2 = 0.61$; Figure 8C, middle), a trend that was opposite for perceived category differences (Figure 8C, left).

This raises the possibility that features like sharpness and contrast are important for determining gloss level but play a reduced role for material categorisation relative to colour features. That is, it is possible that colour is driving much of the perceived difference in category profile. To test this, we compared the performance of the full model to a model with only coverage, sharpness, and contrast, and also to a model that had only the colour features of lowlight value, lowlight saturation, and highlight saturation (Figure 8C, right). Compared to the full model, which explained 32% of the variance in category profiles, a model with the original gloss features explains 24% of the variance in category profiles, $R^2 = 0.24$, $F(3,426422) = 44329$, $p < 0.001$, whereas a model with only colour features performs much worse, $R^2 = 0.08$, $F(3,426422) = 13487$, $p < 0.001$. Thus, while colour of the highlights and lowlights provide some information about category differences, gloss features like contrast and sharpness play a much more pronounced role in the perceived categorical differences between materials.

Image features that predict gloss level change across category

The results in the previous section suggest that the same visual features are responsible for both perceived gloss level and category judgments, albeit with different contributions. If these features are simultaneously but independently used to determine gloss level and material category, then the contribution of these features to perceived gloss level should be stable across categories. Figure 7C plots the correlation between gloss ratings and visual features for the top 36 stimuli from

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

each emergent category (x-axis) and for all stimuli in an emergent category (y-axis). This plot shows that the predictiveness of each feature to perceived gloss level (and indeed, *whether* there is a correlation, and the *direction* of the correlation) changes with material class. The grey zone shows significant correlations, which are also indicated with a star below the plots in Figure 7A (negative correlations are shown in red). Note that coverage, contrast, and sharpness predict gloss level even for emergent categories that show no relationship between gloss and factor scores (symbols in parentheses), providing further evidence against gloss as a component dimension of material category (*feedforward approach*). Also note that coverage, which showed almost no relationship with perceived gloss in the combined RSA, now shows correlations with perceived gloss for many emergent categories, and these correlations are often negative. However, such negative correlations are not consistent with previous literature (Marlow et al., 2012; Marlow & Anderson, 2013), and since coverage did not vary much for our stimulus set, we believe they arise from coverage negatively covarying with sharpness within these categories (Supplementary Figure 9).

Since the range of coverage, sharpness, and contrast differs with material class, one possibility is that the predictiveness of each gloss feature might rely on the availability of that feature within each category (in line with suggestions by Marlow & Anderson, 2013). To test this idea, we plotted the predictiveness of each feature against the interquartile range of each feature, for each category (Figure 7D). Surprisingly, there is no relationship between the predictiveness and “availability” (variance) of the gloss features. These results suggest that the pockets of feature space occupied by different material classes affect the processing of those very features when estimating surface glossiness. Our results do not support a traditional feedforward view that assumes that material perception proceeds from low-level image measurements, to mid-level estimates of surface properties, to high-level material classes, nor the idea that material properties like gloss and material class are simultaneously “read out” from visual features. Rather, our data suggest that the perception and neural processing of material properties like surface gloss should be considered in the context of material recognition. We discuss implications of this in the Discussion section.

DISCUSSION

Most surfaces in our environment reflect at least some amount of light specularly, and the image structure caused by specular reflections is often perceived as surface gloss. Recent advances in computer rendering software have led to a surge in studies investigating our perception of surface gloss, such as what are the photogeometric constraints for something to appear glossy versus matte (i.e. the relationship between intensity and shape; Beck & Prazdny, 1981; Blake & Bülthoff, 1990;

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Wendt et al., 2008; Todd et al. 2004; Anderson & Kim, 2009; Kim et al., 2011; Marlow et al., 2011; Kim et al., 2012), and what image features are responsible for the perceived level of surface gloss (Marlow et a., 2012, Marlow & Anderson, 2013). Thinking about surface gloss in this way (i.e. what makes a surface look more/less glossy) often leads to gloss being treated as a unitary concept (Motoyoshi et al., 2007), where there are certain image cues (e.g. coverage, contrast, sharpness) to gloss (Marlow et a., 2012). Here we showed that the image structure caused by specular reflections not only affects how glossy a surface looks, it can also alter the material quality, or category, of the material altogether. Critically, such specular reflection structure not only differentiates materials like plastic and metal (Todd & Norman, 2018); we showed that parametrically changing reflectance parameters also leads to qualitative changes in material appearance beyond those expected by the reflectance function used. A potential reason for this is that a material's surface reflectance properties create some of its most salient optical characteristics and relying on such characteristics could carry ecological importance when other cues to material are not available. For instance, the visual effects of translucency can be greatly diminished with frontal lighting (Xiao et al., 2014); in such cases, image features caused by specular reflections might remain diagnostic of translucent materials (e.g. porcelain). Similarly, meso-scale details like fibres on cloth or scratches on metal might not be visually resolvable when seen at a distance; yet such materials might still be recognised from specular image structure. Indeed, these additional sources of images structure (from mesostructure or translucency) are absent from the stimuli used in the present study, and although such details might render the materials more compelling (when say compared to a photograph), the stimuli nevertheless convincingly resemble silk-like, porcelain-like, wax-like, brushed metal-like, etc. materials. Interestingly, these results are in line with findings in the computer graphics literature that show that visual effects from different types of fabrics come predominantly from specular reflections, and diffuse reflection and shadowing-masking play a much less pronounced role even for relatively matte-looking fabrics (Irawan & Marschner, 2012). Indeed, our data support the notion that the visual system does not try to estimate the physical sources of image structure (after all, we do not have access to this information); instead, vision assigns perceptual qualities to statistically varying image structure (Anderson, 2011; Fleming, 2012, 2014, 2017; Purves, Morgenstern, & Wojtach, 2015).

Direct manipulation of images (Experiment 3) can better test which features are causally responsible for perceived category shifts, relative to correlational tests with measured image features (Experiment 2). For example, we found that operationalising contrast as a linear multiplication of intensity values (in the pure feature manipulation condition) was not adequate for two of the emergent categories; silver metals and fabrics seem to contain a different kind of contrast

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

than cannot be captured by such manipulations. However, despite the success of the feature manipulations for other categories, caution should be taken against taking these features at face value. For example, sharpness manipulations relied on surface normal information to preserve the photogeometric relationship between specular reflections and 3D shape. Since both object models that we used had a similar kind of mesostructure detail, any effect of changes in mesostructure detail that might contribute to categorisation could be missed. For example, "brokenness" of specular highlights could possibly be an important factor when comparing stimuli in the present study to smoother surfaces, or surfaces with a different kind of mesostructure. Similarly, as already mentioned, to properly investigate the role of coverage in both material categorisation and gloss perception, a wider range of stimuli with different shapes and light fields needs to be tested. Here, the purpose of identifying visual features that predict and transform material category was to act as a proof of concept that the appearance of specular reflections plays a much more pronounced role in material categorisation than previously appreciated. Though we sampled a much larger range of stimuli than previous studies, future studies should investigate which features discriminate material category for a wider range of stimuli with more diverse shapes, mesostructure detail, environment lighting, and reflectance parameters, and how the dimensionality of gloss appearance expands with these additional factors. Furthermore, how a change in measured or manipulated image features relates to the visual system's sensitivity to such changes (e.g. via perceptual judgments of such features; Marlow et al., 2012) is an important topic for future investigations.

One might wonder whether the emergent categories from our stimulus set could reflect cognitive decisions. For example, "chocolate" and "plastic" could have similar "gloss types" (i.e. similar specular reflection characteristics) and the different labels might result from a cognitive decision by participants based on body colour (brown versus yellow). However, we think that such a clear perceptual scission of image structure into different layers (e.g. a gloss layer and a body colour layer) and then subsequent "cognitive reassembly" is unlikely, especially for the complex-shaped stimuli and wide sampling of reflectance parameters used in the present study. This kind of layered appearance of gloss that can be construed as a form of transparency perception arguably only applies to smooth surfaces with little variation in surface curvature, like spheres (Kim et al., 2012). Notably, these are the kinds of surfaces that have been used previously to study the dimensionality of gloss perception. For example, Pellacini et al. (2000) manipulated the diffuse and specular components of glossy spheres and found three "gloss" dimensions corresponding to different aspects of the appearance of specular reflections – contrast, sharpness, and body colour (that is, the perceptual dimensions are similar to some of our measured image features). However, Pellacini et al. (2000) measured psychophysical relationships between stimuli within a narrow range of

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

parameter manipulations that likely encompassed the range of plastic- and ceramic-looking materials. Our results showed that expanding the range of stimuli that are normally tested and asking participants to judge material appearance greatly increased the perceptual dimensionality of glossy materials. Furthermore, for the present stimuli, image structure from diffuse and specular reflections seems to interactively combine to create a gestalt-like quality for each material. Like with object recognition, this material quality can be given a label like “gold” or “pearl”, but nonetheless reflects a holistic impression, rather than cognitive combination of features (see also Okazawa, Koida, & Komatsu, 2011, who provide further evidence for this argument). This holistic impression is likely why we could successfully causally manipulate features to transform one category into another (Experiment 3), but why linear models like LDA, and RSA (Experiment 2) did not better account for the data.

Another reason why we do not think our results merely reflect cognitive processes is that the predictiveness of each feature to perceived gloss changes with material category, suggesting that the pockets of feature space occupied by different material classes affect the processing of those very features when estimating surface glossiness. This suggests that the perception and neural processing of material properties like surface gloss should in fact be considered in the context of material recognition. For example, neurons in the lower bank of the superior temporal sulcus (STS) in monkeys have been shown to preferentially respond to specific combinations of manipulated gloss parameters (Nishio et al., 2012). It is possible that a categorical impression of materials is triggered by these specific combinations, driving this response. Our results also have implications for the object recognition literature (Grill-Spector & Weiner, 2014; Bracci, Ritchie, & Op de Beeck, 2017; Long, Yu, & Konkle, 2018), which often investigates how visual features like texture (Long, Yu, & Konkle, 2018) and form (Bracci, Ritchie, & Op de Beeck, 2017) influence object category representations, but thus far have neglected the role of surface properties in object recognition, for example by not controlling for changes in surface properties (Kaiser, Azzalini, & Peelen, 2016; Bracci et al., 2017, Zeman et al., 2019), or “scrambling” the images in a way that breaks the perception of surfaces (Long, Yu, & Konkle, 2018). Our results suggest that the specific photogeometric constraints on image structure that trigger our perception of surface gloss play an important role in visual categorisation, and suggest that a fruitful direction for neuroimaging could be to focus on identifying the neural mechanisms that represent objects holistically (Schmid & Doerschner, 2019).

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

METHODS

Stimuli

Stimulus generation: free-naming, multi-AFC, and gloss rating experiments (Experiments 1, 2, and 4)

We generated our stimulus set by computer rendering complex glossy objects under natural illumination fields. Object meshes were the Stanford Bunny and Dragon from the Stanford 3D Scanning Repository (Stanford University Computer Graphics Laboratory; http://graphics.stanford.edu/data/3Dscanrep/). Wavefront .obj files of these models were imported into the open-source modelling software Blender (v2.79) and rendered using the Cycles render engine, which is an unbiased, physically based path-tracing engine.

Stimuli were illuminated by the "kitchen" and "campus" light fields from the Debevec Light Probe Image Gallery (Debevec, 1998), and the resulting scenes were rendered at a resolution of 1080 × 1080 pixels. Interactions between light and surfaces were modelled using the Principled BSDF shader, which is based on Disney's principled model known as the "PBR" shader (Burley, 2012). The Principled BSDF shader approximates physical interactions between illuminants and surfaces with a diffuse and specular component (for dielectric materials), and a microroughness parameter that controls the amount of specular scatter. An advantage of the Principled BSDF shader over other models like the Ward model is that it accounts for the Fresnel effect, appropriately adjusted for microroughness level. The microfacet distribution used is Multiple-scattering GGX, which takes multiple bounce (scattering) events between microfacets into account, giving energy conserving results. Although there are many parameters to adjust in the Principled BSDF shader, we manipulated only the following (details can be found at https://docs.blender.org/manual/en/dev/render/shader_nodes/shader/principled.html):

- *Base Color:* Proportion of light reflected by the R, G, and B diffuse components.
- *Specular:* Amount of light reflected by the specular component. The normalised range of this parameter is remapped linearly to the incident specular range 0-8%, which encompasses most common materials. Values above 1 (i.e. above 8% specular reflectance) are also possible. A value of 5 (the maximum value used) translates to 40% specular reflectance.
- *Specular tint:* Tints the facing specular reflection using the base colour, while glancing reflection remains white.
- *Roughness:* Specifies microfacet roughness of the surface, controlling the amount of specular scatter.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

- *Anisotropic:* Amount of anisotropy for specular reflections. Higher values give elongated highlights along the tangent direction. This direction was set to radial for our stimuli.
- *Anisotropic rotation:* Rotates the direction of anisotropy, with a value of 1.0 indicating a rotation of 360°.

Rendering parameters: free-naming experiment (Experiment 1)

A total of 270 stimuli were generated for the free-naming experiment. Stimuli were generated by rendering combinations of the parameters for a dragon model embedded in the kitchen light field:

- *Base Color:* Six diffuse base colours based on two levels of saturation (greyscale, saturated yellow) and three levels of value (dark, medium, and light). For the greyscale stimuli, RGB values were equal for each lightness level (RGB = 0.01, 0.1, and 0.3 for dark, medium, and light stimuli, respectively). The RGB values for saturated stimuli were [0.01, 0.007, 0.001] for dark stimuli, [0.1, 0.074, 0.01] for medium stimuli, and [0.3, 0.221, 0.03] for light stimuli.
- *Specular:* Three specular levels: 0.1, 0.3, and 1, which correspond to 0.8%, 2.4%, and 8% specular reflectance, respectively.
- *Specular tint:* Two specular tint levels for saturated stimuli: 0 (no specular tint) and 1 (full specular tint).
- *Roughness:* Four roughness levels: 0, 0.1, 0.3, and 0.6.
- *Anisotropic:* Two anisotropic levels for stimuli with roughness levels greater than zero: 0 and 0.9.
- *Anisotropic rotation:* Two anisotropic rotations for anisotropic stimuli: 0 (no rotation) and 0.25 (90° rotation).

Rendering parameters: multi-AFC and gloss rating experiments (Experiments 2 and 3)

A total of 924 stimuli were generated for the multi-AFC and gloss rating experiments. Stimuli were generated by rendering combinations of the parameters for two shapes (dragon, bunny) and two light fields (kitchen, campus):

- *Base Color:* Six diffuse base colours based on two levels of saturation (greyscale, saturated yellow) and three levels of value (dark, medium, and light). For the greyscale stimuli, RGB values were equal for each lightness level (RGB = 0.01, 0.03, and 0.3 for dark, medium, and

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

light stimuli, respectively). The RGB values for saturated stimuli were [0.01, 0.007, 0.001] for dark stimuli, [0.03, 0.022, 0.003] for medium stimuli, and [0.3, 0.221, 0.03] for light stimuli.

- *Specular:* Four specular levels for the darkest four diffuse shading levels: 0.1, 0.3, 1, and 5, which corresponds to 0.8%, 2.4%, 8%, and 40% specular reflectance, respectively. The light-coloured stimuli were only rendered with the first three specular levels.
- *Specular tint:* Two specular tint levels for saturated stimuli: 0 (no specular tint) and 1 (full specular tint).
- *Roughness:* Three roughness levels: 0, 0.3, and 0.6.
- *Anisotropic:* Two anisotropic levels for stimuli with roughness levels greater than zero: 0 and 0.9.
- *Anisotropic rotation:* Two anisotropic rotations for anisotropic stimuli: 0 and 0.25.

Stimulus generation: feature manipulation experiment (Experiment 3)

We generated the feature-manipulated stimulus set with an approach similar to compositing techniques used in visual effects. For our manipulation we need several image components of an input scene: a full rendered image, an image of the object rendered with diffuse shading only (diffuse component), a binary mask image, and an image of the surface normals. All of these components were rendered with the Cycles engine in Blender: A manipulated image is obtained via a sequence of steps that are depicted in Figure 9A: a specular component is first obtained by subtracting the diffuse component from the full rendered image; the sharpness of this specular component is reduced via a normal-based blur operator; the diffuse and specular components are then optionally multiplied by a colour; finally, the intensity and saturation of the resulting specular and diffuse components are adjusted and the final image is obtained by addition of these components (specular + diffuse). These steps are explained in more detail below.

The normal-based blur operator is implemented with a cross bilateral filter (Paris et al., 2009) on the specular component and takes into account the orientation of the surface normals. Such a filter is controlled by a pair of so-called spatial and range weighting functions. The space weight is given by a box kernel of a large static size (160x160 pixels), which merely serves to limit the set of pixels that are considered for blurring. The range weight is given by an exponentiated normal dot product: $(\mathbf{n} \cdot \mathbf{n_0})^s$, where $\mathbf{n_0}$ is the normal at the centre pixel in the kernel, $\mathbf{n}$ is the normal at a neighbour pixel, and s is a shininess parameter. This simple range weighting formula is directly inspired by Phong shading: large shininess values result in a sharp angular filter, whereas small shininess values result in a broad angular filter, which has the effect of blurring specular reflections.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In practice, we use a blur parameter *b* in the [0..1] range, which is empirically remapped to shininess using $s = 1 + 1/((b/5)^4*100+\varepsilon)$, where ε is a small constant used to avoid division by 0.

Some particular categories of materials such as gold or chocolate require a colorized diffuse or specular component. This is obtained by simply multiplying the image component by a colour: *sCol* for specular, *dCol* for diffuse. In some cases (gold and silver), the diffuse component is entirely discarded, which is equivalent to having *dCol*=(0,0,0) as indicated in Figure 8A in the manuscript. We have made the choice of using bright and saturated colours in all other cases in order to make the parameters of the next step more easily comparable.

For the last intensity and colour adjustment step, each image component is first converted to the HSV colour space. The following manipulations apply either to the specular or diffuse component, with the corresponding parameters either prefixed by 's' or 'd'. The value channel is manipulated by a combination of gamma exponentiation (controlled by *dGamma* and *sGamma*) and multiplication (by *sBoost* or *dBoost*) using the following simple formula: $Boost*v^{Gamma}$, where *v* stands for value in HSV colour space. The saturation channel is also multiplied (by *sSat* or *dSat*). Both components are converted back to RGB and added together to yield the final image.

We use an additional filter for the specific case of velvet/satin. It is applied to the specular component right after normal-based blurring. The specular component is multiplied by a first mask m_1 that is computed via a non-linear mapping of the luminance *l* of the specular component. We use the following formula: $m_1 = f(0, l_m/2, l)$ if $l<l_m/2$, and $m_1=1-f(l_m/2, l_m, l)$ otherwise, with $f(a,b,x)$ a smooth Hermite interpolation between 0 and 1 when $a<x<b$, and l_m the maximum luminance in the specular component image. The effect of multiplication by m_1 is to darken the core of the brightest highlights, hence producing elongated highlight loops as seen in Supplementary Figures 5 and 7. It is similar in spirit to the sinusoidal modulation of Sawayama and Nishida (Figure 11 in Sawayama & Nishida 2018), except that it is only applied to the specular component and with a different non-linear remapping. We have also found it necessary to multiply the specular component by a second mask m_2 to slightly attenuate some of the elongated highlights. This mask is computed using a simple normal dot product: $m_2=(\mathbf{n} \cdot \mathbf{n_m})$, where **n** is the normal at a pixel and $\mathbf{n_m}$ is a manually-set direction that is chosen per scene.

Manipulation parameters: feature manipulation experiment (Experiment 3)

For each of the four scenes showing either of two objects (bunny and dragon) in either of two lighting environments (campus and kitchen), we have taken as input a material configuration previously classified as 'Glazed Ceramic', from which we have produced 12 manipulations for each of

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the other material categories. There were also two manipulation conditions (see below), yielding a total of 104 stimulus images for the feature manipulation experiment.

The input material was systematically chosen to have a grayscale *Base Color* of 0.1, a *Specular* level of 0.3, a *Specular Tint* of 0, a *Roughness* of 0, and an *Anisotropic* level of 0.

The manipulation parameters are listed in the table below. In the "pure" manipulation condition, all the *Gamma* parameters were set to 1 and the special velvet filter was discarded. In the "extra" manipulation condition, setting some of the *sGamma* and *dGamma* parameters away from 1 had an impact on the intensity of either the specular or diffuse component; as a result, we also had to adjust the *sBoost* and *dBoost* parameters. We applied the velvet filter only for the velvet manipulation in this condition, using the following $\mathbf{n_m}$ directions: (0.57,-0.53,0.63) for Bunny/Campus, (-0.22,-0.48,0.85) for Bunny/Kitchen, (-0.62,0.7,0.35) for Dragon/Campus, and (-0.72,0.57,0.39) for Dragon/Kitchen.

Category	Blur	sCol	sSat	sBoost	sGamma	dCol	dSat	dBoost	dGamma
Glazed ceramic	0	(1,1,1)	1	1	1	(1,1,1)	1	1	1
Brass/Bronze/Copper	0.45	(1,0.78,0)	0.6	1.8	1	(1,0.87,0)	0.25	0.75	1
Porc./Stone/Chalk	1	(1,1,1)	0	0.4	1	(1,1,1)	0	5.5	1
Chocolate (liquid)	0.05	(1,0.84,0)	0.4	0.5	1	(1,0.75,0)	0.4	1.4	1
Chocolate (solid)	0.5	(1,0.84,0)	0.2	0.75	1	(1,0.75,0)	0.4	1.4	1
Glazed Porcelain	0	(1,1,1)	0	0.25	1	(1,1,1)	0.3	6	1
Gold metals	0.15	(1,0.74,0)	0.6	2.25	1	(0,0,0)	0.75	2.75	1
Pearl	0.525	(1,1,1)	0.5	1.625	1	(1,1,1)	0.7	3	1
Plastic	0.35	(1,1,1)	0.625	1	1	(1,1,1)	0	1.75	1
Latex/Rubber/Playdoh	0.75	(1,1,1)	1.25	0.75	1	(1,1,1)	1	1.5	1
Silver metals	0.125	(1,1,1)	0.25	1.25	1	(0,0,0)	1	3	1
Wax/Soap	0.35	(1,1,1)	0	0.25	1	(1,0.84,0)	0.5	6.25	1
Velvet/Silk/Fabric	0.6	(1,0.95,0)	0.33	1.25	1	(1,0.82,0)	0.6	3.75	1

Table 1: Filter parameters for the "pure" image manipulation condition

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Category	Blur	sCol	sSat	sBoost	sGamma	dCol	dSat	dBoost	dGamma
Glazed ceramic	0	(1,1,1)	1	1	1	(1,1,1)	1	1	1
Brass/Bronze/Copper	0.45	(1,0.78,0)	0.6	**3**	**1.4**	(1,0.87,0)	0.25	**1.25**	**1.2**
Porc./Stone/Chalk	1	(1,1,1)	0	0.4	1	(1,1,1)	0	5.5	1
Chocolate (liquid)	0.05	(1,0.84,0)	0.4	0.5	1	(1,0.75,0)	0.4	1.4	1
Chocolate (solid)	0.5	(1,0.84,0)	0.2	**1.6**	**1.5**	(1,0.75,0)	0.4	1.4	1
Glazed Porcelain	0	(1,1,1)	0	**1**	**1.8**	(1,1,1)	0.3	**2.3**	**0.5**
Gold metals	0.15	(1,0.74,0)	0.6	**1.625**	**0.42**	(0,0,0)	0.75	0	1
Pearl	0.525	(1,1,1)	0.5	**1.1**	**0.6**	(1,1,1)	0.7	2.37	1
Plastic	0.35	(1,1,1)	0.5	**0.8**	**1.3**	(1,1,1)	0	**1.5**	**0.9**
Latex/Rubber/Playdoh	0.75	(1,1,1)	1.25	0.75	1	(1,1,1)	1	1.5	1
Silver metals	0.125	(1,1,1)	0.25	**1.25**	**0.3**	(0,0,0)	1	0	1
Wax/Soap	0.35	(1,1,1)	0	**0.13**	**0.7**	(1,0.84,0)	0.5	**2.2**	**0.5**
Velvet/Silk/Fabric	0.6	(1,0.95,0)	0.33	2.9	1	(1,0.82,0)	0.6	3.75	1

Table 2: Filter parameters for the "extra" image manipulation condition. Modified parameters compared to Table 1 are shown in bold.

Procedure

Participants and stimulus presentation

Participants were students at Justus Liebig University Giessen in Germany and were compensated 8 euros per hour. The experiments followed the guidelines set forth by the Declaration of Helsinki.

In the free naming task, stimuli were projected onto a white wall in a classroom. Thick black cloth was used to block light from the windows, so that the only source of light came from the projected image. For all other experiments the stimuli were presented on a Sony OLED monitor running at a refresh rate of 120 Hz with a resolution of 1920 x 1080 pixels controlled by a Dell computer running Windows 10. Stimuli were viewed in a dark room at a viewing distance of approximately 60cm. The only source of light was the monitor that displayed the stimuli.

Stimulus presentation and data collection were controlled by a MATLAB script (release 2018b, Mathworks, Natick, MA) using the Psychophysics Toolbox (Brainard, 1997).

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Free-naming task (Experiment 1)

Fifteen participants completed free naming experiment. All participants gathered in a classroom and completed the task at the same time. They viewed stimuli one at a time and were asked to classify the material of each stimulus, with no restrictions. They were provided with sheets of paper with space for each trial number to write down their answers. The experimenter controlled stimulus presentation with a keyboard press. Each trial was presented for as long as it took for all participants to finish writing down their responses. A blank screen was shown between each trial for one second. The experiment took approximately three hours to complete, including breaks.

The instructions, written on a sheet of paper in both English and German, were as follows:

> You will be shown 270 images, and your task is to write down your impressions of the material of each object, i.e. what does it look like it is made of? Below are some suggestions of materials that might help prompt you. Your answers might not include, nor are they limited to, the suggestions below. There are no right or wrong answers and you can respond however you like. You can be as specific (e.g. aluminium foil, polyethylene), or general (e.g. metal, plastic) as you like. You can also write down more than one material (e.g. "looks most like glazed ceramic but could also be plastic"), or even say that it doesn't look like anything you know. If you can't remember the name of a material, you can write e.g. "the stuff that X is made out of".

The following examples were provided below the instructions:

- Metal, e.g. silver, gold, steel, iron, aluminium, chrome, foil
- Textiles, e.g. velvet, silk, leather
- Plastic, e.g. PVC, nylon, acrylic, polyethylene, Styrofoam
- Ceramic, e.g. porcelain, china
- Minerals, e.g. stone, concrete, rock
- Coatings, e.g. glazed, painted, enamel, PTFE/Teflon
- Other: soap, wax, chalk, pearl, composite materials

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Multi-AFC task (Experiment 2)

Eighty native-level German speakers participated in the multi-AFC experiment. The 924 stimuli were randomly split into 4 sessions, so that each participant categorised a quarter of the stimuli (20 participants per stimulus). The experiment was self-paced with no time constraints, and for most participants the experiment lasted approximately 1-1.5 hours.

On each trial, observers were presented with an object in the centre of the screen (29° visual angle) with 18 categories displayed along the sides of the stimulus (Figure 4A). They were asked to choose the category that best applied to that stimulus. If they were unhappy with their choice, they could change the confidence rating at the bottom right of the screen. Before the experiment, participants were asked to read carefully through the list of materials and were shown several examples of the stimuli to be presented. The purpose of this was so observers got a sense of the range of materials in the experiment. Observers were restricted to choosing only one category for each stimulus, and were given the following instructions verbally by the experimenter:

> Use the mouse to click on the category that best describes the material of each object. When you are satisfied with your choice press the space bar to proceed to the next trial. In the case of categories with multiple items (e.g. velvet/silk/fabric), the perceived material only needs to apply to one, not all, the categories. There are no right or wrong answers, as the experiment is about the perception of materials. Not all categories will have an equal number of stimuli – you may choose one category more or less than others (or not at all). If you are not satisfied or confident with your choice, change the confidence rating at the bottom of the screen. This should be used in the case you feel that none of the available categories fit; if you think more than one category applies then just choose the most suitable option.

Feature manipulation experiment (Experiment 3)

Twenty-two native-level German speakers participated in the feature manipulation experiment. Each participant categorised all 104 stimuli, and the task was identical to the multi-AFC task in Experiment 2. The experiment was self-paced with no time constraints, and for most participants the experiment lasted approximately 45 minutes.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Gloss rating task

Twenty-two native-level German and English speakers participated in the gloss ratings experiment. One participant was excluded because they did not understand the task instructions, resulting in 21 participants. The experiment was self-paced with no time constraints, and for most participants the experiment lasted approximately 1 hour.

Before the experiment, participants were shown real world examples of glossy objects (some of which are shown in Figure 1A). As they were shown these images, they were given the following instructions:

> Many objects and materials in our environment are glossy or shiny. Glossy things have specular reflections, and different materials look glossier/shinier than others. You will be shown different objects and asked to rate how glossy each object looks. This experiment is about visual perception so there is no right or wrong answer – just rate how glossy each object looks to you.

Participants were shown many example trials before starting the experiment so that they could experience the full range of stimuli and calibrate their ratings to the range of gloss levels in the experiment. On each trial, observers were presented with an object in the centre of the screen (29° visual angle). The task was to rate how glossy the object was by moving the mouse vertically to adjust the level of a bar on the right side of the screen. They were told to look at many points on the object before making their decision. The starting level of the rating bar was randomly set on each trial.

ANALYSES

Free-naming experiment (Experiment 1)

A student assistant went through participants' responses of the free-naming task and extracted all category terms (nouns) and descriptors (e.g. adjectives describing fillings, coatings, finishes, and states), translating them to English. Each participant often used multiple terms per stimulus, all of which were recorded as separate entries. Similar terms (like foil/metal foil/aluminium foil, or pearl/pearlescent/mother of pearl/bath pearl) were combined into a single category. Some terms like "Kunstuff" and "Plastik" were considered duplicates because they translated to a single term (plastic) in English.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Multi-AFC experiment (Experiment 2)

Reduced set of category terms

Two of the authors (AS and KD) worked together to reduce the set of category terms that were generated in the free-naming task (the 29 category terms that at least 5 out of 15 participants used; see Figure 3A). Note that this was an arbitrary cut off with the aim of being quite inclusive while at the same time maintaining some consensus between participants. Materials that were visually or semantically similar were combined (specifically: fabrics, including velvet and silk; unglazed porcelain, stone, and chalk; latex and rubber; wax and soap; brass, bronze, and copper; iron and chrome; and silver and aluminium). This was guided by correlations between the categories (based on the number of participants that used each term for each stimulus; Supplementary Figure 2). The superordinate category "metal" was not included separately. In the free-naming task, non-category terms were often required to distinguish particular coatings (glazed/varnished), finishes (polished), or states (liquid/melting; see Figure 3B). For the multi-AFC task, we separated categories where these terms applied; for example, we included both glazed porcelain and unglazed porcelain, both liquid and solid chocolate, and also included a "covered in wet paint" category. Thus, the final set of 18 category terms (Figure 4) that participants could choose from in the multi-AFC task was as inclusive as possible while not being too large to overwhelm participants.

Factor analysis

The correlations between different categories in terms of their stimulus profiles (Supplementary Figure 3) indicate that some category terms were not independent. This suggests the existence of underlying common dimensions; that is, participants used the same underlying criteria for different category terms. An exploratory factor analysis was performed on the stimulus profiles (Figure 4B), which allowed us to explain some of this covariation and reveal the underlying category space of our stimulus set. Figure 5A shows that there is a steady increase in the common variance explained by each additional factor. The amount of extra variance explained by each additional factor did not drop off after a certain number of factors, indicated by the absence of a plateaux in this plot. Therefore, we arbitrarily chose to extract 12 factors on the basis that they were interpretable and explained over 80% of the common variance between stimulus profiles. Although this was an arbitrary cut-off, it is similar to or greater than the amount of variance explained by the factors/components retained in other material perception studies that have used factor analysis or principle components analysis (see Schmid & Doerschner, 2018, for a discussion of these studies).

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 5B shows a heat map of the factor loadings for each category, and Figure 5C shows the 13 emergent categories from the reduced category space after factor analysis, with example stimuli from each emergent category. Note that the 13th category results from the negative loadings onto Factor 4. We labelled these emergent categories based on the original categories that loaded most strongly onto each factor. Factor scores were calculated using a weighted least-squares estimate (also known as the "Bartlett" method).

Visual feature measurements

Calculations for most of these features relied on specular highlight coverage maps. For glossy surfaces with uniform reflectance properties (like the stimuli used in the present study), specular reflections cover the entire surface. However, for low-gloss objects we often only see specular highlights, or "bright" reflections, which are usually reflections of direct light sources like the sun, or a lamp. For very shiny surfaces (like metal) and in some lighting conditions we also see lowlights, or "dark" reflections (Kim et al., 2012), which are reflections of indirect light coming from other surfaces in the scene. We chose to define *coverage* as the amount of the surface covered in specular highlights, excluding lowlights, which is consistent with previous literature (Marlow et al., 2012; Cicco, Wijntjes, & Pont, 2019). Marlow and colleagues measured *coverage*, *contrast*, and *sharpness* of specular highlights using perceptual judgments from participants, due to the difficulty in segmenting the specular component of an image from other sources of image structure (such as diffuse reflectance and shading). An acknowledged concern of this approach is that it uses one perceptual output (perceived coverage, contrast, and sharpness) to match another (perceived gloss). It is unclear what image structure observers use to judge each feature, and participants might conflate their judgments of the visual features with each other and with perceived gloss (Marlow & Anderson, 2013; see also van Assen, Barla, & Fleming, 2018, for a similar use of this method). Therefore, we wanted to develop objective measures of specular reflection features. Currently there is no established method for segmenting specular reflections and diffuse shading from a single image that is robust across different contexts (e.g. changes in surface albedo, shape, and illumination conditions). To help with this segmentation we rendered additional images that isolated specular and diffuse components.

Extra rendered images. For each stimulus, a purely specular image was rendered, which had the same specular reflectance as the original (full rendered) image but with diffuse component turned off. Two purely diffuse images were rendered, which we call "diffuse image 1" (used to calculate *coverage*) and "diffuse image 2" (used to calculate *sharpness* and *contrast*; Figure 6). Kim

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

et al. (2012) separated specular highlights and lowlights by subtracting an image of a rendered glossy surface (with a specular and diffuse component) from an "equivalent" fully diffuse image with the same total reflectance as the glossy surface. Kim et al. used the Ward model (Ward, 1994) where the total reflectance could be easily matched between glossy and diffuse renderings because the glossy surface reflected the same amount of incident light at all viewing angles. Since the Principled BSDF simulates the Fresnel effect (whereby specular reflectance increases at grazing angles, depending on surface roughness and index of refraction), the diffuse (1) renderings were matched to the purely specular renderings in total reflectance at facing surface normal (i.e. surface orientations that face the camera). The second diffuse image (diffuse image 2) was created by rendering only the diffuse component of the original (full rendered) stimulus, with the specular component turned off.

Segmentation. Specular reflections were segmented into specular highlights and lowlights by subtracting the diffuse image 1 from the purely specular image, which resulted in a subtracted image (Figure 6b). This subtracted image was thresholded to serve as a coverage mask. For best results for our stimulus set, the threshold was set to one third of the maximum diffuse shading for each stimulus. Pixels above this threshold were considered highlights, and the remaining pixels were considered lowlights. For calculations of sharpness and contrast, specular reflections were segmented from diffuse shading by subtracting the diffuse image 2 from the original "full" stimulus.

Features. Coverage was defined as the proportion of object pixels that were calculated to be specular highlights (excluding lowlights). *Contrast* was the sum of root-mean-squared (RMS) contrast of extracted specular reflections at different spatial frequency bandpasses. *Sharpness* of extracted specular reflections was calculated for each pixel within the highlight regions using a measure of local phase coherence (Hassen, Wang, & Salama, 2013), then these values were averaged. *Highlight saturation* and *lowlight saturation* were calculated as the average colour saturation of pixels within the highlight region, and outside of the highlight region (which we call the lowlight region), respectively. *Lowlight value* was calculated as the average value of pixels in the lowlight region.

Feature manipulation experiment (Experiment 3)

The same visual feature measurements were used for the rendered stimuli (from Experiment 2) and the manipulated images (from Experiment 3), so that the measurements would be comparable (see Supplemental Figure 5A and C). However, slight modifications had to be made. For the rendered stimuli (Experiment 2), the segmentation between highlights and lowlights (described in the previous section) relied on two additional rendered images (the two input images

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

at top left in Figure 6 in the main text): a rendered specular image and a rendered diffuse image (1) with the same total reflectance. For Experiment 3, the feature manipulations are all applied to the same "Glazed Ceramic" material (see Methods). We reused the corresponding rendered specular and diffuse images from the stimulus set from Experiment 2, with one additional step: the specular image was further modified by the normal-based blur filter to account for the change in highlight coverage induced by a change in sharpness. Apart from this, the rest of the feature measurement routines remained unchanged.

ACKNOWLEDGEMENTS

A.S. and K.D. are supported by a Sofja Kovalevskaja Award endowed by the German Federal Ministry of Education, awarded to K.D. P.B. is supported by ANR project VIDA (ANR-17-CE23-0017). We thank Chris Baker for helpful comments on earlier versions of this manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

REFERENCES

Anderson, B. L. (2011). Visual perception of materials and surfaces. *Current Biology, 21*(24), R978-R983.

Anderson, B. L., & Kim, J. (2009). Image statistics do not explain the perception of gloss and lightness. *Journal of Vision, 9*(11):10, 1-17

Beck, J., & Prazdny, S. (1981). Highlights and the perception of glossiness. *Attention, Perception, & Psychophysics, 30*(4), 407-410.

Blake, A., & Bülthoff, H. (1990). Does the brain know the physics of specular reflection? *Nature, 343*(6254), 165–168.

Bracci, S., Ritchie, J. B., & Op de Beeck, H. (2017). On the partnership between neural representations of object categories and visual feature sin the ventral visual pathway. *Neuropsychologia, 105,* 153-164.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433-436.

Burley, B. (2012). Physically-based shading at Disney. In ACM SIGGRAPH 2012 Course: Practical Physically-based Shading in Film and Game Production. SIGGRAPH'12. <https://disney-animation.s3.amazonaws.com/library/s2012_pbs_disney_brdf_notes_v2.pdf>

Chadwick, A. C., & Kentridge, R. W. (2015). The perception of gloss: A review. *Vision Research, 109,* 221-235.

Cicco, F. D., Wijntjes, M. W. A., & Pont, S. C. (2019). Understanding gloss perception through the lens of art: Combining perception, image analysis, and painting recipes of 17th century painted grapes. *Journal of Vision, 19*(3):7, 1-15.

Debevec, P. (1998). Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In J. Grimes & G. Lorig (Eds.), *Proceedings of SIGGRAPH 98* (pp. 189–198). Orlando, FL: ACM.

Fleming, R. W. (2014). Visual perception of materials and their properties. *Vision Research, 94,* 62-75.

Fleming, R. W. (2017). Material perception. *Annual Review of Vision Science, 3,* 365-388.

Fleming, R. W., Storrs, K. R. (2019). Learning to see stuff. *Current Opinion in Behavioral Sciences., 30,* 100-108.

Foster, D. H. (2011). Color constancy. *Vision Research, 51,* 674-700.

Grill-Spector, K. & Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience, 15*(8), 536-548.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Hassen, R., Wang, Z., & Salama, M. M. A. (2013). Image sharpness assessment based on local phase coherence. *IEEE Transactions on Image Processing, 22*(7), 2798-2810.

Hunter, R. S. (1937). Methods of determining gloss. *Journal of Research of the National Bureau of Standards, 18*(1), 19–41.

Irawan, P., & Marschner, S. (2012). Specular reflection from woven cloth. *ACM Transactions on Graphics, 31*(1):11, 1-20.

Kim, J., Marlow, P. J., & Anderson, B. L. (2011). The perception of gloss depends on highlight congruence with surface shading. *Journal of Vision, 11*(9):4, 1-19.

Kim, J., Marlow, P. J., & Anderson, B. L. (2012). The dark side of gloss. *Nature Neuroscience, 15*(11), 1590-1595.

Komatsu, H., & Goda, N. (2018). Neural mechanisms of material perception: Quest on Shitsukan. *Neuroscience, 392,* 329-347.

Kriegeskorte, N., Mur, M. & Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2*(4), 1-28.

Long, B. Yu, C. P. & Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences, 115*(38) E9015-E9024.

Marlow, P. J., & Anderson, B. L. (2013). Generative constraints on image cues for perceived gloss. *Journal of Vision, 13*(14):2, 1-23.

Marlow, P. J., Kim, J., & Anderson, B. L. (2011). The role of brightness and orientation congruence in the perception of surface gloss. *Journal of Vision, 11*(9):16, 1-12.

Marlow, P. J., Kim, J., & Anderson, B. L. (2012). The perception and misperception of specular surface reflectance. *Current Biology, 22*(20), 1909–1913.

Motoyoshi, I., Nishida, S. Y., Sharan, L., & Adelson, E. H. (2007). Image statistics and the perception of surface qualities. *Nature, 447*(7141), 206–209.

Nishio, A., Goda, N., & Komatsu, H. (2012). Neural selectivity and representation of gloss in the monkey inferior temporal cortex. *The Journal of Neuroscience, 32*(31), 10780-10793.

Okazawa, G., Koida, K., & Komatsu, H. (2011). Categorical properties of the color term “GOLD”. *Journal of Vision, 11*(8):4, 1-19.

Paris, S., Kornprobst, P., and Tumblin, J. (2009). *Bilateral Filtering.* Now Publishers Inc., Hanover, MA, USA.

Pellacini, F., Ferwerda, J. A., & Greenberg, D.P. (2000). Toward a psychophysically-based light reflection model for image synthesis. *Proceedings of SIGGRAPH,* 55–64.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.892083; this version posted January 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Purves, D., Morgenstern, Y., & Wojtach, W. T. (2015). Will understanding vision require a wholly empirical paradigm? *Frontiers in Psychology, 6*, 1-6.

Sawayama, M., & Nishida, S. (2018). Material and shape perception based on two types of intensity gradient information. *PLoS Computational Biology, 14*(4).

Schmid, A. C., & Doerschner, K. (2018). Shatter and splatter: The contribution of mechanical and optical properties to the perception of soft and hard breaking materials. *Journal of Vision, 18*(1):14, 1-32.

Schmid, A. C., & Doerschner, K. (2019). Representing stuff in the human brain. *Current Opinion in Behavioral Sciences, 30,* 178-185.

Tamura, H., Higashi, H., & Nakauchi, S. (2018). Dynamic visual cues for differentiating mirror and glass. *Scientific reports, 8*(8403), 1-12.

Todd, J. T., & Norman, J. F. (2018). The visual perception of metal. *Journal of Vision, 18*(3):9, 1-17.

Todd, J. T., & Norman, J. F. (2019). Reflections on glass. *Journal of Vision, 19*(4):26, 1-21.

Todd, J. T., Norman, J. F., & Mingolla, E. (2004). Lightness constancy in the presence of specular highlights. *Psychological Science, 15*(1), 33–39.

van Assen, J. J. R., Barla, P. & Fleming, R. W. (2018). Visual features in the perception of liquids. Current Biology,

Vangorp, P., Barla, P., & Fleming, R. W. (2017). The perception of hazy gloss. *Journal of Vision, 17*(5):19, 1-17.

Ward, G. J. (1994). The RADIANCE lighting simulation and rendering system. *Proceedings of SIGGRAPH, 28,* 459-472.

Wendt, G., Faul, F., & Mausfeld, R. (2008). Highlight disparity contributes to the authenticity and strength of perceived glossiness. *Journal of Vision, 8*(1):14, 1-10.

Xiao, B., Walter, B., Gkioulekas, I. Zickler, T., Adelson, E. (2014). Looking against the light: How perception of translucency depends on lighting direction, *Journal of Vision, 14*(3):17, 1-22.

Ziglari, L. (2017). Interpreting multiple regression results: β weights and structure coefficients. *General Linear Model Journal, 43*(2), 13-22.