Perceptual qualities and material classes

Roland W. Fleming Department of Psychology, Justus-Liebig-Universität Gießen, Gießen, Germany

Christiane Wiebel Department of Psychology, Justus-Liebig-Universität Gießen, Gießen, Germany

Karl Gegenfurtner Department of Psychology, Justus-Liebig-Universität Gießen, Gießen, Germany

Under typical viewing conditions, we can easily group materials into distinct classes (e.g., woods, plastics, textiles). Additionally, we can also make many other judgments about material properties (e.g., hardness, rigidity, colorfulness). Although these two types of judgment (classification and inferring material properties) have different requirements, they likely facilitate one another. We conducted two experiments to investigate the interactions between material classification and judgments of material qualities in both the visual and semantic domains. In Experiment 1, nine students viewed 130 images of materials from 10 different classes. For each image, they rated nine subjective properties (glossiness, transparency, colorfulness, roughness, hardness, coldness, fragility, naturalness, prettiness). In Experiment 2, 65 subjects were given the verbal names of six material classes, which they rated in terms of 42 adjectives describing material qualities. In both experiments, there was notable agreement between subjects, and a relatively small number of factors (weighted combinations of different qualities) were substantially independent of one another. Despite the difficulty of classifying materials from images (Liu, Sharan, Adelson, & Rosenholtz, 2010), the different classes were well clustered in the feature space defined by the subjective ratings. K-means clustering could correctly identify class membership for over 90% of the samples, based on the average ratings across subjects. We also found a high degree of consistency between the two tasks, suggesting subjects access similar information about materials whether judging their qualities visually or from memory. Together, these findings show that perceptual qualities are well defined, distinct, and systematically related to material class membership.

Introduction

In everyday life, we are usually extremely good at recognizing different materials, such as wood, plastic, or soap, based on their visual appearance. For example, if we look at an office chair, we not only identify that the object as a whole is a chair, but can also readily identify what the component parts are made out of—stainless-steel legs, textile covering, plastic armrests, and so on. Indeed, based on our subjective experience, it seems likely that our ability to identify different classes of material probably rivals our ability to identify different classes of object. The range of materials we encounter is vast, and yet we can make many remarkably subtle material judgments, such as whether fruit is freshly cut or half an hour stale or whether furniture is veneered with real or fake wood.

One interesting related observation is that most materials that we encounter appear to belong to some kind of natural class, such as "plastic," "metal," "stone," or "fabric." A few example materials are presented in Figure 1. Even if we do not have a readily available verbal name for a given class of material (e.g., "the stuff on the inside surface of a banana skin"), we can usually relate a given example of the material to a psychological concept and liken the material to other similar exemplars we have seen previously. This is an impressive achievement because the members of a given class can vary widely in terms of their visual appearance. For example, the natural class "metal" includes such diverse appearances as mercury, copper, lead, and bronze. Metal objects come in an enormous variety of shapes and sizes, from needles to manhole covers to helicopters, and yet somehow we are able to group metal materials together and make inferences about new exemplars based on our experiences with

Citation: Fleming, R. W., Wiebel, C., & Gegenfurtner, K. (2013). Perceptual qualities and material classes. *Journal of Vision*, *13*(8):9, 1–20, http://www.journalofvision.org/content/13/8/9, doi:10.1167/13.8.9.

doi: 10.1167/13.8.9 Received December 12, 2012; published July 11, 2013 ISSN 1534-7362 © 2013 ARVO

Figure 1. Examples of different materials (from left to right: marble, fabric, bark, paper). We can generally assign the materials to distinct psychological classes and also make judgments about the perceptual qualities of individual samples.

other members of the class. This is a major challenge for the visual system to overcome.

To make matters more difficult, in image terms, members of different classes may sometimes be more similar to one another than to members of the same class. For example, the reflectance properties, shape, and mesoscale texture of a piece of limestone may be more similar to bread or sponge than to a quartz crystal, and yet we most probably group both limestone and quartz into the class of "stones" while "sponge" and "bread" are quite different categories. Clearly, the process of assigning a given material to an appropriate class is computationally a difficult problem for the brain to solve. Rather little is known about how we recognize and classify materials, at least relative to what is known about the recognition and categorization of faces or objects (although we review some work on the topic below). Here we aim to gain some insights into material classification.

Another important observation about material perception is that it is not limited to assigning a material to a specific class. We can also make many judgments about the *perceived qualities* of different materials irrespective of their class membership. For example, we can tell whether a material is soft or hard, rough or smooth, glossy or matte, flexible or fragile, etc. In some cases, qualities may be perceptual counterparts of objective material attributes, such as viscosity or elasticity. In other cases, the qualities may be highly subjective, such as whether the material appears "beautiful" or "comforting," for example. Nevertheless, in both cases, estimating or attributing perceptual qualities represents a parallel type of material perception judgment, which is somewhat independent from assigning materials to classes. Some qualities are clearly shared by materials from different categories (e.g., woods and stones are both usually perceived to be hard) while some qualities may vary substantially within a category (e.g., stone can be completely opaque, like chalk, or highly translucent, like jade). Furthermore, we can subjectively compare the surface properties (e.g., degree of glossiness) of two materials that may belong to different perceived classes, again suggesting that judgments of material qualities can be independent of the material class.

At the same time, these two types of judgments (assigning class membership and estimating material qualities) clearly interact with one another. For example, identifying that a material is a fabric gives us access to all kinds of stored semantic knowledge about fabrics, such as the fact that most fabrics are flexible, rather than rigid. In other words, assigning class membership can aid in the estimation of subjective qualities. This is especially important for properties such as density or friction, which may not easily be estimated visually from a given sample of the material or viewpoint. At the same time, the reciprocal interaction is also quite likely. In other words, judging the properties of a material probably plays a key role in assigning the material to a given class. For example, perceiving that a surface is soft, flexible, and fibrous presumably helps us to work out that the material is probably a textile even if we have never seen this particular type of textile before.

The relationship between judgments of material qualities and assignment of materials to classes raises a number of interesting questions about how materials are estimated and represented in the human mind. How distinct are different classes from one another in terms of their perceptual qualities? In other words, to what extent can class assignments be predicted by appearance qualities? What is the relationship between visual estimates of material qualities and stored knowledge about material classes? How distinct are different material qualities from one another? Are there a small number of basic dimensions or a large number of different properties that can be judged independently of one another?

To gain insight into these questions, we performed two experiments on the relationship between material qualities and class membership. In the first experiment, subjects were shown images of different exemplars from 10 different material classes and had to rate the extent to which each sample manifested nine different appearance qualities. This allowed us to gain insight into how different members of a given class are related to one another and how appearance quality judgments relate to class assignments. In the second experiment, subjects were given the verbal name of a given material class and a questionnaire sheet containing 42 different adjectives describing material qualities. Their task was again to rate the extent to which they believed members of that class manifested each material quality. By design, the two experiments are complementary. The visual experiment includes multiple samples from each category but fewer material qualities. By contrast, the semantic experiment contains only class labels but allows us to test the relationships between many more material qualities. By comparing responses in the two experiments, we can determine the extent to which judgments of material qualities for each class are consistent across visual and semantic domains.

Previous work

As mentioned above, relatively little is known about material classification—at least compared to object classification. A number of authors have attempted to derive feature spaces for describing the relationships between different textures much like color spaces describe the relationships between different colors. For example, Heeger and Bergen (1995) and Portilla and Simoncelli (2000) have developed texture analysis and synthesis algorithms that describe the statistical properties of texture patches parametrically. Although they are not explicitly designed to provide metrics for comparing images, the feature spaces used by these algorithms could be used to measure the similarity between different surfaces for classifying images of materials or for relating them to one another.

These statistical texture models have a large number of parameters that are designed to capture image structure in general rather than to identify the specific degrees of freedom relating different textures or materials to one another. To identify the "psychological dimensions" of texture, Rao and Lohse (1996) assessed 56 textures from the Brodatz album along 12 intuitive dimensions, including contrast, roughness, coarseness, and regularity. Subjects rated the images based on visual inspection. Eight texture classes were derived, which could be represented along three major axes by analogy to the cardinal axes of color space. The authors interpreted these axes as "repetitiveness," "directionality," and "contrast" as well as coarseness and complexity. In a follow-up study, Bhushan, Rao, and Lohse (1997) asked subjects to categorize words describing textures and tested whether there was a correspondence between the ratings in the verbal and visual domain. Results confirmed a three-dimensional data structure with similar semantic axes for categorizing texture words. Moreover, a strong correspondence between texture words and texture image dimensions was found when subjects were asked to map them onto each other. Here, we take a similar approach to the more general problem of assessing *material* properties and classes rather than focusing solely on texture.

Matusik, Pfister, Brand, and McMillan (2003) made detailed measurements of the reflectance properties (BRDFs) of over 100 physical materials and used nonlinear dimensionality reduction techniques to identify a low-order embedding of the materials. He found that the BRDFs contained many statistical redundancies, allowing him to accurately approximate the materials using just 10 to 15 dimensions. The authors asked subjects to rate a variety of visual qualities (e.g., "glossiness," "redness") for all their samples. The resulting ratings were used to define visual "trait vectors" spanning the low-dimensional material manifold, which expressed how materials are related to one another in terms of the different traits. Based on this, it was possible to synthesize novel BRDFs that exaggerated or attenuated specific visual qualities (e.g., making a metal appear rustier). In follow-up work, Matusik, Zwicker, and Durand (2005); Ray, Levy, Wang, Turk, and Vallet (2009); and Ruiters, Schnabel, and Klein (2010) applied similar techniques to textures (rather than homogeneous BRDFs), enabling them to synthesize smooth transitions in texture appearance between different samples. These results strongly suggest that the visual system can attribute continuously varying perceptual qualities to different materials. However, the work does not establish how qualities are related to perceptual classes.

To gain insights into material classification, Sharan, Rosenholtz, and Adelson (2009) have developed the MIT-Flickr material database, consisting of 100 images from 10 different classes of materials downloaded from the photo-sharing site Flickr. Sharan and colleagues showed that subjects are surprisingly fast and accurate in assigning these images to their distinct material classes even with short presentation times. When presented for 40 ms, subjects correctly classified 83% of the images in a two-alternative forced-choice paradigm. This shows that material recognition can be achieved reasonably fast, and it has been suggested that a set of low-level and mid-level features can be used to characterize natural images of different material classes. Liu et al. (2010) developed an algorithm to classify material classes from natural images. The algorithm achieves 44.6% correct classification on the MIT-Flickr material database. Although not as good as humans, this performance is impressive given the wide variety of appearances present in the 10 material classes. In other words, the database is much less homogeneous than conventional texture databases.

Here, we also use a subset of the MIT-Flickr database, but instead of asking subjects to assign the images to classes, we asked them to rate subjective material qualities. We then infer class structure from the ratings to assess how the two types of tasks (classification and perceptual quality ratings) are related to one another. Thus, whereas previous work has investigated either classification or quality ratings, here we try to understand the connection between the two.

Experiment 1: Visual judgments of perceptual qualities

Methods

Overview

Subjects viewed photos of different materials and rated the materials for various perceptual qualities (e.g., glossiness, hardness, fragility).

Stimuli

Stimuli consisted of 512 $\times$ 384 JPEG photos, selected from the MIT-Flickr Materials Database (Sharan et al., 2009). Thirteen exemplars were selected from each of the 10 material categories in the database (fabric, foliage, glass, leather, metal, paper, plastic, stone, water, and wood), making a total of 130 items.

The order of the images was scrambled, and they were then compiled into a single PDF, one image per page. The PDF was presented to the subjects using the Apple Mac application "Preview," in slideshow mode with the "Actual Size" option (as opposed to "Full Screen") to reduce the visibility of JPEG artifacts. The images were presented using an Acer PD528 DLP beamer, projecting onto a white wall in the classroom. The sRGB profile was used but slightly adjusted to increase brightness. The shutters of the windows were lowered to improve contrast although the overall light level was still typical of a classroom or office (i.e., not a darkened room). To confirm that image intensity was homogeneous, we measured the luminance of a uniform white image at the four corners and center of the projection screen. The values ranged from 69.2 to 73.1 cd/m^2 with a mean of 71.76 and a variance of 2.46 cd/m2 (i.e., 3.4% of the mean).

Subjects

Nine master's students from the University of Giessen (eight female, one male) performed the experiment as part of a seminar course on color and material perception. All participants reported normal or corrected-to-normal visual acuity.

Procedure

All subjects sat together in a single room and viewed the images simultaneously on the projector screen. They entered their responses on laptops running Microsoft Excel. The experiment was organized into nine blocks of 130 trials. In each block of trials, a different perceptual quality was assessed, and in each trial within a block, subjects rated that quality for a single image. An alternative approach would have been to provide ratings for several perceptual qualities for each trial, but this has the disadvantage of constant task switching and would have made pacing the presentation of images more difficult.

The 130 images were shown in the same order in each block. In each trial, the subjects' task was to assess the image for the perceptual quality of the current block and enter a rating from one to six into the spreadsheet to record their response. Having assigned a value for a given perceptual quality to all 130 images, the subjects took a short break, and then the next block (i.e., next perceptual quality) was started.

Before each block, the perceptual quality to be judged in the forthcoming block was defined, and the polarity of the six-point scale (i.e., what low and high values correspond to) was explained. The subjects were encouraged to ask questions to clarify their understanding of the material property to be rated and the rating scale. Importantly, the subjects were not informed that the materials were grouped into distinct classes; they were simply told that they would see 130 images of various different materials. The following nine qualities were assessed with the following definitions:

- Glossiness: How glossy or shiny does the material appear to you? Low values indicate a matte, dull appearance; high values indicate a shiny, reflective appearance.
- Transparency: To what extent does the material appear to transmit light? Low values indicate an opaque appearance; high values indicate the material allows a lot of light to pass through it.
- Colorfulness: How colorful does the material appear to you? Low values indicate a grayish, monochrome appearance; high values indicate a colorful appearance, which could be either a strong single color or several colors.
- Roughness: If you were to reach out and touch the material, how rough would it feel? Low values indicate that the surface would feel smooth; high values indicate that it would feel rough.
- Hardness: If you were to reach out and touch the material, how hard or soft would it feel? How much

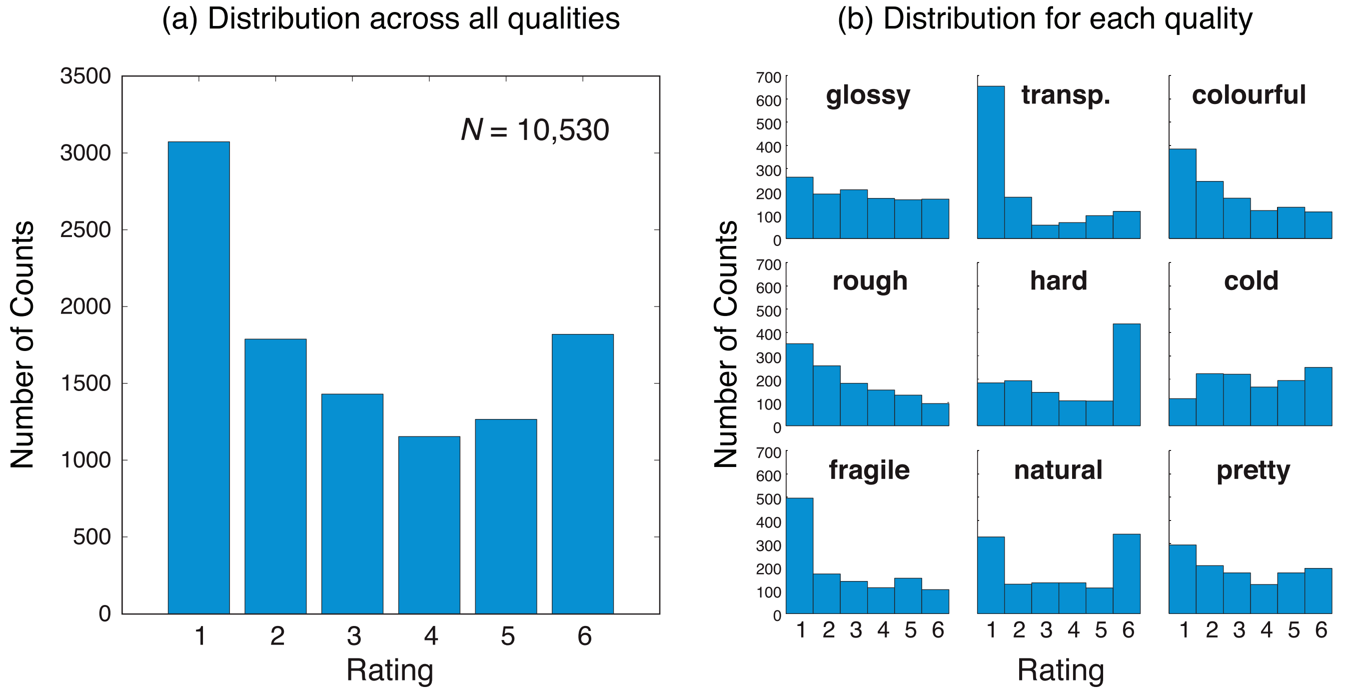


Figure 2. Histograms of rating values. (a) Overall distribution of rating values across all subjects, material classes, and qualities. (b) Distribution of rating values for each quality.

force would be required to change the shape of the material? Low values indicate that the surface would feel soft; high values indicate that it would feel hard.
- Coldness: To what extent would you expect the surface to feel cold to the touch? Low values indicate that the material would typically feel warm or body temperature; high values indicate that the material would feel cold to the touch.
- Fragility: How fragile or easy to break is the material? High values indicate that a small amount of force would be required to break, tear, or crumble the material; low values indicate that the material is highly resistant and could not easily be broken.
- Naturalness: How natural does the material appear to be? To what extent is the material in its most natural, common state? Low values indicate that the material appears unnatural; high values indicate that it appears natural.
- Prettiness: How pretty or visually attractive is the material to you? Low values indicate the material is ugly or unattractive; high values indicate that it is attractive or beautiful to the eye.

Within each block, the experimenter manually progressed through the images in relatively rapid succession (about 2 s per image). From time to time, when a student made a mistake or could not keep up, he or she shouted out, and we backtracked a few images as necessary, allowing the subject to correct the errors. Other than that, there was no communication whatsoever between participants: They were explicitly instructed not to confer during the experiment, and the experimenter monitored this. The other subjects did not adjust their ratings to the images that were seen again.

Results

Response distribution

Figure 2a shows the overall distribution of ratings pooled across all subjects and all trials, i.e., the total number of times subjects used each of the six different values on the rating scale. There are two notable aspects of the distribution. First, the histogram appears to be bimodal with subjects favoring extreme values (one and six) over intermediate values. Put another way, there does not appear to be a pronounced regression to the mean in the use of the scale. The fact that subjects were willing to assign extreme values suggests that they had strong, categorical impressions of the perceptual qualities and could determine with some confidence whether or not a material exhibited the quality of interest in each block. This suggests that overall, the various perceptual qualities that subjects rated could be interpreted in a meaningful way and that they could meaningfully be applied to materials.

The second observation is that the distribution is markedly asymmetrical with more ones than sixes. In other words, materials tended to exhibit extremely low

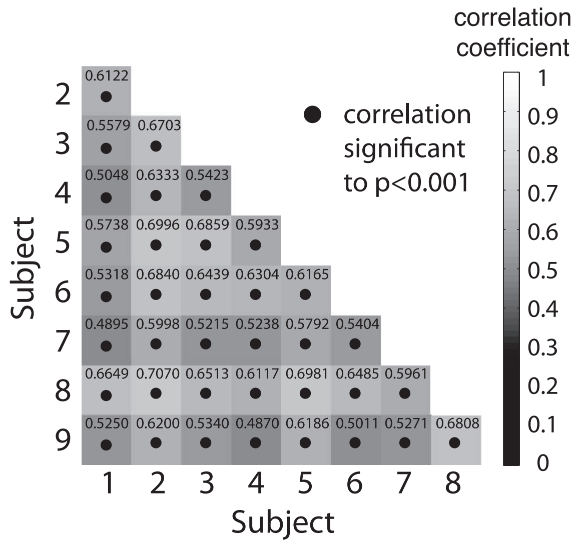


Figure 3. Intersubject correlations. Gray level indicates the correlation coefficient as specified by the color bar. Correlation coefficient values are stated in each cell. Dots indicate that the correlation in the corresponding cell is statistically significantly different from zero at the $p < 0.001$ level. Note that all correlations are significant and positive.

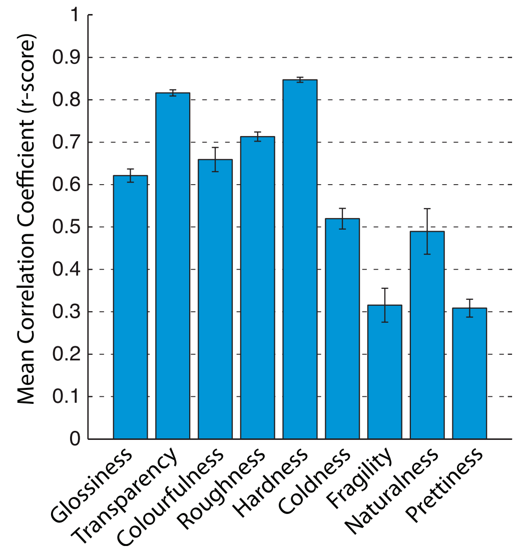


Figure 4. Mean correlations between subjects for each perceptual quality. Error bars indicate standard errors of the *r* scores.

values of the qualities more often than high values. Some caution is required in interpreting the meaning of this observation. The polarity of the scales is in some sense arbitrary. For example, subjects were asked to rate "glossiness" with high values assigned to glossy materials and low values assigned to matte materials. However, we could just as easily have asked subjects to rate the reciprocal quality (i.e., "matteness"), in which the glossiness scale is simply inverted. If all scales were inverted, the asymmetry would be reversed. By inverting only some of the scales, it would be possible to largely remove the asymmetry. Thus, this asymmetry probably does not indicate something profound about the distribution of materials within the nine-dimensional (9-D) quality space. An alternative explanation of the asymmetry would be a response bias; namely, for some reason, the subjects preferred to give extreme low values irrespective of the perceptual quality. However, inspection of the score distributions separated for each perceptual quality (Figure 2b) suggests this is not the case. Most of the individual distributions are skewed and unimodal, but the direction of skew varies from quality to quality. For example, "roughness" is dominated by low values (i.e., most materials were smooth) while "hardness" is dominated by high values. Other qualities (e.g., "transparency" and "naturalness") are bimodal, indicating that materials tended to exhibit either high or low values but rarely intermediate ones. Thus, a constant response bias cannot account for the asymmetry.

Consistency across subjects

It is interesting to ask to what extent subjects are consistent in their judgments of material qualities. It could be that each subject interpreted the meaning of each perceptual quality differently or that there is a highly subjective aspect to some of the qualities (e.g., "prettiness"), such that a given image could appear to have a high value to one subject but a low value to another. This would lead to a high degree of inconsistency between the different subjects. By contrast, if the subjects agreed on the values assigned to each material, it suggests that the perceptual qualities are in some sense objective and meaningful. To test this, we estimated the correlations between subjects' responses across all images and perceptual qualities.

Figure 3 plots the correlation matrix comparing each subject's ratings to all the other subjects. Low correlations are plotted as dark grays ($r = 0$ would be black), higher correlations as lighter grays ($r = 1$ would be white). As can be seen, all subjects are substantially positively correlated with one another: All correlations are significantly above zero with *r* scores ranging from 0.4870 to 0.7070. The mean and standard deviation of all the pair-wise *r* scores are 0.5974 and 0.0669, respectively. This suggests that subjects were substantially consistent with one another in their assignment of perceptual quality ratings to the 130 different materials in the experiment.

We can also separate the correlations between subjects for each perceptual quality to measure the extent to which the different perceptual qualities elicited similar ratings from the subjects. We find that

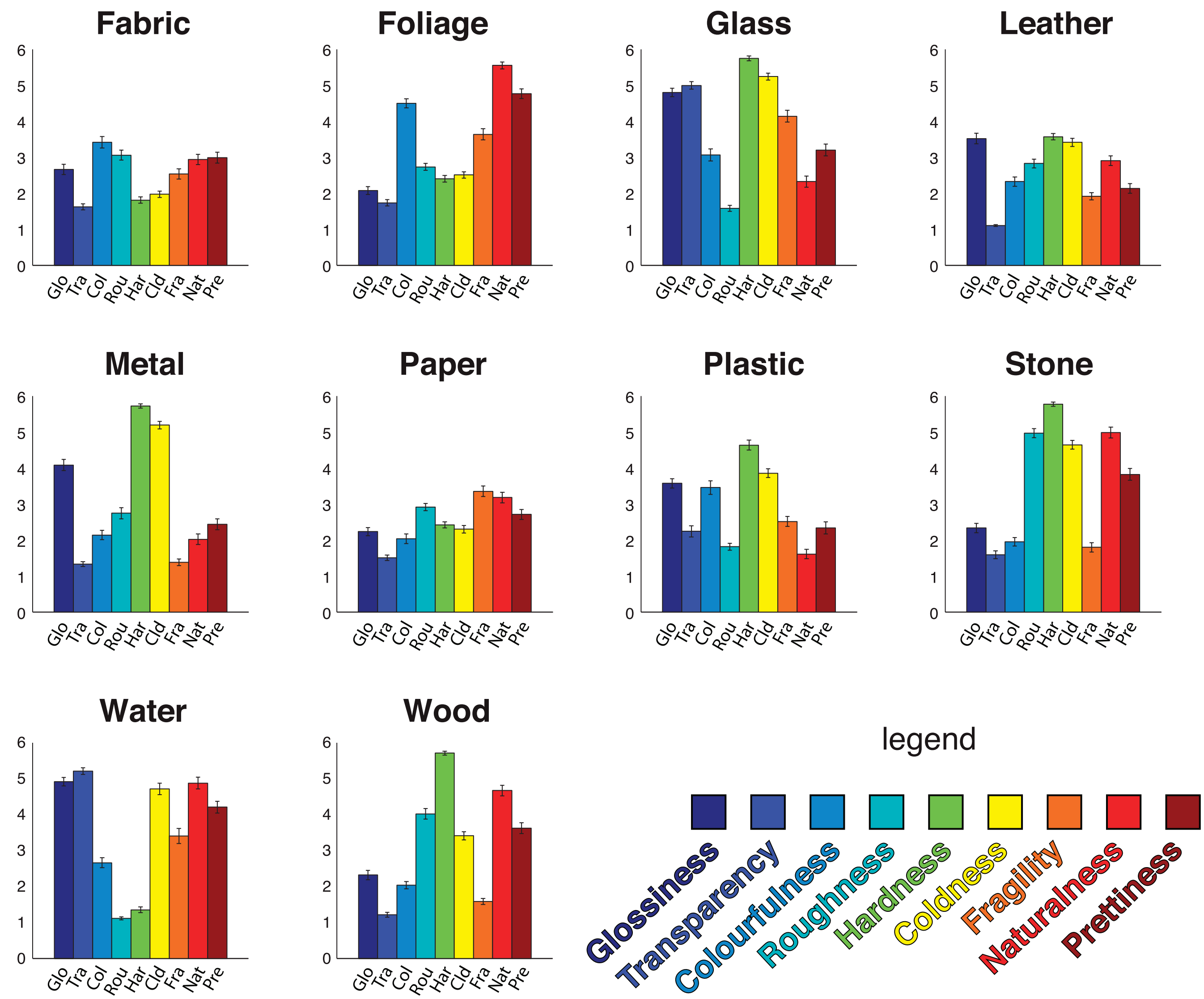


Figure 5. Mean quality scores for each material class. Error bars represent standard errors of the mean. Different material classes have distinctly different feature signatures.

some qualities, notably "transparency" and "hardness," are highly consistent across subjects. Other qualities (e.g., "naturalness") are largely consistent but with one or two subjects that differ from the others. Finally, some qualities (e.g., "prettiness," "fragility") are much less consistent across subjects. These differences are summarized in Figure 4, which plots the mean of the r scores in each correlation matrix for each perceptual quality. Error bars indicate standard errors of the r scores.

Ratings for each material class

Subjects were not informed that the 130 different images consisted of 10 distinct material classes. It is interesting to ask to what extent the different material classes nevertheless exhibited distinctive patterns of responses across qualities. Intuitively, we might expect materials within a class to share certain qualities that distinguish them from other material classes. For example, objects made out of glass tend to be quite glossy, highly transparent, variable in colorfulness, generally rather smooth, very hard, cold to the touch, fragile, artificial (as opposed to natural), and with varying values of prettiness (one's predilection for cut glass, or lack thereof, notwithstanding). By contrast, water, which is also glossy and transparent, probably differs from glass in terms of "hardness" and "naturalness." Thus, the ratings of different qualities could form a distinctive feature "signature" for each class of materials.

Figure 5 shows the mean ratings of each quality for each material class, averaged over the nine subjects and 13 different exemplars within each class. Error bars

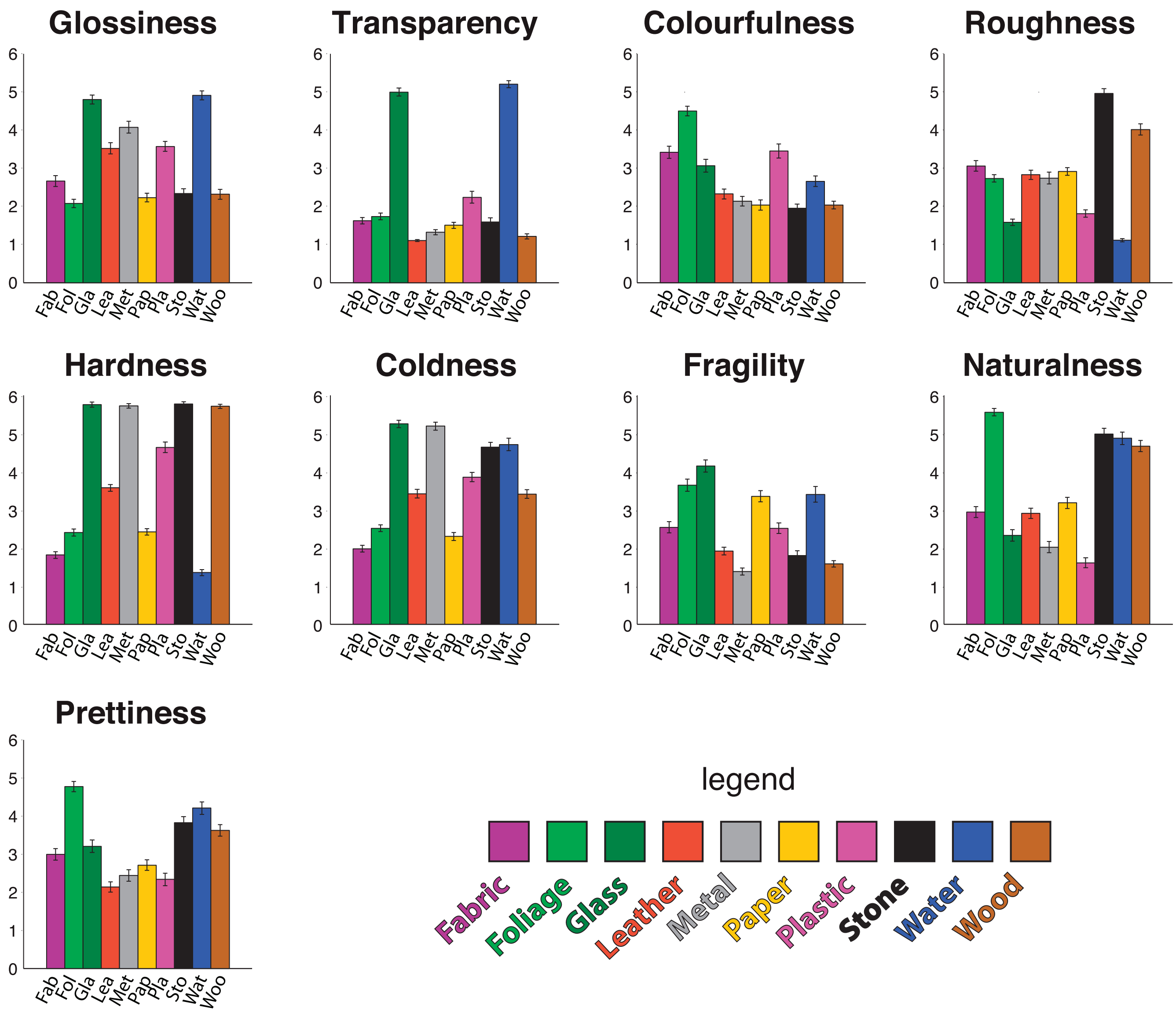


Figure 6. Mean ratings for each perceptual quality (same data as in Figure 5 but regrouped by perceptual quality). Error bars represent standard errors of the mean.

indicate standard errors. As expected, different classes of materials tend to have distinctive signatures of different qualities. For example, water is high in "glossiness" and "transparency" but very low in "roughness" and "hardness" whereas stone is roughly the opposite for these qualities. Other materials also have broadly intuitive signatures.

It is interesting to note at this stage that there seem to be some important correlations between different classes, suggesting a smaller number of underlying degrees of variation along which the different classes may be clustered. For example, stone, wood, and, to some extent, metal have somewhat similar signatures. We return to these correlations below.

We can also plot the same data grouped by the perceptual qualities as shown in Figure 6. Here, each bar indicates the average ratings for a different material class, and error bars again indicate standard errors.

The results are again broadly intuitive. For example, for the perceptual quality "transparency," most material classes (fabric, foliage, leather, metal, paper, plastic, stone, and wood) receive low scores whereas glass and water materials receive high average ratings. The overall variance is relatively large.

Correlations between qualities

As we have seen, different qualities have different distributions across the material classes, suggesting that

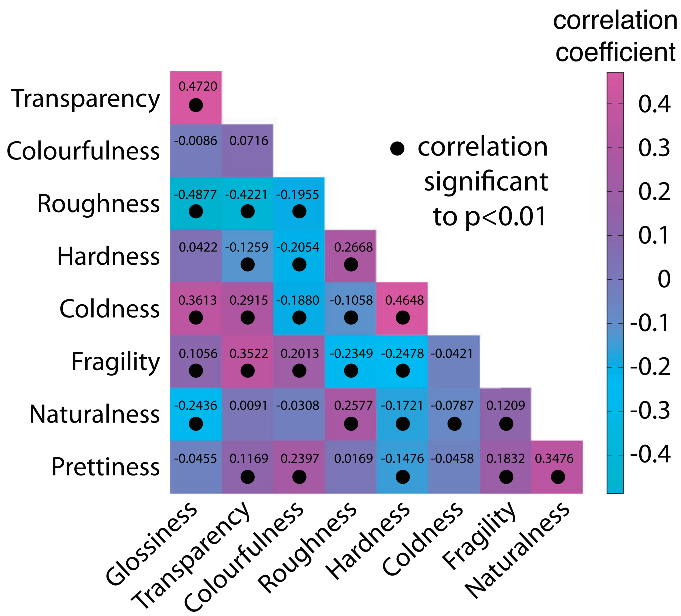


Figure 7. Correlation matrix relating the nine different perceptual qualities to one another. Colors indicate correlation coefficients as specified by the color bar. Pinks indicate positive correlation; blues indicate negative correlation. Correlation coefficient values are stated in each cell. Dots indicate that the correlation in the corresponding cell is statistically significantly different from zero at the $p < 0.01$ level.

these qualities provide a means to distinguish between types of material. However, to what extent are these different perceptual qualities truly independent from one another, and to what extent are they correlated? A priori, we might expect some qualities to correlate with one another. For example, "glossiness" and "transparency" are likely to be somewhat correlated because the physics of dielectrics means that materials that are highly transparent are usually also specular. By contrast, "colorfulness" and "naturalness" are presumably independent properties as both natural and artificial materials can be strongly or weakly colored. The correlation matrix relating the perceptual qualities to one another is shown in Figure 7.

The correlation coefficients range from −0.4877 to 0.4720, meaning that although most of the qualities are significantly correlated with one another the correlations are nevertheless small in magnitude. The most strongly positively correlated qualities are "glossiness" and "transparency" ($r = 0.4720$), followed by "coldness" and "hardness" ($r = 0.4648$). Both of these correlations make intuitive sense from a physical point of view. Materials that are hard, like metal, glass, and stone, are often also good conductors of heat, causing them to feel cold to the touch. By contrast, soft materials, like fabrics and paper, tend to trap air in their fibers, making them good insulators.

The most strongly negatively correlated qualities are "roughness" and "glossiness" ($r = -0.4877$) and "roughness" and "transparency" ($r = -0.4221$). Again, this makes intuitive sense because roughening a surface strongly affects the way it scatters light. Rough surfaces tend to be either almost completely diffuse (Oren & Nayar, 1994) or have very broad specular lobes (Torrance & Sparrow, 1967; Ward, 1992). This leads to extremely blurry, low-contrast highlights and, in the case of transparent surfaces, a frosted, translucent appearance. By contrast, smooth surfaces, like polished metal or glass, tend to make sharp, clearly visible highlights and refracted features.

Overall, however, it is important to note that the qualities are generally only weakly correlated with one another. Half of the correlation coefficients have an unsigned magnitude of less than 0.16, and about 80% have an unsigned magnitude less than 0.35. This indicates that while correlations do occur, broadly, the perceptual qualities are weakly correlated with one another, and subjects can treat them as distinct and independent attributes of materials. The correlations that are found seem to relate to true underlying correlations between the physical and functional properties of different materials rather than a psychological confounding of the different qualities in the subjects' minds.

Correlations between material classes

We have just considered correlations between different perceptual qualities and found that subjects can rate materials along the nine different qualities as if they were distinct dimensions. To what extent is the same true of different material classes? Are different material classes highly distinct, or do some classes correlate with one another? As we have already seen from the pattern of average ratings shown in Figure 5, some material classes (e.g., wood and stone) do tend to correlate relatively well with one another across the nine perceptual qualities we tested, whereas others (e.g., metal and fabric) appear to be quite independent. In Figure 8, we assess this in more detail by plotting the correlation matrix relating each material class to each other.

The correlation coefficients range from −0.3408 to 0.5815, indicating that the correlations between different material classes are generally relatively modest. The most strongly positively correlated material classes are stone and wood ($r = 0.5815$), followed by metal and leather ($r = 0.4141$). The correlation between stone and wood is fairly unsurprising: Of the material classes considered here, these are probably subjectively the most similar, both being naturally hard, often rough, brownish in color, and grained. Although, in everyday life, we would rarely confuse wood with stone, intuitively it seems that when compared with other classes (e.g.,

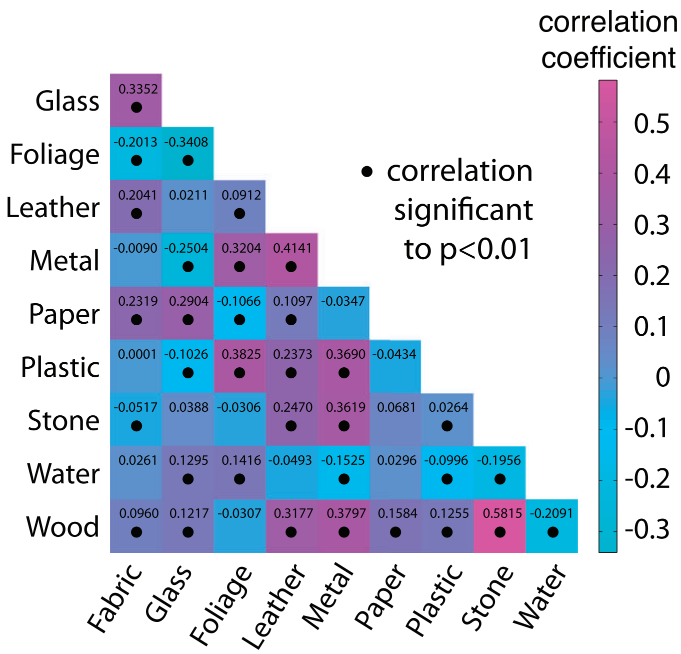


Figure 8. Correlation matrix relating the 10 different material classes to one another. Colors indicate correlation coefficients as specified by the color bar (note scale is not the same as in Figure 7). Pinks indicate positive correlation; blues indicate negative correlation. Correlation coefficient values are stated in each cell. Dots indicate that the correlation in the corresponding cell is statistically significantly different from zero at the $p < 0.01$ level.

fabric, water, or glass), wood and stone are rather similar to one another. The correlation between metal and leather is somewhat more surprising: When asked verbally which classes are most similar to one another, none of the subjects reported these classes. At the same time, it is also important to note that correlation does not capture the absolute similarity in ratings between two classes of material; it simply measures the tendency for different perceptual qualities to increase or decrease in union. Paired sample t tests between the ratings for each perceptual quality for leather and metal materials reveals that the two classes were significantly different for six of the nine perceptual qualities: "glossiness" ($t = -3.3698$; $p < 10^{-3}$), "transparency" ($t = -3.3360$; $p < 10^{-3}$), "hardness" ($t = -20.1050$; $p < 10^{-39}$), "coldness" ($t = -11.8454$; $p < 10^{-21}$), "fragility" ($t = 4.1446$; $p < 10^{-4}$), and "naturalness" ($t = 4.9169$; $p < 10^{-6}$). Thus, correlations alone do not capture the true perceived similarities between different material classes. To evaluate this, we must consider the distribution of the samples in the 9-D feature space defined by the different perceptual qualities.

Clustering of material classes in the space of perceptual qualities

To aid visualizing how the images are distributed in the feature space of perceptual qualities, we performed principal component analysis (PCA) on the mean ratings across subjects. The factor loadings of the first two principal components (PCs, see Figure 9a) indicate that PC1 is strongly positively loaded by "glossiness" and "transparency" and negatively by "roughness" whereas PC2 contrasts positive loading on "naturalness" against stronger negative loadings on "hardness" and "coldness." Although it makes intuitive sense that glossiness and transparency may tend to be correlated and that glossy, transparent things tend to be smooth rather than rough, the factor loadings do not lead to a clear and decisive interpretation of the underlying psychological dimensions. In other words, the factor loadings probably depend heavily on the specific

(a) Factor Loadings

	PC 1	PC 2
Glossiness	0.4736	-0.2213
Transparency	0.5637	0.0406
Colourfulness	0.1361	0.2901
Roughness	-0.5101	-0.0283
Hardness	-0.1753	-0.6863
Coldness	0.2063	-0.3865
Fragility	0.2585	0.2264
Naturalness	-0.1976	0.3558
Prettiness	-0.0030	0.2573

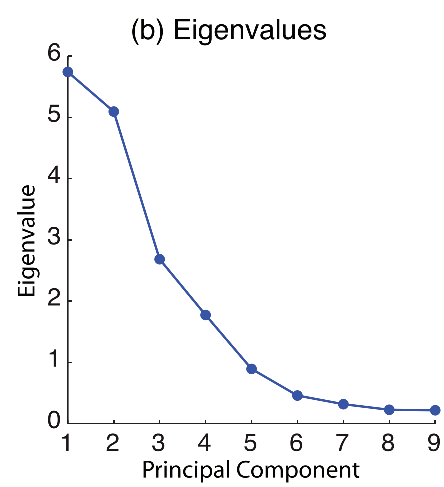


Figure 9. Results of PCA on the mean ratings across subjects. (a) Factor loadings for the first two PCs. (b) Variances of the PCs (i.e., eigenvalues of the covariance matrix).

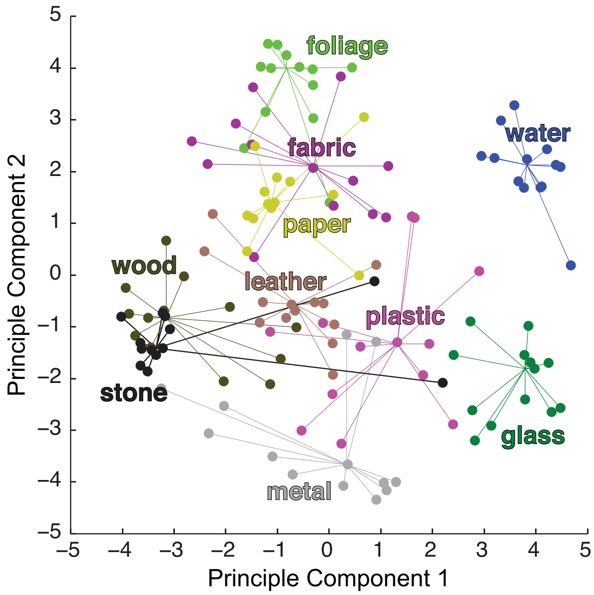


Figure 10. Distribution of the samples in the first two PCs. Circles represent projected positions of individual samples (13 images per class); lines join each sample to the projected mean location of each cluster. Color coding is based on true class membership, which was not told to the participants.

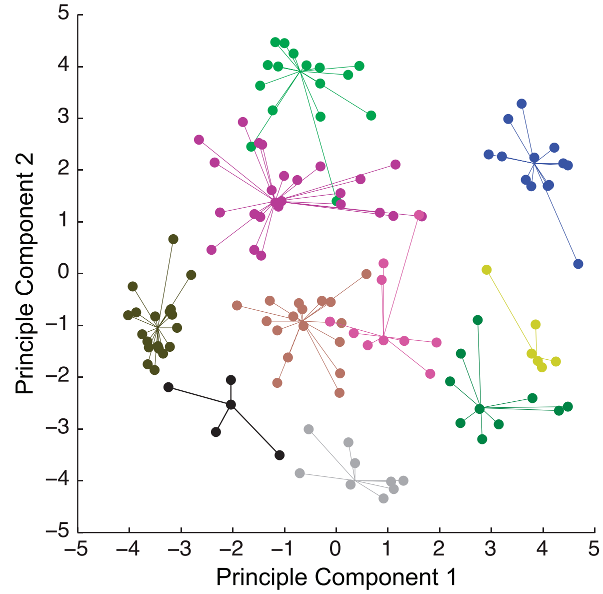


Figure 11. Clustering of the samples using a k-means clustering algorithm. Color coding is based on the nearest true cluster center (from Figure 10).

stimulus set used rather than revealing cardinal axes of the mental representation of materials.

In Figure 9b, we plot the eigenvalues of the PCs. Interestingly, there is a large drop of almost a factor of two between PC2 and PC3, and the first two PCs together account for 62% of the variance. The first five PCs account for 93% of the variance. This means that we can get an approximate impression of the overall distribution using just the first few PCs. In Figure 10, we plot ratings for each image projected onto the first two PCs and color code each image by its true class membership. As we have just argued, we suggest that the specific orientation of the distribution of images in the PCA space is not the most important aspect of the distribution. Instead, what is notable is the extent to which the samples are clustered and the proximity relationships between the different classes.

It is important to note that the visually apparent distances are not a perfect representation of the true distances in the space as the residual 38% of the variance in the distribution falls along the other seven dimensions. Nevertheless, it is striking how clearly the different samples are distributed within the space. Recall that the participants were not informed that the different samples belonged to 10 distinct material classes. Despite this, we see that the samples within each class are generally closely clustered in the space and distinct from the other clusters. The previously noted correlation between stone and wood shows up as a close proximity of the two classes. Water clearly stands out as perceptually distant from the other classes.

For comparison, we also plot the output of a k-means clustering algorithm (MATLAB function kmeans with 10 replications and the initial conditions specified by a preclustering), which was set to identify 10 clusters based on the spatial distribution of the samples in the full 9-D PCA space (Figure 11). We have color coded the samples based on the proximity of the cluster means to the cluster centers derived from the true class labels.

One way to evaluate the extent to which the participants' ratings of different material samples are naturally clustered is to compare the clusters returned by the k-means algorithm to the ground truth labels. We find that 90.13% of the samples have the same cluster members in common in the rating data as in the k-means analysis. In other words, only about 1 in 10 of the samples' cluster memberships cannot be fully accounted for by simple proximity relations in the 9-D perceptual feature space. This suggests that the different material classes are easily separated into distinct clusters based on the nine perceptual qualities considered here.

Together, these findings suggest that there is a close coupling between visually inferred perceptual qualities of materials and the visual categories to which the materials belong. In the following experiment, we test the extent to which the same is also true for semantic knowledge of materials accessed through verbal class labels. In comparison to the visual experiment, we asked subjects to rate a larger number of material qualities.

1. Rough	15. Fine	29. Hard
2. Smooth	16. Coarse	30. Soft
3. Transparent	17. Warm	31. Elastic
4. Opaque	18. Cold	32. Firm
5. Bendable	19. Chromatic	33. Flexible
6. Rigid	20. Achromatic	34. Inflexible
7. Granular	21. Matte	35. Oriented
8. Homogeneous	22. Glossy	36. Unoriented
9. Caliginous	23. Undirected	37. Circular
10. Clear	24. Directed	38. Line-like
11. Simple	25. Irregular	39. Multicolored
12. Complex	26. Regular	40. Uni-colored
13. High-contrast	27. Sharp-edged	41. Systematic
14. Low-contrast	28. Stubby	42. Random

Table 1. List of material properties used in the questionnaire. Adjectives were originally presented in German—here translated into English.

Experiment 2: Semantic ratings of material classes

Methods

Subjects

Sixty-five second-year psychology students from the University of Giessen took part in the survey. They received course credit for their participation. An additional 22 questionnaires were discarded due to missing values.

Material and procedure

Students were seated in a classroom of the university. A short written instruction was given prior to the questionnaire, stating that they would be given a list of adjectives describing various different appearance or surface qualities of materials. Subjects were told to rate six different material classes according to these adjectives on a six-value scale. The procedure was demonstrated with an example.

The questionnaire consisted of six stages, presented on separate sheets, each addressing one of the following material classes: wood, stone, metal, glass, plastic, and fabric. For each material, the same list of 42 adjectives was presented. Each adjective represented the opposite of one of the other adjectives (e.g., "warm" and "cold"). The order of the adjectives was randomized between the materials. A six-value scale was used to rate the extent to which each adjective applied to the given material. The lower end was labeled "trifft zu" (agree), and the upper end was labeled "trifft nicht zu" (disagree). Subjects were asked to rate each material according to the 42 adjectives on the six-value scale in the order that was given.

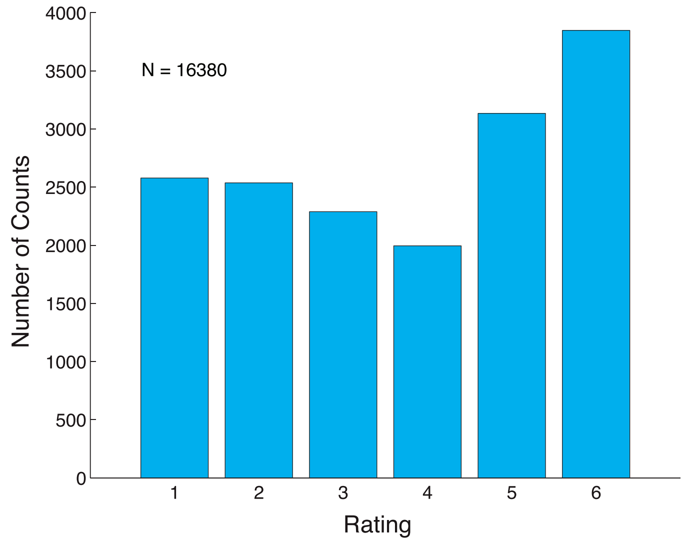


Figure 12. Distribution of rating values for Experiment 2.

Adjectives were collected based on extensive literature review (Picard, Dacremont, Valentin, & Giboreau, 2003; Rao & Lohse, 1996; Tamura, Mori, & Yamawaki, 1978) and on a small pretest. In the pretest, six subjects were asked to imagine a certain material and to write down all adjectives that came to mind regarding this material. Every subject completed this task for all six materials used in the questionnaire. Subjects were given as much time as they needed. Based on this, we selected a wide range of adjectives that can be used to describe materials visually as well as haptically (see Table 1).

Results

Response distribution

Figure 12 shows the overall distribution of responses across subjects and trials of Experiment 2. As in Experiment 1, the data pattern does not suggest a regression to the mean, implying again that the subjects could assign values clearly and decisively. Interestingly, they more often assigned high values ("disagree") than low values ("agree"). Given that we tested a large number of possible qualities, this is not entirely surprising as some qualities might be hard to imagine or only weakly relevant for the semantic representations of materials. The asymmetry of the rating distribution occurred despite the fact that adjectives were selected as opposing pairs. If all qualities were relevant to all materials, this would predict that a high rating for a given adjective should be matched by an equally low rating for its opposing counterpart. The subjects' willingness to disagree with both adjectives in a pair presumably reflects the fact that they considered those adjectives to be inapplicable to the given material.

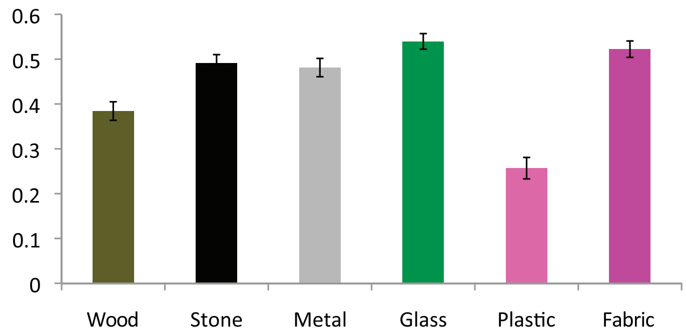


Figure 13. Mean intersubject correlation coefficients for the six material types. Error bars indicate standard errors of the mean.

Despite this, overall, this result is broadly consistent with the findings of Experiment 1.

Consistency across subjects

It is interesting to ask how consistent subjects were in rating the 42 different qualities for each material class. Unlike in Experiment 1—in which subjects saw identical images when making their ratings—in Experiment 2, each subject's concept or mental image of the material was determined entirely independently, presumably based on his or her previous experiences. Furthermore, subjects were asked to rate many more qualities than in Experiment 1, which means there are potentially more degrees of freedom along which the subjects could differ from one another. With this in mind, it seems reasonable to expect lower intersubject correlations than in Experiment 1. We summarize mean intersubject correlations for each material class in Figure 13.

Somewhat surprisingly, we find that the average degree of consistency across subjects was not dramatically lower than in Experiment 1. Mean intersubject correlations were higher than $r = 0.3$ for all material classes except for plastic. This suggests that subjects seem to have consistent mental representations of the six material classes we tested. The fact that plastic seems to yield systematically lower correlations is probably due to the fact that plastic is an artificial material, which can occur in a very diverse variety of forms (e.g., polythene shopping bags vs. polystyrene packing materials). While stone is typically hard and glass is mostly transparent, such general rules often do not hold for plastic, and this likely affects the variance across subjects. In particular, if each subject based his or her response on a mental image of a particular type of plastic, it makes sense that the underlying variance in the appearance of the material would be inherited in the subject's ratings.

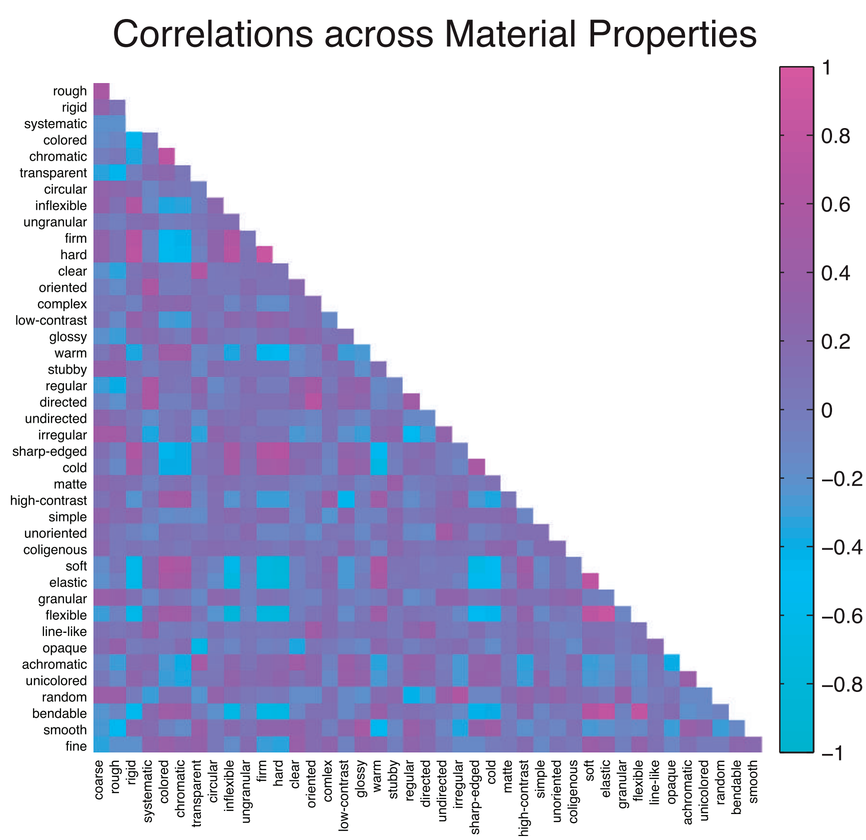
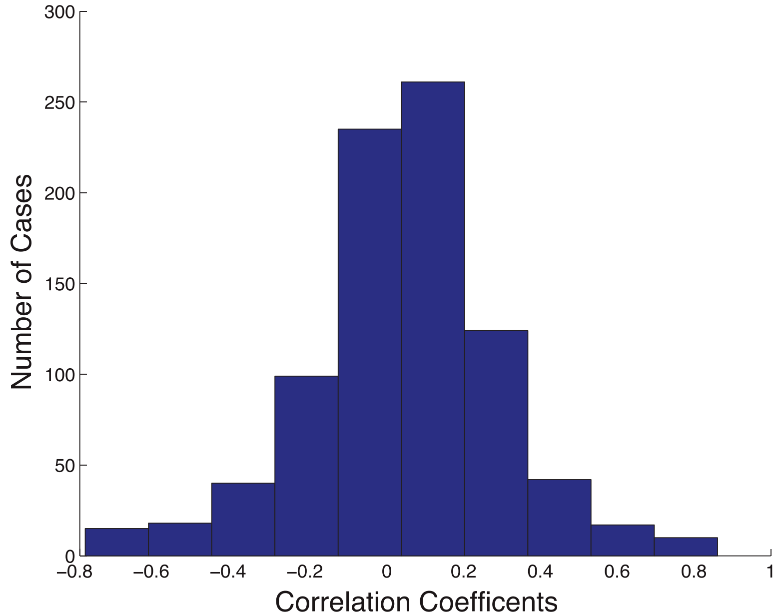


Figure 14. Correlations between properties averaged over the six materials.



largest negative correlations	correlations closest to zero	largest positive correlations
r = -0.79 **soft vs. hard**	*r = -0.0018* **glossy vs. chromatic**	*r = 0.86* **hard vs. firm**
r = -0.76 **flexible vs. inflexible**	*r = -0.0013* **undirected vs. low-contrast**	*r = 0.85* **elastic vs. flexible**
r = -0.76 **elastic vs. hard**	*r = -0.0009* **matte vs. achromatic**	*r = 0.81* **bendable vs. flexible**
r = -0.74 **hard vs. flexible**	*r = -0.0003* **glossy vs. high-contrast**	*r = 0.81* **bendable vs. elastic**
r = -0.74 **elastic vs. firm**	*r = 0.0004* **line-like vs. chromatic**	*r = 0.77* **soft vs. elastic**

Figure 15. Histogram of the correlation coefficients between material properties.

Correlations between qualities (adjectives)

As in Experiment 1, it is unclear to what extent the adjectives used to describe different material properties are truly independent of one another. Given that in Experiment 2 subjects rated many more properties, it is not unreasonable to expect some redundancies between the adjectives. To test this, we averaged material property ratings across all six materials and conducted a correlation analysis over the mean values. In Figure 14, we plot the correlation matrix between adjectives (numerical values are provided in the Supplementary material). In this figure, the $n = 42$ adjectives are organized in complementary pairs, so the k^{th} entry is semantically opposed to the $([n + 1] - k)^{\text{th}}$ entry. For example, the first and last adjectives are "coarse" and "fine," and the second and second-from-last are "rough" and "smooth," respectively. Thus, the diagonal emerging from the bottom left-hand corner contrasts opposing adjectives. As expected, these generally tend to be weakly or negatively correlated with each other as indicated by the predominance of lilac and cyan colors along this diagonal.

To summarize the overall distribution of interproperty correlations, we plot the histogram of correlation coefficients in Figure 15. Correlations ranged between $r = -0.79$ and $r = 0.86$ with a mean of 0.044 and variance of 0.029. As can be clearly seen, a large proportion of the adjectives seem to be only weakly correlated with one another (with 50% of the mass with an r score between −0.087 and +0.183). This suggests that many of the properties examined here could be judged more or less independently of each other, indicating that subjects can make nuanced distinctions between different attributes of materials. At the same time, a smaller number of adjective pairs were relatively strongly correlated, especially the adjectives describing hardness, softness, and malleability.

Correlations between material classes

In addition to correlations between different material properties, it is also interesting to ask to what extent the semantic ratings of different materials are correlated with one another. Figure 16 show the correlation matrix relating the ratings for the different materials to one another.

Overall, correlations ranged between $r = -0.21$ and $r = 0.41$. The highest correlations were found between glass and metal, $r = 0.41$; wood and stone, $r = 0.38$; and stone and metal, $r = 0.35$, and essentially no correlation

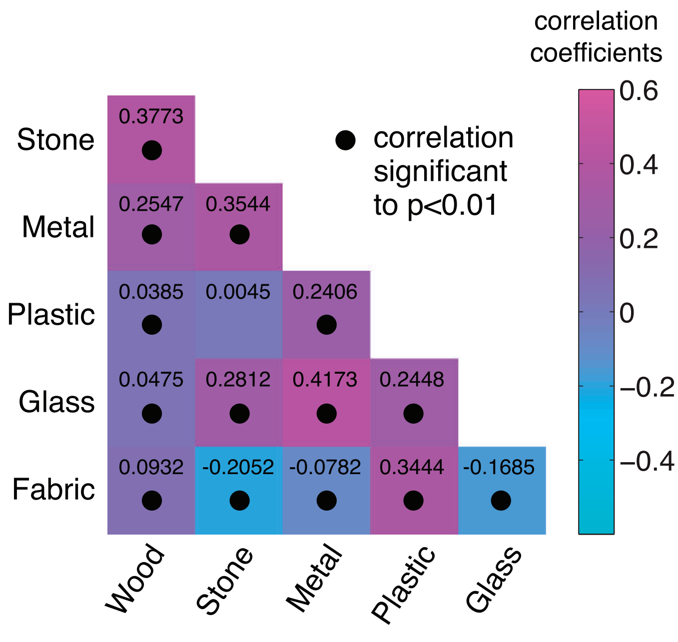


Figure 16. Correlation coefficients between materials.

at all was found between stone and plastic, $r = 0.005$. As before, these are broadly intuitive findings, which are quite similar to those obtained in Experiment 1, in which subjects rated single, unlabeled sample images. The similarities in the correlation structure between the two experiments suggest that semantic representations of entire classes of material are derived from—or at least substantially related to—sensory ratings attributed to individual members of the class. Alternatively, it could be that when asked to rate a class of materials based on a verbal label, subjects bring to mind (perhaps through mental imagery) one or a few representative samples and use these to assign their ratings. Either way, the consistency suggests that verbal and visual tasks access similar stored knowledge about material qualities.

Clustering of materials in the space of perceptual qualities

Given the relatively weak correlations between adjectives but the moderately large correlations between material classes, it is interesting to ask whether the ratings from Experiment 2 can be used to cluster material classes in a low-dimensional feature space. In order to assess this, we again conducted a PCA over all materials and subjects. Unlike in Experiment 1, in which multiple exemplars were presented for each class, in Experiment 2, each class was represented by only a single label. However, we have ratings from many more subjects, so we can consider how the ratings from different subjects are clustered for each material class.

In Figure 17, we plot the eigenvalues for the PCs. As before, there is a large drop in eigenvalues within the first few PCs and a clear "knee" at the third PC. More than 50% of the variance was represented by the first seven PCs. This again suggests that we can visualize the clustering of the different subjects' ratings of each material class in a low-dimensional subspace derived from the 42-dimensional ratings space.

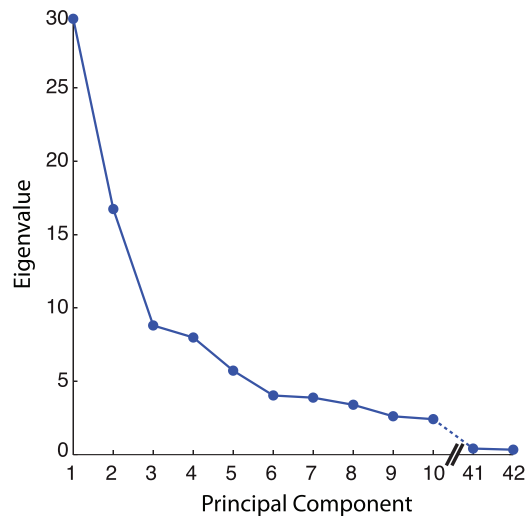


Figure 17. Eigenvalues for the first 10 and last two PCs from Experiment 2. Eigenvalues descend smoothly between PCs 10 and 41 and are therefore omitted from the plot to save space.

In Figure 18, we plot the distribution of ratings for each class in the space defined by the first two principal components. Each point indicates the ratings for a

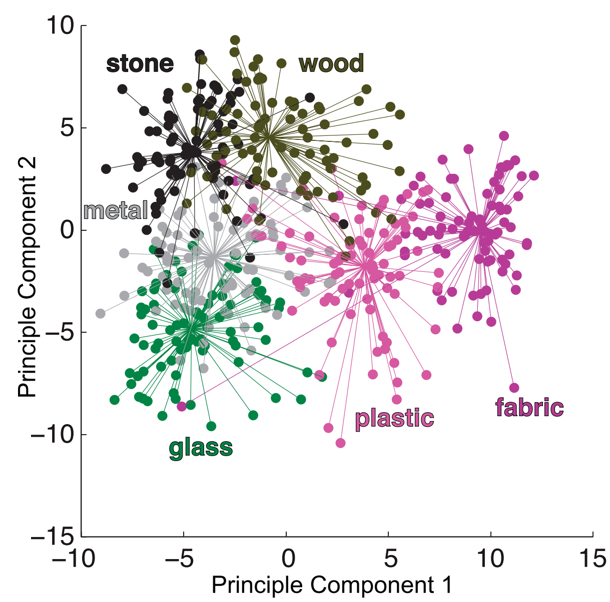


Figure 18. Distribution of ratings of each material class in the 2-D PCA space. Each point represents ratings for a given material class from a single observer. Note that the different observers are broadly consistent in the ratings they assign, leading to a clear clustering of the points belonging to each class.

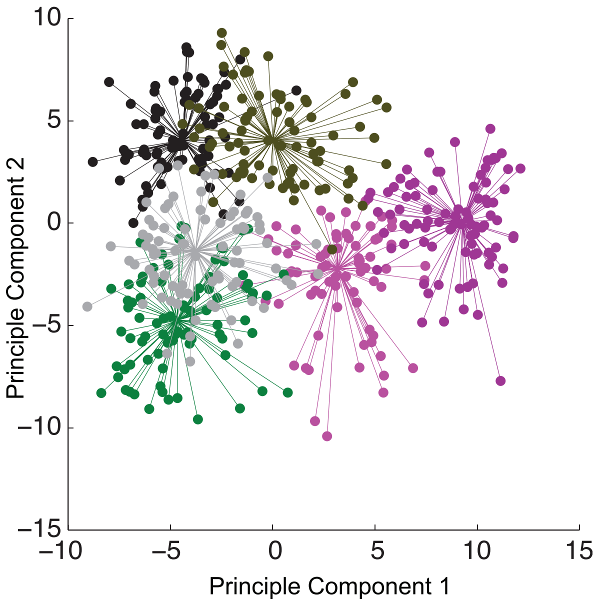


Figure 19. K-means clustering of the data from Experiment 2. Colors are assigned based on the nearest cluster centers from Figure 18.

single observer for a single material class as indicated by the color of the point. The data appear very clearly clustered in this space with dots of the same color appearing close to one another in distinct clouds. Although the clusters overlap to some extent (especially glass, metal, and plastic), they nevertheless appear clearly localized within the space. Thus, in most cases, the subjects' material property ratings lead to quite similar representations of these materials.

As before, it is interesting to ask whether the first few PCs have meaningful interpretations. To do this, we looked at the properties with the highest factor loadings on the two PCs. For the first PC, the highest positive factor loadings were found for the properties hard, unflexible, firm, rigid, and cold. By contrast, the highest negative factor loadings were found for the properties flexible, elastic, soft, bendable, colored, and warm (values ranged between −0.296 and 0.298). This interpretation makes broadly intuitive sense given the alignment of the materials within the space with stone, metal, and glass at one end of the continuum and fabric at the other. The second dimension is best reflected by the properties transparent, clear, regular, systematic, and achromatic on one end, and the other end can be described based on the adjectives irregular, rough, random, coarse, and opaque. Here, values ranged between −0.282 and 0.339. Again, the ordering of the materials along this dimension is broadly intuitive. Stone and wood naturally appear rather rough and coarse, show irregularities, and are opaque. By contrast, glass is usually highly transparent, regular, and not rough. The other three materials tend to be spread over the midrange: Fabric, for instance, can be very smooth, like a piece of silk, but also rough and coarse, like some forms of wool. Thus, broadly speaking, the factor loadings match our intuitive understanding of the properties of the tested classes. However, as in Experiment 1, we should be cautious about interpreting these dimensions as "cardinal axes" of the psychological space of materials. Rather, they more likely reflect the commonalities of the specific samples (or class labels) we tested.

As before, we can apply k-means clustering to assess the extent to which the ratings from different subjects are clustered. K-means clustering derives clusters solely on the proximity of different ratings in the 42-dimensional space. Thus, by comparing the true clusters to those extracted by k-means, we can measure the extent to which members of a given class are clumped together in the space. Figure 19 shows the results of the clustering algorithm with colors assigned based on the nearest true cluster labels. As can be seen, the k-means algorithm returns clusters that are rather similar to the ground truth. As before, we measured the degree of similarity between the true clusters and those returned by k-means based on the percentage of samples that have the same comembership. We find that 94.19% of the samples share the same set of other samples as class members.

Comparison of clustering between Experiments 1 and 2

Having clustered the data from both Experiments 1 and 2, it is interesting to ask to what extent the two distributions are consistent with one another. In Experiment 1, subjects rated material based on visual samples without explicit knowledge that the samples belonged to a limited set of classes whereas in Experiment 2 subjects were asked to describe classes as a whole, based solely on verbal labels. If there is a tight coupling between perceptual qualities and conceptual classes of materials, we might expect a similar embedding of the ratings in the feature space. In other words, we would expect the constellation of cluster centers to have roughly the same spatial configuration in the two spaces. To test this, we used Procrustes analysis to find the best linear transformation of the cluster configuration from Experiment 2 to match the data from Experiment 1. If the cluster configurations are similar, we expect relatively small transformations and a good fit. By contrast, if the configurations are very different, no transformation would provide a good fit. In Figure 20, we plot the PCA data from Experiment 2 mapped into the 2-D PCA space from Experiment 1. Specifically, for each of the six classes that were common to both experiments, we computed the mean position (i.e., cluster centers) of the data from all samples and subjects in the space defined by the first

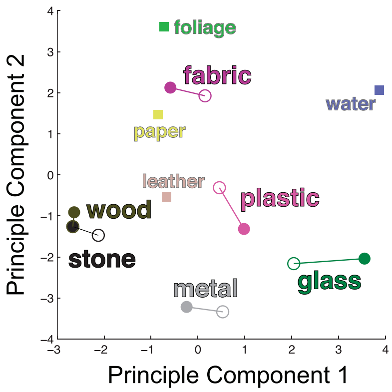


Figure 20. Comparison of cluster centers from Experiments 1 and 2 plotted in the PCA space from Experiment 1. *Filled disks*: cluster centers for each class from Experiment 1. *Open circles*: corresponding cluster centers from Experiment 2 transformed via Procrustes analysis into the PCA space from Experiment 1. *Squares*: cluster centers for classes from Experiment 1 that were not tested in Experiment 2.

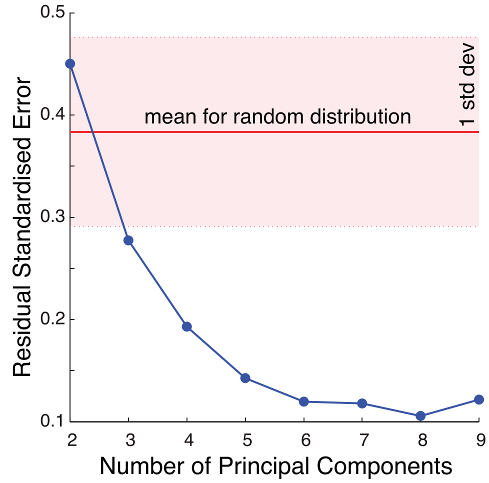


Figure 21. *Blue line*: residual error after fitting the cluster centers from Experiment 2 to those of Experiment 1 using Procrustes analysis as a function of the number of dimensions considered. *Red line and region*: mean residual error ±1 standard deviation for fitting random points distributed within the 9-D space to the data from Experiment 1. The mean and standard deviation were estimated from applying the fitting 1,000 times.

nine PCs from each dataset. We then applied Procustes analysis to map the cluster centers from Experiment 2 into the PCA space from Experiment 1.

Visually, the correspondence between the two configurations in the 2-D space is quite apparent although it is important to note that residual error along the other seven dimensions are not visualized. The residual standardized error for the fit in 9-D is 0.1217. For comparison, we generated random points uniformly distributed within the same range of distances in 9-D and measured the residual error from fitting these points to the cluster centers from Experiment 1. We repeated this procedure 1,000 times and found that the mean error was 0.3835 with a standard deviation of 0.0927. In Figure 21, we plot how the residual error for the true fits changes as a function of the number of dimensions considered. Because the distributions are expressed in PCA space, as expected, adding dimensions yields diminishing returns in terms of fit accuracy. Together, these findings suggest that, on average, visual and semantic representations of the distribution of material classes in the space of material properties are quite similar.

General discussion

We have shown that when presented with photographs of materials, subjects are able to make reliable, systematic judgments of the nine perceptual qualities that we tested. Furthermore, the different material classes are relatively well clustered within the 9-D space defined by the quality ratings, such that if one were given the ratings of the nine different qualities, one could with quite high accuracy determine to which of the 10 material classes the image belonged. Similarly, for semantic ratings of material classes based on verbal class labels, subjects are quite consistent in their assignment of different material qualities to the six tested classes. As with the visual stimuli, the ratings are well clustered in the 42-dimensional feature space, again implying it would be possible to estimate the class given only the ratings assigned to a material. Moreover, we found that the relative spatial locations of the different classes were highly similar in the two experiments. Together, this suggests a strong coupling between the visual estimation of material qualities and the mental representation of different material classes, at least for the classes and features we tested.

It is important to note that we cannot infer the causal directionality of the coupling between qualities and classes. In all likelihood, this relationship is bidirectional. In some cases, the visually inferred material qualities (e.g., glossiness, waviness, strong color, smoothness, etc.) help us to identify the class of a material (e.g., silk) based on its similarity to other members of the class. At the same time, in other cases, identifying that a material belongs to a certain class provides information about the material properties based on stored semantic knowledge (e.g., knowing that silk often feels cool to the touch, which cannot be seen directly). However, our findings do suggest that the two tasks—estimating material properties and assigning materials to mental classes—are intimately related.

Another important caveat is that we should be wary of interpreting the outcome of the PCA as revealing "primary" or "cardinal" dimensions of the space of materials. Although there may exist a mental "material space" (akin to the color spaces), there are grounds for questioning whether it has a fixed set of perceptual dimensions or even whether all materials can be embedded together in a single, monolithic space. In color space, the number of dimensions is based on the transduction mechanisms at the very front end of the vision, and the color space transformations that determine the cardinal axis of color spaces are based on very systematic recombinations of this information. By contrast, there are probably very many factors that contribute to a perceived material property, such as "fragility" or "prettiness," some of which may even be highly subjective. This makes defining cardinal dimensions slippery. Furthermore, while we can relatively easily comprehend similarities and differences between somewhat similar materials and classes (e.g., which is more similar to oak wood: ash or ebony?), it becomes very difficult to make judgments of the similarity between very different materials (e.g., which is more similar to bread: chrome or jade?). This suggests that there may be no single common metric for the "material space" or that large distances are difficult to estimate. Moreover, it seems plausible that the set of features that subjects use to compare materials may vary depending on what samples are to be compared. For example, when comparing a set of very similar materials (e.g., comparing one shampoo to several other shampoos), we may attend closely to subtle differences in appearance that are unimportant when comparing samples that are more different from one another (e.g., comparing shampoo to toothpaste and shaving foam). If the relative weights of different features change—and may be recombined—on the fly, depending on the particular tasks and comparisons to be performed, it seems somewhat doubtful whether materials can be embedded in a fixed "material space" with strict cardinal axes.

A third important limitation of the present study is that it provides little insight into the image features underlying the perception of different material qualities or classes. Despite their small number, the samples within each class are highly diverse in appearance, making it difficult to identify image features that are common to the samples. While some perceptual qualities (e.g., colorfulness) clearly correlate with relatively easily measured image properties (e.g., average color saturation), it seems intuitively less likely that we could identify simple low-level and mid-level image features that consistently predict higher-level—and sometimes more subjective—attributes, such as "prettiness," at least based on the small number of samples we consider here. An informal analysis of the principal components reveals negligible correlations with various statistics derived from color histograms and wavelet marginal distributions. Presumably, combinations of larger numbers of more sophisticated image measurements may prove more fruitful, but the problem of predicting subjective ratings from arbitrary photographs of materials remains extremely challenging.

Previous work (Sharan et al., 2009) has shown that subjects are surprisingly good at classifying materials given brief presentations and perform far better than current computational methods (Liu et al., 2010), which achieve about 45% correct performance with just 10 predefined material classes. Our findings extend these observations by demonstrating a strong connection between material qualities and categorization. The k-means clustering algorithm clustered over 90% of the samples the same way as humans did (i.e., the same mutual class membership) based on the human quality ratings alone. Along with the fact that we obtain similar proximity relations in the semantic task for which no images were presented, this suggests that the features humans use to represent mental classes of materials are not just constellations of low-level and mid-level image features but also of more abstract physical and functional attributes, such as "fragility," "flexibility," and "naturalness." Explaining how the brain is able to estimate (or recognize) such attributes from an image is clearly one of the most important outstanding challenges in the science of material perception.

One possibility is that the visual system estimates parameters of mental models of materials. In other words, given the image data, the visual system estimates material properties by "fitting parameters" of a statistical or physical model to the image data. In this scheme, the mental representation of materials would be like a "generative model," which describes or predicts the possible states or appearances of materials, somewhat akin to the internal models that are thought to underlie motor programs (e.g., Kawato, 1999). This would constitute a "deep" representation of material properties, allowing observers to, for example, imagine likely variations of a given material.

Another possibility is that the visual system does not fit parameters of a generative model but instead recognizes telltale combinations of lower-level features that are diagnostic of the material properties. This heuristic approach could work in two ways. For example, it could enable direct access to stored mental representations of specific materials, facilitating the recovery of stored knowledge about material properties in much the same way as rapid scene recognition does not require fitting a parametric representation of a scene to the image but enables the recovery of knowledge specific to the recognized scene class, such as that offices tend to contain desks (Oliva & Torralba, 2007). Alternatively, the material properties themselves might be "recognized" through combinations of lower-level features. Such approaches have been suggested for the representation of glossiness (Marlow, Kim, & Anderson, 2012; see also Fleming, 2012), viscosity (Fleming & Paulun, 2012), and other material properties. It remains to be seen the extent to which these two broad approaches to estimating material properties—mental models and diagnostic heuristics—can be unified.

Conclusion

Together, our results suggest that subjects are both quite good—and quite consistent—at assigning material qualities to different materials both visually and semantically. We can use ratings of different material properties to identify which class the materials belonged to even when the subjects were not explicitly informed about the classes. Furthermore, we find that the similarity of relationships between different classes are intuitively captured by their proximity to one another in the feature space defined by the subjects' ratings of different material qualities. This suggests that perceptual qualities and material classes are closely related. This is further supported by the fact that we find similar distributions of material classes in the visual and semantic domains, which suggests that perceptual and cognitive representations of material classes are intimately related. Thus, the visual estimation of material qualities and the separation of different material samples into distinct mental classes are two distinct but closely connected tasks.

Keywords: materials, surface perception, object recognition, clustering, image classification, texture perception

Supplementary material

Please see Supplemental Data. The supplementary information contains the complete list of all values presented in Figure 14 of the main text.

Acknowledgments

Supported by DFG grant GE 879/9 to KRG and the EU Marie Curie Initial Training Network "PRISM" (FP7-PEOPLE-2012-ITN, Grant Agreement: 316746) to RWF. The authors wish to thank Urs Kleinholdermann for help with data processing from the questionnaires.

Commercial relationships: none.
Corresponding author: Roland W. Fleming.
Email: roland.w.fleming@psychol.uni-giessen.de.
Address: Justus-Liebig-Universität Gießen, Gießen, Germany.

References

Bhushan, N., Rao, A. R., & Lohse, G. L. (1997). The texture lexicon: Understanding the categorization of visual texture terms and their relationship to texture images. *Cognitive Science: A Multidisciplinary Journal, 22*(2), 219–246.

Fleming, R. W. (2012). Human perception: Visual heuristics in the perception of glossiness. *Current Biology, 22*(20), R865–R866.

Fleming, R. W., & Paulun, V. (2012). Goop! On the visual perception of fluid viscosity. *Journal of Vision, 12*(9):949, http://www.journalofvision.org/content/12/9/949, doi:10.1167/12.9.949. [Abstract]

Heeger, D. J., & Bergen, J. R. (1995). Pyramid-based texture analysis/synthesis. In S. G. Mair & R. Cook (Eds.), *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques. SIGGRAPH '95* (pp. 229–238), New York: ACM. doi:10.1145/218380.218446.

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology, 9*(6), 718–727.

Liu, C., Sharan, L., Adelson, E. H., & Rosenholtz, R. (2010). Exploring features in a Bayesian framework for material recognition. In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* (pp. 239–246), doi:10.1109/CVPR.2010.5540207.

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

Matusik, W., Pfister, H., Brand, M., & McMillan, L. (2003). A data-driven reflectance model. *ACM Transations on Graphics*, *22*(3), 759–769, doi:10.1145/1201775.882343.

Matusik, W., Zwicker, M., & Durand, F. (2005). Texture design using a simplicial complex of morphable textures. *ACM Transactions on Graphics, 24*(3), 787–794, doi:10.1145/1073204.1073262, http://doi.acm.org/10.1145/1073204.1073262.

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences, 11*(12), 520–527.

Oren, M., & Nayar, S. K. (1994). Generalization of Lambert's reflectance model. In *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '94)* (pp. 239–246). New York: ACM. doi:10.1145/192161.192213, http://doi.acm.org/10.1145/192161.192213.

Picard, D., Dacremont, C., Valentin, D., & Giboreau, A. (2003). Perceptual dimensions of tactile textures. *Acta Psychologica, 114*(2), 165–184.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision, 40*(1), 49–71.

Rao, A. R., & Lohse, G. L. (1996). Towards a texture naming system: Identifying relevant dimensions of texture. *Vision Research, 36*(11), 1649–1669.

Ray, N., Levy, B., Wang, H., Turk, G., & Vallet, B. (2009). Material space texturing. *Computer Graphics Forum, 28*(6), 1659–1669.

Ruiters, R., Schnabel, R., & Klein, R. (2010). Patch-based texture interpolation. *Computer Graphics Forum (Proc. of EGSR), 29*(4), 1421–1429.

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

Tamura, H., Mori, S., & Yamawaki, T. (1978). Textural features corresponding to visual properties. *IEEE Transactions on Systems, Man, and Cybernetics, 8*(6), 460–473.

Torrance, K. E., & Sparrow, E. M. (1967). Theory for off-specular reflection from roughened surfaces. *Journal of the Optical Society of America, 57*(9), 1105–1114.

Ward, G. J. (1992). Measuring and modeling anisotropic reflection. *ACM SIGGRAPH Computer Graphics, 26*(2), 265–272.