Visual Perception of Material Properties

Dissertation zur Erlangung des Doktorgrades der Naturwissenschaften der

Justus-Liebig-Universität Gießen
Fachbereich 06 – Psychologie & Sportwissenschaften

Vorgelegt von Jacob R. Cheeseman
am 23.07.2024

Erstbetreuer:
Prof. Dr. Roland W. Fleming (Justus-Liebig-Universität Gießen)
Zweitbetreuer:
Prof. Dr. Karl R. Gegenfurtner (Justus-Liebig-Universität Gießen)

ACKNOWLEDGMENTS

I extend my heartfelt gratitude to my supervisor, Roland Fleming, for his patience, guidance, and overall support throughout this journey. I am also grateful to my second supervisor, Karl Gegenfurtner, for agreeing to meet with me at VSS all these years ago and connecting me with Roland.

My sincere thanks go to the past and present members of Allgemeine Psychologie for being a community of such brilliant and lovely people. I am especially thankful to Rob Ennis, Filipp Schmidt, Thomas Schmidt, Yaniv Morgenstern, Arash Akbarinia, Thorsten Hansen, Guillermo Aguilar, and Jim Ferwerda for their assistance and advice at various stages of this research. Special thanks to Saskia Honnefeller, Britta Fritz, and Jasmin Kleis for their help with data collection; Wendy Adams, Frank Maile, and Jingyang Zhou for being such gracious hosts; my students for inspiring and challenging me; and my friends and family for their support and encouragement.

ABSTRACT

Visual experience allows us to make instant, context-dependent judgments about material properties of surfaces, aiding in recognizing and categorizing materials based on their functional utility. This thesis explores gloss perception and material recognition through investigations into how visual sensitivity to surface properties like specular reflectance varies under different conditions, and how our ability to discriminate and categorize materials is influenced by bottom-up and top-down visual processes.

The first study examines the scaling and discriminability of perceived gloss, focusing on how gloss sensitivity changes with the magnitude of specular reflectance and the role of internal sensory noise. We find that a model based on Maximum Likelihood Difference Scaling (MLDS), a suprathreshold perceptual scaling method, can efficiently and accurately estimate just-noticeable differences in specular reflectance, indicating that human gloss sensitivity operates with the same additive internal noise assumed by the model.

The second study proposes a novel framework for measuring gloss sensitivity, assessing how observer rankings of gloss differences can inform the development of a measurement standard for gloss appearance in diverse contexts. We find that observers consistently rank gloss differences resulting from variations in lighting, shape and viewpoint, while the physical reflectance (in this case, roughness) of the surface is held constant. An image-computable model of visible differences (HDR-VDP-3) accurately predicts observer rankings of suprathreshold gloss differences, which can establish reasonable bounds on gloss sensitivity across these viewing conditions.

The third study demonstrates that, for certain images in which the material identity of surfaces is ambiguous (e.g., a field of wheat vs. a patch of carpet fabric), observers with different assumptions about apparent distance assign different material categories to describe the surfaces. This material-scale ambiguity, a previously undocumented phenomenon in visual perception, has implications for theories of material recognition and categorical perception in general.

Together, these studies suggest that the visual perception of material properties relies on estimating image features diagnostic of material category under typical viewing conditions. Under ambiguous viewing conditions, however, these image features are more open to interpretation than previously thought.

CONTENTS

CHAPTER 1: INTRODUCTION 1

1.1 Motivation 1

1.2 Psychophysics and computational analysis 2

1.3 Material properties and ambiguity in perception 4

CHAPTER 2: SCALING AND DISCRIMINABILITY OF PERCEIVED GLOSS 5

2.1 Introduction 5

2.2 Experiment 1: Establishing a perceptual scale for surface reflectance 7

2.2.1 Participants 7

2.2.2 Stimuli 7

2.2.3 Procedure 10

2.2.4 Results 11

2.3 Experiment 2: Measuring discriminability on a perceptual scale 12

2.3.1 Participants 12

2.3.2 Stimuli 12

2.3.3 Procedure 14

2.3.4 Results 15

2.4 Discussion 18

2.5 Conclusions 20

CHAPTER 3: GLOSS DISCRIMINATION: TOWARDS AN IMAGE-BASED PERCEPTUAL MODEL 21

3.1 Introduction 22

3.2 Experiment 1: Predicting apparent gloss differences across viewing conditions 26

3.2.1 Participants 26

3.2.2 Stimuli 26

3.2.3 Procedure 28

3.2.4 Image metrics 29

3.2.5 Results 30

3.3 Experiment 2: Image metric validation in a lab-based control experiment 33

3.3.1 Participants 33

3.3.2 Stimuli 34

3.3.3 Procedure 36

3.3.4 Results 37

3.4 Discussion 40
3.4.1 Towards ‘reasonable bounds’ on JNDs for surface reflectance 41
3.4.2 Limitations and future directions 43
3.5 Conclusions 45

CHAPTER 4: SCALE AMBIGUITIES IN MATERIAL RECOGNITION 46
4.1 Introduction 47
4.2 Experiment 1: Unbiased judgements of distance and material 49
4.2.1 Participants 49
4.2.2 Stimuli 50
4.2.3 Procedure 52
4.2.4 Results 52
4.3 Experiment 2: Biased judgements of distance and material 57
4.3.1 Participants 57
4.3.2 Stimuli 57
4.3.3 Procedure 58
4.3.4 Results 58
4.4 Discussion 60
4.4.1 Image cues and material-scale ambiguity 61
4.4.2 Limitations and future directions 62
4.5 Conclusions 64

CHAPTER 5: GENERAL DISCUSSION 65
5.1 Gloss and material perception 65
5.2 Signal and noise 67
5.3 Conclusions 71

REFERENCES 73

APPENDIX 91
Supplementary material for Chapter 4 91
List of publications 103
Declaration 104

CHAPTER 1: INTRODUCTION

1.1 Motivation

Gloss perception plays a crucial role in both industrial applications and fundamental vision science. In industry, precise assessment of gloss is important for quality control in various manufacturing processes, such as automotive, coatings, and packaging (Linke & Das, 2016; Wu et al., 2016). In vision science, gloss perception reflects the inner workings of human visual processing, including how the brain processes and interprets complex visual information related to material properties (Fleming, 2017). The challenges in defining and measuring gloss stem from the multidimensional nature of gloss perception, including its dependence on illumination, surface properties, and individual differences between observers. Unlike color, which has relatively well-defined physical and perceptual dimensions (Fairchild, 2005), there is currently no standardized method for measuring and communicating gloss appearance, making it more difficult to compare gloss across different contexts and applications.

The central challenge in creating a metric for gloss perception is to account for the interactions between physical properties, visual estimation, and subjective interpretation. The Bidirectional Reflectance Distribution Function (BRDF) offers a comprehensive physical description of how light reflects from surfaces (Nicodemus et al., 1977), enabling realistic depiction of gloss and other material properties using computer graphics techniques (Ngan et al., 2005). While the BRDF describes the physical signal (reflected light), it does not describe how this signal is perceived. In other words, the human visual system interprets light and reflectance in a manner that is not strictly bound to the physical accuracy of these models. Instead, it relies on a complex set of perceptual cues that can differ significantly from the assumptions of the models. This suggests that, despite the vast range of potential (measured or analytical) BRDFs, none can be descriptions of *material appearance*, because this involves how light interacts with the human visual system.

The visual estimation process in gloss perception is crucial for making discriminations between materials with similar surface properties. For example, when selecting between two samples of varnished wood, we can discern slight surface differences that index their aesthetic quality. However, gloss perception does not equate to estimating a specific value of reflectance from a few different angles, as if measured by a ‘glossmeter’. Instead, it involves neural

computations that extract visual cues from retinal image information, and these cues are correlated with different physical parameters; for example, lighting environment (Adams et al., 2018; Fleming et al., 2003; Ged et al., 2020; Ho et al., 2006; Morimoto et al., 2023b; Motoyoshi & Matoba, 2012; Olkkonen & Brainard, 2011; Pont & te Pas, 2006; Wendt & Faul, 2017), shape (Berzhanskaya et al., 2005; Marlow & Anderson, 2024; Morimoto et al., 2023b; Nishida & Shinya, 1998; Olkkonen & Brainard, 2011; Tiedemann, 2018), viewpoint (Ho et al., 2007), and the presence of dynamic motion or Fresnel effects (Doerschner et al., 2011; Faul, 2019, 2021; Ferwerda & Padhye, 2021; Shiwen et al., 2023; Wendt et al., 2010; Wendt & Faul, 2018).

Understanding these multidimensional neural computations is complicated by the inherently subjective nature of gloss perception. For instance, early works on gloss perception aimed at developing measurement standards focused more on subjective dimensions rather than on understanding visual computations (Harrison, 1945; Hunter, 1937). In particular, Hunter's six dimensions of gloss (specular, sheen, contrast, haze, image distinctness, and surface texture) still dominate how gloss is measured and communicated today (for a thorough review of this topic, see Chadwick & Kentridge, 2015). Recent research has focused on measuring gloss constancy across changes in illumination, shape, and body color (Fleming et al., 2003; Marlow & Anderson, 2013; Morimoto et al., 2023b). For example, the orientation congruence between highlights and shading patterns, or the relationship between intensity gradients and surface normals, can influence the perception of glossiness. However, purely photometric measurements do not necessarily distinguish the causal origin of surface features. This limitation points to a need for further research that combines psychophysics with computational analysis to establish boundary conditions for measuring sensitivity to dimensions of gloss.

1.2 Psychophysics and computational analysis

A psychophysical image analysis that estimates Just Noticeable Differences (JNDs) in gloss perception is needed to understand how well – and under what specific viewing conditions – observers can discriminate between surfaces with subtle differences in reflectance properties. Previous studies (e.g., Ferwerda et al., 2001; Pellacini et al., 2000) have focused on judgments of suprathreshold appearance differences under symmetric viewing conditions, in which surfaces differ in only one variable, such as specular reflectance. These studies have shown how perceived surface reflectance varies with physical surface reflectance, and other factors, such as lighting and

shape. However, understanding how finer differences in gloss are perceived, particularly at near-threshold levels, is crucial for practical applications (e.g., industrial quality control). As described above, surface gloss discrimination involves estimation of local image features, such as specular highlights. Observers may also employ different strategies when evaluating different dimensions of gloss appearance, focusing on image information that correlates with the dimension of interest (Hunter, 1937; Kildau, 2016; Toscani et al., 2020). A fundamental question, therefore, is how these suprathreshold judgments of surface gloss correlate with near-threshold discrimination of specular reflectance, and how sensitivity to gloss varies with the magnitude of specular reflectance.

It has been suggested that near-threshold image differences can be used to predict suprathreshold differences in complex attributes like overall image quality (Mantiuk et al., 2011). However, it is less clear whether this approach can extend to the domain of material appearance. The experiments of Ramanarayanan et al. (2007) demonstrated that two images can be noticeably different, yet depict surfaces that appear to be made of the same material, suggesting that just-noticeable changes in surface reflectance are insufficient for judging overall material similarity. Moreover, while methods like Maximum Likelihood Difference Scaling are effective for assessing image similarity (Charrier et al., 2012), they may not be valid for estimating the discriminability of specific features like specular highlights (Protonotarios et al., 2016). Yet, other experiments on the watercolor effect (Devinck & Knoblauch, 2012a) and visual contrast (Kingdom, 2016) have found that discrimination performance is well-predicted by suprathreshold scaling, which indicates that these methods offer an alternative way to accurately measure discriminability. This is significant because the method avoids the need for laborious experiments typically required for measuring discrimination thresholds.

Chapter 2 of this thesis focuses on measuring JNDs along a single dimension of gloss (i.e., specular reflectance). This is particularly important, given the challenges in defining JNDs for a distal stimulus property like gloss. That is, visual sensitivity to relatively simple image features (e.g., contrast, edges) is more directly related to proximal stimulation of the retina, but for gloss, a distal surface property that is inherently influenced by diverse viewing conditions, the measurement of sensitivity becomes more complex. For example, two objects with the same intrinsic surface reflectance can have markedly different levels of discriminability based on variations in extrinsic factors such as lighting, viewpoint, and shape. Chapter 3 of this thesis aims to investigate gloss perception across these viewing conditions, establishing a framework for

estimating reasonable bounds on gloss sensitivity. To this end, our experiments tested HDR-VDP-3 (Chapiro et al., 2024; Mantiuk et al., 2023), a popular metric designed to predict the visibility of image differences, both in uncontrolled online settings and in optimally controlled laboratory conditions.

1.3 Material properties and ambiguity in perception

The ability to reliably and efficiently recognize materials is a fundamental aspect of visual perception. However, this process can become complicated under certain viewing conditions where different materials produce similar image features that are scale-dependent, leading to perceptual ambiguities. This phenomenon suggests that our perception and categorization of materials are influenced not just by their intrinsic reflectance properties, but also by the context in which they are viewed, and the subjective interpretation of the observer (e.g., assumed viewing distance). This ambiguity is rooted in the multiscale nature of material appearance and how different materials interact with light at various spatial scales (Pont & Koenderink, 2002, 2005). For instance, the appearance of water droplets or scratches on brushed metal can vary dramatically depending on the viewing distance. Such changes in appearance with distance can sometimes deceive observers about the true material properties, leading to potential mis-categorization. Understanding these scale-appearance dependencies is critical for developing more accurate models of material perception and for applications in fields like computer graphics, where the realistic rendering of materials is essential.

Chapter 4 of this thesis is about how contextual information, particularly viewing distance, influences material recognition. The investigation utilizes a set of photographs that allow multiple interpretations, some of which depend on the assumed distance from the camera to surfaces. If conflicting assumptions about viewing distance can cause identical images to be recognized as completely different materials, this challenges the theoretical proposition of a straightforward mapping between image features and material categories (Bell et al., 2015; Fleming et al., 2013; Sharan et al., 2013).

CHAPTER 2: SCALING AND DISCRIMINABILITY OF PERCEIVED GLOSS

A similar version of this chapter has been published as:

Cheeseman, J. R., Ferwerda, J. A., Maile, F. J., & Fleming, R. W. (2021). Scaling and discriminability of perceived gloss. Journal of the Optical Society of America A, 38(2), 203–210. https://doi.org/10.1364/JOSAA.409454

While much attention has been given to understanding biases in gloss perception (e.g., changes in perceived reflectance as a function of lighting, shape, viewpoint and other factors), here we investigated sensitivity to changes in surface reflectance. We tested how visual sensitivity to differences in specular reflectance varies as a function of the magnitude of specular reflectance. Stimuli consisted of renderings of glossy objects under natural illumination. Using Maximum Likelihood Difference Scaling (MLDS), we created a perceptual scaling of the specular reflectance parameter of the Ward reflectance model. Then, using the Method of Constant Stimuli and a standard 2AFC procedure, we obtained psychometric functions for gloss discrimination across a range of reflectance values derived from the perceptual scale. Both methods demonstrate that discriminability is significantly diminished at high levels of specular reflectance, thus indicating that gloss sensitivity depends on the magnitude of change in the image produced by different reflectance values. Taken together, these experiments also suggest that internal sensory noise remains constant for suprathreshold and near-threshold intervals of specular reflectance, which supports the use of MLDS as a highly efficient method for evaluating gloss sensitivity.

2.1 Introduction

The perception of real and virtual surface gloss has been investigated with a variety of experimental and analytical techniques, including the Method of Paired Comparisons (Marlow et al., 2012), Multidimensional Scaling (Pellacini et al., 2000), Maximum Likelihood Difference Scaling (Obein et al., 2004) and Maximum Likelihood Conjoint Measurement (Ho et al., 2008). These studies have focused on judgments of suprathreshold appearance differences and/or asymmetric viewing conditions to test how perceived surface reflectance varies as a function of physical surface reflectance, and other factors such as lighting and shape. Yet, for many practical purposes it is important to know not only which reflectance a given surface appears to have, but also how well observers can discriminate between surfaces that differ only in their intrinsic

reflectance properties. Surface gloss discrimination is believed to involve fine-scale examination of local image features, such as specular highlights (Phillips et al., 2010). However, it is also known that observers may adopt different strategies when tasked to evaluate the "gloss" of a surface, which consists of multiple appearance dimensions (Hunter & Harold, 1987; Leloup et al., 2012). To what extent do suprathreshold judgments of surface gloss predict near-threshold discrimination of specular reflectance? How does sensitivity to gloss vary as a function of the magnitude of specular reflectance?

Mantiuk, Kim, Rempel, and Heidrich (Mantiuk et al., 2011) demonstrated that near-threshold image differences can predict suprathreshold differences of complex attributes such as overall image quality. However, it remains unclear whether this also applies in the domain of material appearance. Given that two images can depict surfaces that appear to be made of the same material despite visible differences (Ramanarayanan et al., 2007), just-noticeable changes in surface reflectance may not be relevant for judging the overall similarity of material properties such as gloss. Similarly, while suprathreshold perceptual scaling is well-suited to assessing image similarity (Charrier et al., 2012), such methods are not necessarily valid for estimating the discriminability of local image features, such as specular highlights (Protonotarios et al., 2016). Indeed, it is possible that suprathreshold and near-threshold judgments evoke non-trivial differences in sensory representation. For example, Maximum Likelihood Difference Scaling (MLDS) is a popular suprathreshold perceptual scaling method in which sensory representations are modeled as independent, Gaussian random variables with equal variance. If this internal sensory noise were multiplicative rather than additive, this would not be evident from the shape of the perceptual scale produced by MLDS (Kingdom & Prins, 2016; Maloney & Yang, 2003), but this could affect discriminability estimates derived from the same perceptual scale (Aguilar et al., 2017). When scaling does predict discrimination, however, performance in both tasks can be modeled under the assumption that suprathreshold and near-threshold judgments share a common transducer function, thus indicating how internal sensory noise grows with stimulus magnitude (Kingdom, 2016).

The following experiments were designed to determine whether suprathreshold scaling can predict just-noticeable differences in surface reflectance. We find, similar to previous studies which directly compared judgments of near-threshold and suprathreshold appearance differences in the watercolor effect (Devinck & Knoblauch, 2012a) and visual contrast (Kingdom, 2016), that

discrimination performance is well-predicted by suprathreshold scaling. Perhaps most notably, our study furnishes evidence that internal sensory noise remains constant for suprathreshold and near-threshold intervals of specular reflectance, which supports the use of MLDS as a highly efficient method for evaluating sensitivity without participants having to perform tedious discrimination experiments. These findings have potentially important implications for future studies of material appearance across the fields of industrial manufacturing, computer graphics, and vision science.

2.2 Experiment 1: Establishing a perceptual scale for surface reflectance

We first sought to construct and verify a perceptual scale for surface gloss. Our approach was similar to that taken by Pellacini, Ferwerda, and Greenberg (2000), who applied Multidimensional Scaling (MDS) to judgments of computer-simulated glossy spheres under artificial illumination. With this data they constructed a perceptually-scaled gloss space consisting of two dimensions (contrast and distinctness of the reflected image), which they later used to derive just-noticeable differences (JNDs) in gloss (Ferwerda et al., 2001). Here we employed Maximum Likelihood Difference Scaling (MLDS; see Maloney & Yang, 2003), more naturalistic (although still computer-generated) stimuli, and we varied only the specular reflectance of the target object, while all other scene variables were fixed.

2.2.1 Participants

Ten adults (5 males and 5 females; age range: 19 to 40 years; M = 24 years, SD = 6.2 years) with normal or corrected-to-normal visual acuity participated in the experiment and were paid 8€ per hour. All participants provided informed consent prior to the following experiments, which were approved by the ethics review board at Justus Liebig University Giessen and conducted in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki).

2.2.2 Stimuli

Seven stimulus images were created with the Mitsuba v0.5 physically-based renderer (Jakob, 2010). The rendered scene (see Figure 1) consisted of a central target object (a laser-scanned 3D model of a bell pepper; see Norman & Phillips, 2016), on a marble-textured pedestal with four golf balls positioned in the foreground, all lit by a high dynamic range illumination map of an outdoor scene (Adams et al., 2016). Global illumination calculations were performed using photon

mapping (Jensen, 1996), with 16 samples per pixel, and two-bounce interreflections. Surface reflectance properties were represented using the Ward-Dur light reflection model (Geisler-Moroder & Dür, 2010). The model has three parameters that specify the specular reflectance (ρ_s), diffuse reflectance (ρ_d), and microscale roughness (α) of a surface. To produce the seven stimulus images, the specular reflectance of the target object was varied in seven equal steps (ρ_s = {0.017, 0.031, 0.044, 0.058, 0.072, 0.085, 0.099}; this matches the range of values used by Pellacini, Ferwerda, and Greenberg (2000). As in previous studies of gloss perception, a dark green diffuse color (RGBρ_d = {0.1, 0.3, 0.1} and low surface roughness (α = 0.04) were used to ensure that surfaces had visible specular highlights (see Fleming et al., 2003; Pellacini et al., 2000). In previous studies where participants were required to estimate lighting conditions (Xia et al., 2014), golf balls have been used as probe objects; therefore we included these objects in our scene to provide supplementary information about scene lighting, and to anchor judgments about surface reflectance properties (e.g., Gilchrist et al., 1999). Each golf ball had high specular reflectance (ρ_s = 0.099), achromatic diffuse reflectance (ρ_d = 0.9, 0.45, 0.225, 0.113 left-to-right), and low surface roughness (α = 0.04). The matte gray pedestal object had reflectance parameters (ρ_d = 0.5, ρ_s = 0.00) that were modulated by a marble-patterned texture map. The stimulus parameters are summarized in Table 1. The 720 × 720 rendered images were converted to the sRGB color space, tone mapped using the method described in Reinhard, Stark, Shirley, and Ferwerda (2002) with parameter values (key = 0.18; burn = 0; gamma = 2.0), and stored in the PNG image format. The complete set of stimulus images for Experiment 1 is contained in Dataset 1 (see Cheeseman et al., 2020).

The images were displayed on an Eizo ColorEdge CG277 LCD monitor (27” diagonal; 2560 × 1440 resolution). At a viewing distance of 50 cm, each image subtended approximately 19 degrees of visual angle. The display was calibrated to have an sRGB color gamut, 80 cd/m2 D65 white point, and a gamma of 2.0. With these settings, changes in the specular reflectance of the target object produced proportional changes in displayed image luminance. Of particular interest for the purposes of this study is the specular contrast of the target object, which is the increase in the contrast of the image of the target object above the base contrast in the image of a diffusely reflecting target object. The base contrast in the displayed image of the target object was (61%) and the specular contrasts produced by our chosen ρ_s values were (9%, 14%, 17%, 20%, 21%, 23%, 24%). The displayed stimulus image luminances, chromaticities, and contrasts are

summarized in Table 1. The display was viewed in a dark room, and the images were presented against a uniform middle-gray background.

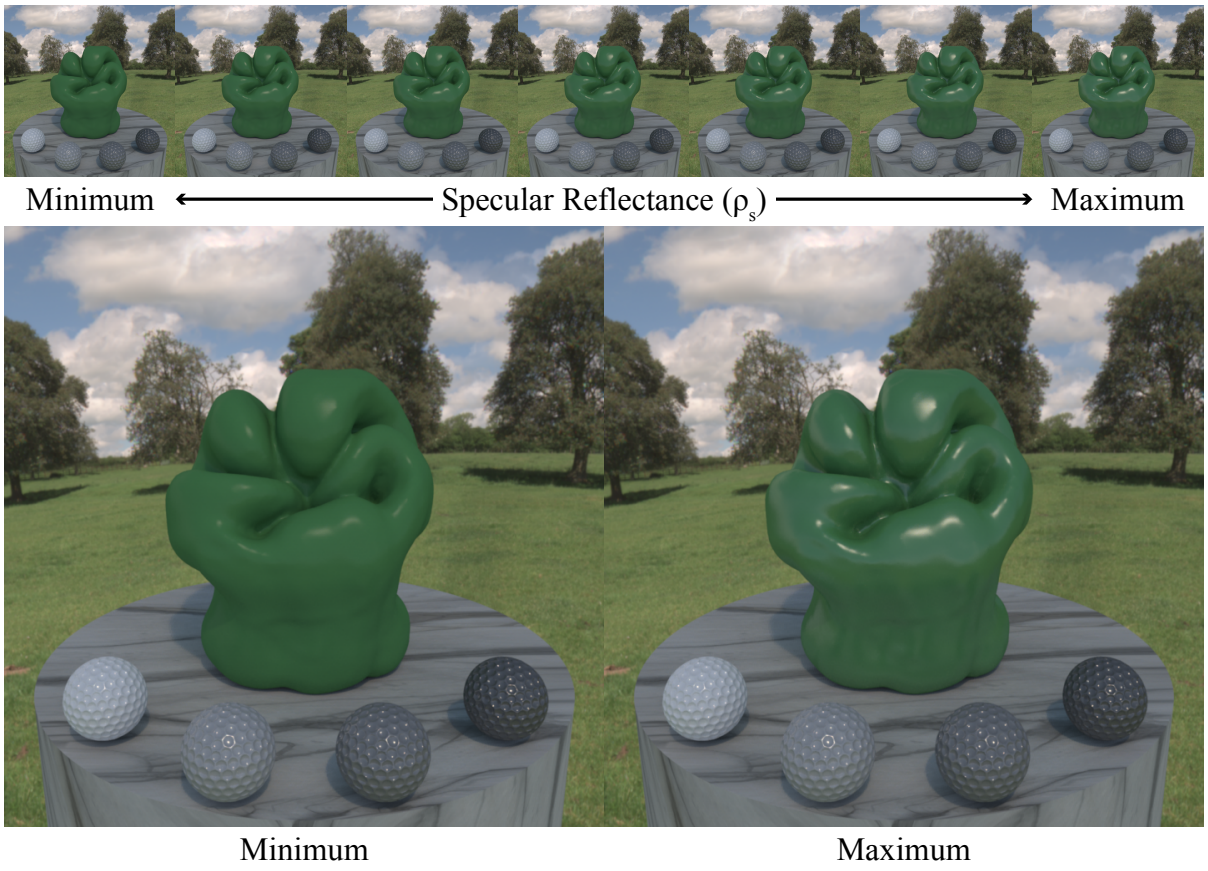

Figure 1. Stimulus images shown in Experiment 1. The specular reflectance of the green target object is varied in seven equal steps from low ($\rho_s = 0.017$) to high ($\rho_s = 0.099$). The scene consists of a 3D model of a bell pepper (Capsicum annuum) seated on a marble pedestal under natural illumination. Golf balls are arrayed in the foreground to provide information about the illumination field.

Table 1. Stimulus Properties

Target object (pepper)									
	Surface reflectance properties			Chromaticity		Displayed image properties			
Obj. ID	ρ_d	ρ_s	α			Min. lum.	Max. lum.	Diff. cont.	Spec. cont.
	(0.1,0.3,0.1)	0.000	0.04	0.2979	0.4370	1.62	6.61	0.61	0.00
1	"	0.017	"	0.2978	0.4309	1.66	9.39	"	0.09
2	"	0.031	"	0.2977	0.4267	1.70	11.79	"	0.14
3	"	0.044	"	0.2977	0.4232	1.75	13.99	"	0.17
4	"	0.058	"	0.2976	0.4197	1.80	16.32	"	0.20
5	"	0.072	"	0.2975	0.4166	1.85	18.60	"	0.21
6	"	0.085	"	0.2974	0.4138	1.89	20.68	"	0.23
7	"	0.099	"	0.2973	0.4111	1.93	22.90	"	0.24

Anchor objects (golf balls)									
	Surface reflectance properties			Chromaticity		Displayed image properties			
Obj. ID	ρ_d	ρ_s	α			Min. lum.	Max. lum.	Diff. cont.	Spec. cont.
1	(0.9,0.9,0.9)	0.099	0.04	0.2942	0.3121	10.06	35.48	0.42	0.14
2	(0.45,0.45,0.45)	"	"	0.2956	0.3150	6.53	28.65	0.32	0.30
3	(0.225,0.225,0.225)	"	"	0.2952	0.3155	3.27	21.08	0.37	0.36
4	(0.113,0.113,0.113)	"	"	0.2937	0.3135	1.63	19.08	0.44	0.40

2.2.3 Procedure

The experiment was controlled by a Dell Precision T3500 desktop computer running Windows 10 v1809 (OS Build 17763.503) and PsychoPy v3.0.7 (Peirce & Macaskill, 2018). Following the Method of Triads variant of MLDS, three images were simultaneously presented on each trial, which remained visible until the participant selected the (left or right) pair of images that depict the smallest difference in gloss relative to the central target object. After a response was entered, the images were replaced by a central white fixation cross for 750 ms, and the next trial would begin with a new combination of images. With three images presented on each trial, and seven different images in the stimulus set, each participant completed a total of 35 trials (i.e., one trial per distinct combination of three out of seven images). Observers typically completed the experiment in less than 5 minutes.

2.2.4 Results

The pooled responses from all 10 participants were treated as trial repetitions and analyzed using the implementations of MLDS by Kingdom and Prins (Kingdom & Prins, 2016) and Aguilar et al. (Aguilar et al., 2017). As can be seen in Figure 2a (orange data), MLDS reveals that for this particular scene, linear steps in stimulus magnitude (specular reflectance) are nonlinearly related to differences in perceptual magnitude (perceived gloss). Previous studies have found the relationship between physical reflectance and perceived gloss to be approximately linear (Pellacini et al., 2000), or a complex nonlinear function (Obein et al., 2004), while here we observe a very mild compressive function. The assumed form of this function and its best-fitting coefficients (determined by nonlinear least squares) are shown in Eq. 1

$$\psi = -1.8472 \exp(-14.27\, S) + 1.4495 \tag{1}$$

where perceptual magnitude $\psi \in \mathbb{R} : \psi \in [0,1]$ and stimulus magnitude $S \in \mathbb{Z} : S \in [0.017,0.099]$. However, when this perceptual scale is plotted against the specular contrasts (SC) of each image (Figure 2b; green data), there is a linear relationship ($\psi = 6.83SC - 0.68$). This is in agreement with the findings of Pellacini, Ferwerda, and Greenberg (2000) who used multidimensional scaling methods in their studies.

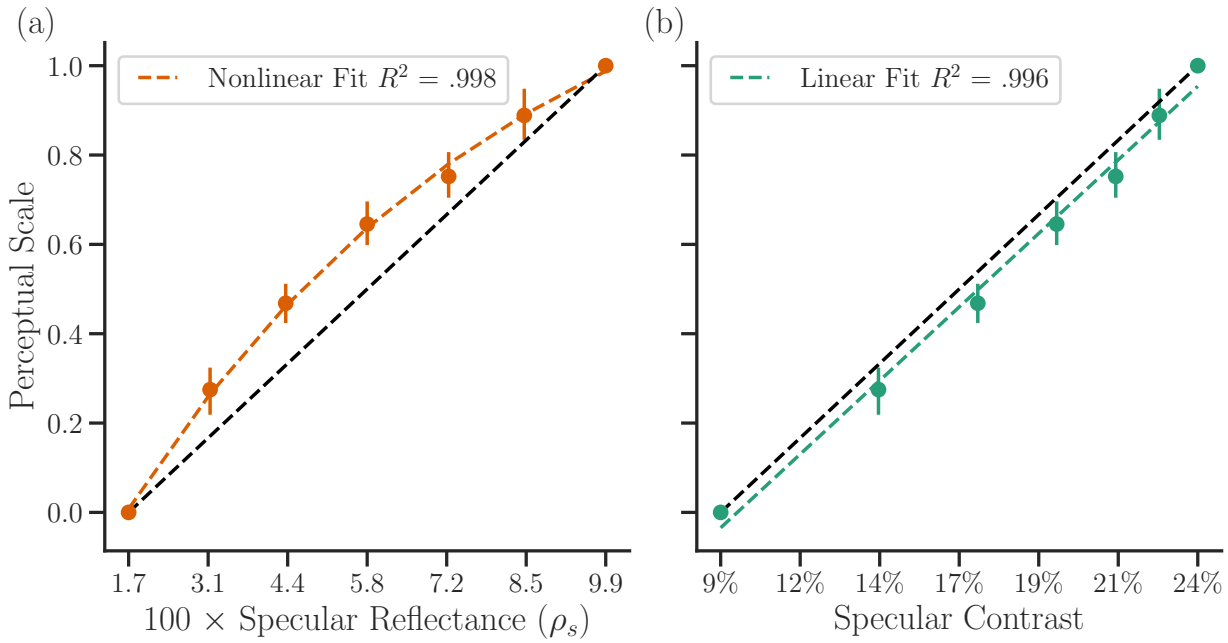

Figure 2. Maximum Likelihood Difference Scaling results. The underlying MLDS data is provided in Data File 1 (Cheeseman et al., 2020). (a) According to the MLDS perceptual scale, perceived magnitudes of gloss are related to physical magnitudes of specular reflectance by a compressive non-linearity (orange data), which deviates from a perfectly linear relationship (black diagonal). (b) Perceived gloss is linearly related to specular contrast (green data). Error bars indicate bootstrapped 95% confidence intervals.

2.3 Experiment 2: Measuring discriminability on a perceptual scale

With a suprathreshold perceptual gloss scale in hand, we sought to characterize discriminability at equidistant locations on this scale. However physical and perceptual magnitudes are quantitatively related for a given set of conditions, it is often assumed that the tasks employed to estimate discriminability, or to construct a perceptual scale, involve qualitatively similar kinds of judgments. In other words, the difference between suprathreshold and near-threshold judgments should be one of degree and not of kind. In the following experiment discriminability is estimated with the Method of Constant Stimuli, which unlike MLDS, requires values of specular reflectance that probe the full range of discriminability in order to determine just-noticeable differences of this parameter. This experiment therefore tests whether suprathreshold scaling (MLDS) can predict differences in discriminability that normally accompany absolute changes in stimulus magnitude.

2.3.1 Participants

A distinct group of twenty-three adults (10 males and 13 females; age range: 18 to 29 years; *M* = 22.8 years, *SD* = 3.2 years) with normal or corrected-to-normal visual acuity participated in the experiment and were paid 8€ per hour. All participants provided informed consent prior to the following experiments, which were approved by the ethics review board at Justus Liebig University Giessen and conducted in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki).

2.3.2 Stimuli

The virtual scene from the previous experiment was also used here; however, subthreshold and suprathreshold intervals of specular reflectance were used to vary the gloss of the target object. Three equidistant standard parameter values of specular reflectance were calculated using the perceptually uniform scale obtained in Experiment 1. This was accomplished by inputting five linearly spaced perceptual magnitudes ($\psi \in \mathbb{R} : \psi \in [0,1]$) to the inverted form of Eq. 1

$$S = \frac{\log\left(\frac{\psi - 1.4495}{-1.8472}\right)}{-14.27} \tag{2}$$

and retaining the middle three values ($\rho_s = \{0.030, 0.047, 0.068\}$). The perceived difference in gloss between each of the three standard values of specular reflectance is therefore equivalent. Ten comparison values of specular reflectance were also calculated for each standard, with five values above and five below each corresponding standard value. To ensure the perceptual uniformity of each set of comparison values, the minimum and maximum comparison values for each standard were calculated using the perceptually uniform scale, while intermediate comparison values were scaled logarithmically. The complete stimulus set (3 standards + 30 comparisons = 33 images; available in Dataset 1; see Cheeseman et al., 2020) was rendered with the values of specular reflectance listed in Figure 3.

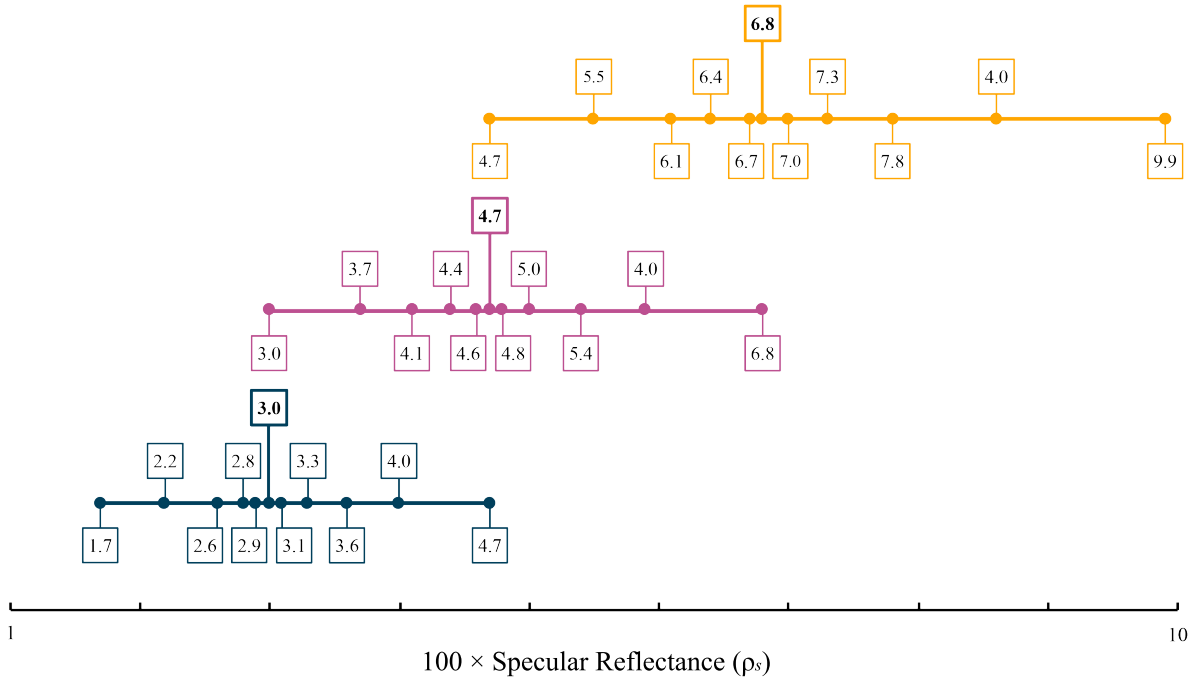

Figure 3. Rendered values of the Ward specular reflectance parameter (ρ_s) used to estimate discriminability with the Method of Constant Stimuli. Three standard values (shown in bold) and ten logarithmically scaled comparison values for each standard were calculated at equidistant locations on the perceptually uniform scale obtained in Experiment 1 with MLDS. On any given trial in the experiment observers visually discriminated between a low, medium, or high reflectance standard image and a randomly selected comparison image from the corresponding subset.

2.3.3 Procedure

Observers were tested under the same conditions described for Experiment 1, except for the following important differences. Here, with the goal of measuring discriminability on our perceptually uniform scale, we employ the Method of Constant Stimuli in a 2AFC task, wherein two images (i.e., a standard and comparison stimulus) are presented on each trial, and the observer selects the left or right image depicting the target object with the greater degree of gloss. The low, medium, or high reflectance standard images were only paired with images from the corresponding subset (e.g., if the standard reflectance $\rho_s = 0.030$, then the comparison reflectance $\rho_s \in \{0.017, 0.022, 0.026, 0.028, 0.029, 0.031, 0.033, 0.036, 0.040, 0.047\}$). The standard stimulus image appeared at random on either side of the screen. Stimulus pairs for each of the three standards were randomly interleaved, and the observers were shown 15 repetitions of the entire set (30 image pairs

× 15 repetitions = 450 trials per observer). Once the observer ended the current trial by entering a response using the left or right arrow key, the screen was cleared for 1 second, and the images for the next trial were displayed. In order to limit the total duration of the experiment to approximately 1 hour, the images were displayed for a maximum of 5 seconds before disappearing from the screen, after which the observer could advance to the next trial by entering a response.

2.3.4 Results

The proportion of trials in which the target object was judged to be glossier in the comparison image was calculated separately for the low, medium, and high ranges of specular reflectance. Logarithmic curves were then fit to these proportions at each value of specular reflectance via Bayesian estimation (Schütt et al., 2016). The psychometric function slopes for each observer (Figure 4a) illustrate that significant differences in discriminability were found at equidistant locations on our perceptual scale. A one-way repeated measures ANOVA confirmed that for the majority of observers, the slope of the psychometric function decreases with greater magnitudes of specular reflectance ($F(2,44) = 46.3$, $p < .001$, $\eta^2_p = .678$). Our observers were therefore less sensitive to increasing values of specular reflectance. Differences in discriminability can be seen when psychometric functions estimated from pooled data for each standard are plotted on the physical axis (Figure 4b). However, these differences in slope are eliminated when the psychometric functions are plotted on the perceptual scale (Figure 4c). This result demonstrates that the perceptual scale is responsible for the pattern of discriminability across our range of specular reflectance, and further suggests that MLDS may be used to compensate for such differences in discrimination performance.

The perceptual scale generated by MLDS in Experiment 1 was then used to calculate discriminability estimates that could be directly compared with those obtained in the current experiment. This was accomplished by reparametrizing the perceptual scale in *d'* units and reading out discrimination thresholds at specified levels of performance (a detailed technical explanation is provided in Aguilar et al., 2017; analysis code available at http://github.com/TUBvision/mlds). Six *d'* values ($d' \in \{-2.0, -1.0, -0.5, 0.5, 1.0, 2.0\}$) were used to estimate thresholds from the MLDS perceptual scale at each of the three standard values of specular reflectance. In a standard 2AFC paradigm these *d'* values correspond to correct response rates of 8%, 24%, 36%, 64%, 76%, and 92%, respectively. Discrimination thresholds at these performance levels were then read out

from the psychometric functions obtained for low, medium, and high specular reflectance standards in the current experiment (Schütt et al., 2016). According to this between-subjects analysis, there is broad agreement between the thresholds predicted by MLDS and those obtained using a 2AFC task and the Method of Constant Stimuli. This can be seen in Figure 4d, where the thresholds for both methods are plotted against each other for the low, medium, and high standards. Note that the 95% confidence intervals for all of the estimated thresholds cross the identity line, thus indicating that negligible differences exist between these methods, at least when directly compared on a common metric.

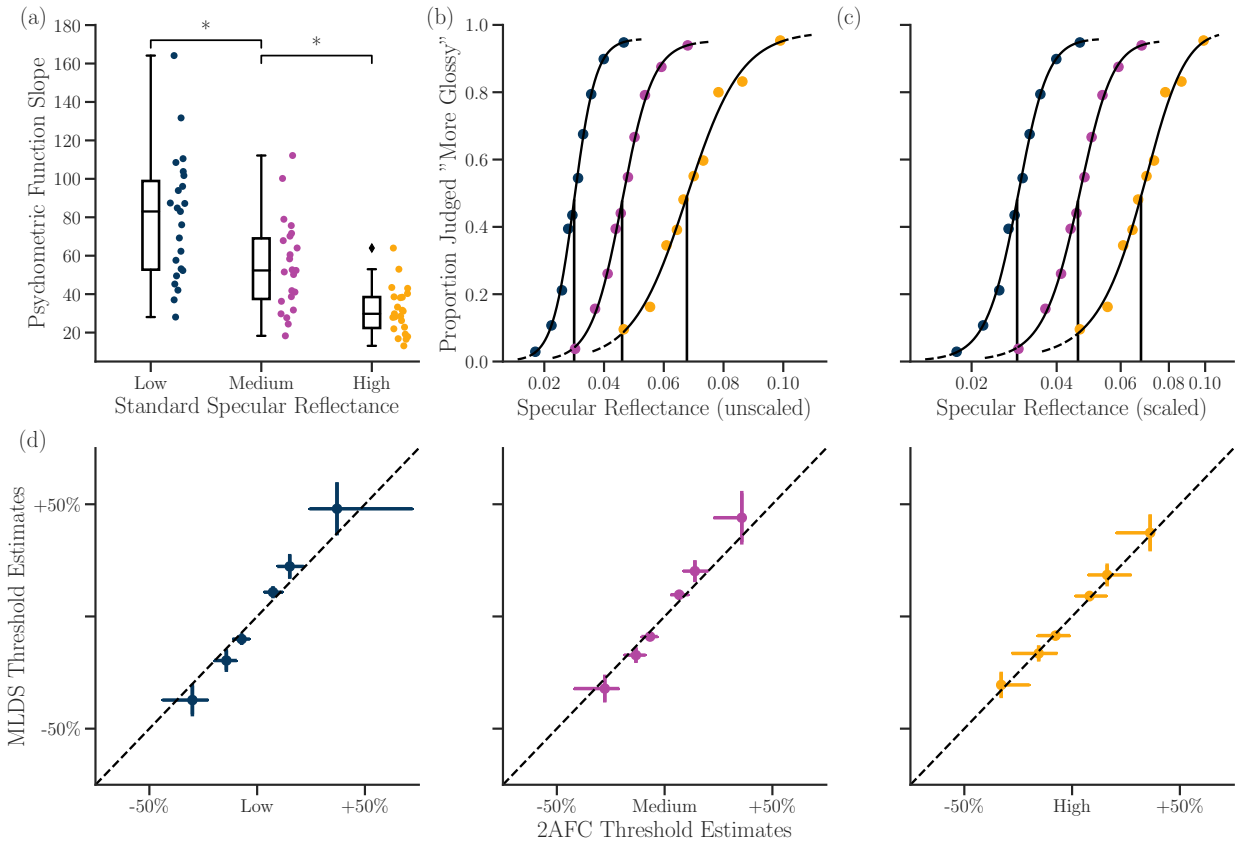

Figure 4. Discriminability estimates obtained with the Method of Constant Stimuli. The underlying 2AFC data is provided in Data File 2 (Cheeseman et al., 2020). (a) Psychometric function slopes for individual participants (colored data points) and corresponding box plots for the three standards. Asterisks represent a significance level of $p < .01$. (b) Psychometric functions (pooled across participants) for each of the three standard parameter values, here plotted on the unscaled physical axis. (c) Differences in the slope of these psychometric functions are eliminated when plotted on the perceptual scale. (d) Discrimination threshold estimates for the three standard parameter values obtained from the reparametrized MLDS perceptual scale (Experiment 1) and the Method of Constant Stimuli in a 2AFC task (Experiment 2). The thresholds are expressed as differences relative to each standard. Vertical and horizontal lines indicate bootstrapped 95% confidence intervals. The confidence intervals for all estimates cross the (black diagonal) identity line, thus indicating that the estimates from each method are not significantly different.

2.4 Discussion

If surface specular reflectance signals the only difference that could be seen between two otherwise identical surfaces, how does the magnitude of this difference affect what visual information observers use to judge these surfaces? The current study set out to answer this fundamental question in two experiments. First, we established a perceptual scaling of specular reflectance using MLDS, which involves judging the similarity of suprathreshold image differences. We then characterized discriminability along this scale using the Method of Constant Stimuli in a 2AFC task, in which discrimination thresholds are estimated by presenting observers with image differences that span the full range of discriminability. Taken together, our results provide convergent evidence that MLDS can scale both small and large image differences, which allows for successful prediction of discrimination thresholds.

In the formalism of MLDS, sensory representations are modeled as independent, Gaussian random variables with equal variance, while the precision of each trial decision (i.e., which pair is more similar) is estimated by the fitting procedure. Simulated violations of this equal variance assumption about internal sensory noise do not affect the shape of the perceptual scale produced by MLDS (Kingdom & Prins, 2016; Maloney & Yang, 2003), but may affect discriminability estimates derived from the same perceptual scale (Aguilar et al., 2017). The model assumptions underlying MLDS may also interact with stimulus complexity and dynamic range (Aguilar & Maertens, 2020; Protonotarios et al., 2016), both of which have been shown to affect the perception of gloss (Adams et al., 2018; Doerschner et al., 2010; Obein et al., 2004; Phillips et al., 2009). It is also plausible that scaling and discrimination tasks induce—or draw on—non-trivial differences in stimulus representation. In the case of surface gloss, near-threshold discrimination involves attending to local features that signal small differences in the proximal stimulus, while suprathreshold scaling involves attending to whole objects and abstracting similarity from multiple dimensions of the distal stimulus (Maloney & Knoblauch, 2020; Phillips et al., 2010). Such task-dependencies may be particularly relevant when the stimulus property in question ("gloss") consists of multiple appearance dimensions (Hunter & Harold, 1987; Toscani et al., 2020), and is thus more open to interpretation. Then again, under symmetric viewing conditions, where the only visible differences between otherwise identical images are to be found in the relative magnitudes of specular reflectance, the complexity of surface gloss is boiled down to a manipulation of local contrast (Figure 2b). Our experiments are therefore analogous to those described by Kingdom

(2016), who compared scaling and discrimination data from experiments (originally published by Whittle (1986, 1992) in which observers judged the difference in luminance of a disk superimposed against a uniform background. Analyses of those data revealed a remarkable degree of agreement between the scaling and discrimination tasks, which was taken as evidence that the sensory representation of contrast is governed by additive noise. Similarly, if it is assumed that a common transducer function mediates scaling and discriminability of perceived gloss, our results indicate that internal sensory noise remains constant for suprathreshold and near-threshold intervals of specular reflectance. Given the potential limitations of MLDS described above, it is reassuring that our findings agree with previous studies that demonstrated agreement between MLDS perceptual scales and discrimination performance for other appearance characteristics (Devinck & Knoblauch, 2012b; Kingdom, 2016). This suggests that, at least for comparisons of surfaces that differ only in specular reflectance, MLDS is well able to model judgments of suprathreshold and near-threshold differences in surface appearance.

The results of our experiments also indicate that gloss sensitivity cannot be captured by a single point estimate, since discriminability of gloss critically depends on the magnitude of surface specular reflectance. In this regard, gloss sensitivity would seem to follow Weber's Law, which assumes that discriminability is invariant if and only if physical magnitudes are varied in constant proportion to perceptual magnitudes (Fechner, 1860/1966). Weber's Law has inspired considerable debate about the transducer functions that relate stimulus and sensation (Stevens, 1961), yet from its inception, Fechner acknowledged that the lawfulness of Weber's Law depends on the nature of the stimulus. For example, he comments that while the law could be demonstrated with experiments in pitch perception, a case for its existence in color perception could not then be made (Fechner, 1860/1966). This early observation suggested that a perceptually uniform color space would be a complex mathematical entity, and these complexities were not fully appreciated until the next century, when it was discovered that small differences in chromaticity could only be adequately specified within local regions of the CIE 1931 color space (MacAdam, 1943; Smith & Guild, 1931). Similarly, the prospect of a uniform perceptual space for surface gloss remains elusive because changes in illumination, shape, and viewpoint can drastically alter the perception of surface material properties (Fleming et al., 2003; Ho et al., 2007; Norman et al., 2016, 2020; Vangorp et al., 2007; Zhang et al., 2020), which therefore means that the validity of any gloss space will be constrained by the viewing conditions chosen for its construction (Fores et al., 2014).

Despite these difficulties, our finding that MLDS provides a solution for both scaling and discriminability of gloss indicates that the construction of a perceptually uniform gloss space is a tractable problem. Moreover, MLDS offers considerable efficiency advantages. To evaluate sensitivity at just three reflectance values using the method of constant stimuli we used 450 trials per participant, many of which were close to threshold performance and therefore potentially frustrating for the participants. While this could be made somewhat more efficient through an adaptive sampling procedure (Kontsevich & Tyler, 1999; Lieberman & Pentland, 1982; Watson, 2017), in contrast, MLDS delivered quite accurate sensitivity estimates with just 35 trials per participant. This makes it feasible to compare sensitivity across many conditions, a prerequisite for future studies investigating how factors such as lighting, shape and other reflectance parameters influence sensitivity to gloss.

2.5 Conclusions

Returning to our central question: to what extent do suprathreshold judgments of surface gloss predict near-threshold discrimination of specular reflectance? It has been argued that just-noticeable differences can predict suprathreshold differences in complex visual properties (Mantiuk et al., 2011), and also that such small image differences are not necessarily relevant to the task of scaling material appearance (Ramanarayanan et al., 2007). Our results demonstrate that MLDS, a method of perceptual scaling that works with suprathreshold appearance differences, not only predicts discriminability of specular reflectance, but also provides a means for improving the perceptual uniformity of discriminability estimates. Future work will need to characterize the extent to which estimates of gloss discriminability can generalize across asymmetric viewing conditions, in which multiple dimensions of gloss are varied in addition to changes in illumination, shape, and viewpoint. Yet in the long run, a model of surface gloss perception will only be complete if it can correctly predict variations in discriminability as well as suprathreshold appearance.

CHAPTER 3: GLOSS DISCRIMINATION: TOWARDS AN IMAGE-BASED PERCEPTUAL MODEL

A similar version of this chapter has been uploaded to a public pre-print server as:

Cheeseman, J. R., Ferwerda, J. A, Morimoto, T., & Fleming, R. W. (2024). Gloss discrimination: Towards an image-based perceptual model. PsyArXiv Preprint. https://doi.org/10.31234/osf.io/anx4q

Gloss is typically considered the perceptual counterpart of a surface's specular reflectance characteristics, much as color is the perceptual counterpart of a surface's diffuse reflectance spectrum. In many contexts, it is tempting to ask how discriminable two surfaces are given their reflectance properties. Yet, as we argue here, this is a poorly-posed question, as factors other than reflectance (e.g., lighting, shape, viewpoint) can have substantial effects on how discriminable two images of glossy surfaces are to human participants. This fundamental difficulty with predicting gloss discrimination, whether from a physical measurement or from proximal image data, has so far hobbled efforts to establish a rigorously defined perceptual standard for surface gloss, like those that exist for color. Here, we propose an experimental framework for making this problem tractable, starting from the premise that any perceptual standard of gloss discrimination must account for how distal scene variables influence the statistics of proximal image data. With this goal in mind, we rendered a large set of images in which shape, illumination, viewpoint, and surface roughness were varied. For each combination of viewing conditions, a fixed difference in surface roughness was used to create a pair of images showing the same object (from the same viewpoint and under the same lighting) with high and low gloss. Human participants (N = 150) completed a paired comparisons task in which they were required to select image pairs with the largest apparent gloss difference. Importantly, rankings of the scenes derived from these judgments represent differences in perceived gloss independent of physical reflectance. We find that these rankings are remarkably consistent across participants, and are well predicted by a straightforward Visual Differences Predictor (Daly, 1992; Mantiuk et al., 2023). This allows us to estimate reasonable bounds on visual discriminability for a given surface across a wide range of viewing conditions. This has potential applications in both vision science, computer graphics and industrial contexts.

3.1 Introduction

Determining visual thresholds for proximal stimulus variables—such as luminance (Nachmias & Kocher, 1970), wavelength (Pokorny & Smith, 1970), contrast (Campbell & Robson, 1968), orientation (Appelle, 1972) or spatial frequency (Campbell et al., 1970)—is conceptually straightforward, with well-defined psychophysical methods underpinned by signal-detection theory (Green & Swets, 1966). Yet, it also often happens that we want to know how well participants can distinguish between stimuli that differ in some distal physical property, such as surface gloss. For example, in the pigment and paint industry, it is often necessary to manufacture parts with matching surface appearance, which would require differences in appearance to be 'within tolerance', i.e., below threshold (for a recent review and commentary, see European Coatings Dossier on Testing and Measuring, 2019). Both R&D and quality control require some means to establish whether two samples are perceptually indistinguishable in terms of their gloss. Ideally, it should be possible to do this on the basis of a physical measurement applied to the surfaces. Similarly, computer graphics researchers often need to know how sensitive participants are to reflectance parameters, to determine, for example, whether a given approximation is acceptable (Greenberg et al., 1997; Pellacini et al., 2000). And in vision research, establishing discrimination thresholds for reflectance properties would also be useful for characterizing human perceptual abilities and constraining theories of gloss perception.

However, although the idea of measuring discrimination thresholds for gloss seems intuitive enough, there is a fundamental challenge because surface reflectance is a distal scene property, rather than a proximal stimulus variable like luminance or cone excitation ratios. The images that form the basis of any threshold measurements are the result of complex interactions between multiple distal scene factors in addition to the reflectance: the illumination striking the surface, the surface's shape and the observer's viewpoint. It is not possible to 'leave out' any of these factors; designating values for each factor is a prerequisite for creating the images required for the experiment. Nonetheless, lighting, shape and viewpoint can have potentially enormous effects on the measured thresholds. Under one set of conditions, a given difference in surface reflectance can significantly alter many pixels in the image, yielding very low threshold estimates (Figure 1A). Yet under other view conditions, the exact same difference in reflectance could have little to no effect on the image, and therefore yield infinite threshold estimates (Figure 1B). Thus, although we can experimentally determine whether any two images of surfaces are perceptually

distinguishable, we do not know how the results will generalize to other conditions. In concrete terms: gloss thresholds measured under one illumination may be useless for determining whether two surfaces are perceptually distinguishable under a different illumination. The same holds for changes in shape or even viewpoint. Here, we seek to provide an approach to circumvent this challenge to yield 'reasonable bounds' on discrimination thresholds for gloss.

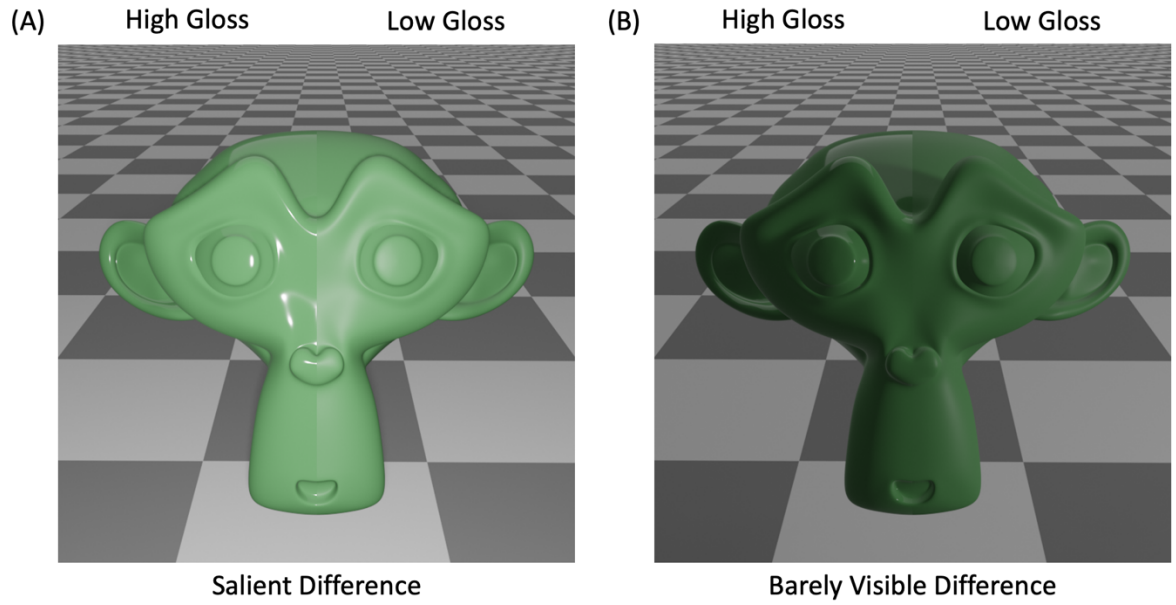

Figure 1. Identical differences in surface reflectance can be salient (A) or barely visible (B) depending on lighting direction.

The practical needs of industry have driven the development of numerous color spaces over the past century. One of the most well-known, the CIE 1931 XYZ color space, was computed from simple color-matching experiments, where participants adjusted lights of different wavelengths to have identical luminosity. Later in 1942, David MacAdam published the results of similar experiments that showed how sensitivity to differences in chromaticity vary within the 1931 CIE XYZ color space. Errors in color-matching performance were found to vary systematically within the space, thus indicating that equal increments within the space do not correspond to equal differences in perceived color. This motivated later researchers to propose color spaces that partially correct for such distortions (e.g., CIELAB, CIECAMO2, CIECAM16). A perceptually-uniform space for gloss would be especially useful for industrial applications, where there is a

need to maintain a consistent material appearance throughout the manufacturing process. However, such a space has remained elusive, owing to the multidimensional nature of gloss, and a lack of agreement about which dimensions of gloss are relevant for particular applications. The six dimensions of perceived gloss set out by Hunter and Harold (1987) have been highly influential, and other results suggest that two to four dimensions account for nearly all variance in subjective comparisons of gloss, at least for the range of surfaces that were considered (Kildau, 2016; Pellacini et al., 2000; K. E. Prokott, 2016; Toscani et al., 2020). However, the number and nature of these dimensions will depend on the set of appearances chosen for testing, and the intended application. While much attention has been given to understanding *biases* in gloss perception (i.e., influences of lighting or shape on the overall level of gloss; see Fleming et al., 2003; Motoyoshi & Matoba, 2012; Nishida & Shinya, 1998; te Pas & Pont, 2005; Vangorp et al., 2007), here we seek to define conditions for measuring *sensitivity* to changes in surface reflectance, paving the way for standards that could serve both industry and vision researchers.

Previous researchers have attempted to characterize gloss perception using a variety of experimental and analytical frameworks, traditionally with real surfaces in controlled lighting environments. For example, Obein et al. (2004) assessed the relationship between the perceived gloss of real surfaces and instrumental measurements of specular reflection using Maximum Likelihood Difference Scaling (MLDS; Maloney & Yang, 2003). Although one of their central claims is that participants' judgments of gloss exhibit constancy under changes in viewing angle, they did not find statistical evidence that a single scaling could be used for multiple incident angles of illumination. Indeed, the evidence for gloss constancy is rather mixed (Chadwick & Kentridge, 2015; Doerschner et al., 2010; Faul, 2019; Fleming et al., 2003; Olkkonen & Brainard, 2011), and it is not obvious how instrumental measurements can possibly generalize much beyond the original scene configuration, especially when shape and illumination are varied in addition to changes in viewpoint. It has long been known within the field that measuring the proportion of reflected light at a sparse sampling of incident angles is an unreliable predictor of perceived gloss (Harrison, 1945). Nevertheless, despite well-documented shortcomings, 'gloss meters' based on this principle remain the industry standard, in part because such measurements can be collected quickly, and better methods are not widely available. On the other extreme, one can measure reflected light at many more incident angles, covering the entire hemisphere above the surface plane, and use this data to estimate a bidirectional reflectance distribution function (BRDF; Nicodemus et al., 1977).

Until very recently, measuring BRDFs has been too costly and inefficient for widespread practical application (Filip & Kolafová, 2019). Despite these recent technical advances, however, it is unlikely that our perceptions of gloss are based on a BRDF-like representation of surface reflectance. Indeed, we have argued that the brain generates heuristic representations, or 'statistical appearance models' of gloss appearance over a range of typical viewing conditions (Fleming, 2014; Fleming & Storrs, 2019).

Advances in computer graphics simulation over the previous three decades have allowed vision researchers to apply these technologies to the study of gloss perception. For example, the study by Pellacini et al. (2000) is notable for its application of Multidimensional Scaling (MDS; Borg & Groenen, 2005) to judgments of glossy spheres shown in simulated illumination. With this data, they constructed a perceptually-uniform gloss space consisting of two dimensions (contrast and distinctness of the reflected image), which they later used to derive just-noticeable differences (JNDs) in gloss (Ferwerda et al., 2001). While these authors were the first to apply this approach to understand gloss perception, the generalizability of their results is limited to the set of appearances used to create the space (Fores et al., 2014). Given that shape and illumination strongly influence material appearance (Vangorp et al., 2007), what is a sufficiently-diverse set of conditions for the purpose of characterizing gloss sensitivity? In the limit, iteratively rendering many combinations of illumination, shape, viewpoint, and surface reflectance will yield a set of images that includes the 'typical' appearance of glossy surfaces across multiple material categories. However, in an industrial manufacturing context (e.g., quality control for surface coatings), often the goal is to measure appearance changes between multiple copies of a single material formulation. Our previous study, Cheeseman et al. (2021), investigated gloss perception in such 'symmetric' viewing conditions, where we measured sensitivity to differences along a single perceptually-uniform dimension (specular reflectance), showing that even with all other variables held constant, sensitivity varies significantly with stimulus magnitude. Here, we pursue a complimentary approach—holding surface reflectance constant while varying illumination, shape and viewpoint—in order to identify viewing conditions where estimates of sensitivity to differences in surface reflectance have optimal generalizability. The current study therefore seeks to establish a framework for characterizing sensitivity to gloss *per se*, and perhaps, to other qualities of material appearance. We show that under symmetric conditions—when all scene parameters except reflectance are held constant—gloss discrimination reduces to an image

discrimination task that can be well predicted by extant image-discrimination models. As a result, we can predict the variations in gloss discrimination that occur as various scene parameters are altered. This provides a route into defining 'reasonable bounds' on gloss discrimination across viewing conditions.

3.2 Experiment 1: Predicting apparent gloss differences across viewing conditions

Many studies have assessed gloss perception by varying the reflectance of surfaces under different viewing conditions. Here, we present participants with a fixed difference in surface reflectance while varying illumination, shape and viewpoint, with the intent of identifying an image metric that can predict perceived differences of gloss across viewing conditions. Importantly, because the difference in reflectance is identical across conditions, any visible differences in gloss are due to extrinsic distal variables that are independent of intrinsic surface reflectance. If an image metric can predict which viewing conditions tend to accentuate or obscure apparent gloss, this could provide a principled basis for establishing tolerances on gloss sensitivity in real world conditions.

3.2.1 Participants

One-hundred-fifty adults (79 males and 71 females; age range: 18 to 68 years; M = 27 years, SD = 8 years) with normal or corrected-to-normal visual acuity participated in the experiment and were paid €10 per hour. Participants were recruited online using Prolific (prolific.co); they were required to have native fluency in English, and a desktop or laptop computer. All experimental procedures were approved by the Justus Liebig University Giessen Psychology Department Ethics Board and conformed with the guidelines of the American Psychological Association (Version 2017) and the Declaration of Helsinki (Version 2013, excluding pre-registration). Informed consent was obtained from all participants.

3.2.2 Stimuli

Stimulus images were created with the Mitsuba physically-based renderer (Jakob, 2010). High dynamic range, linear RGB renderings were tone-mapped to low dynamic range sRGB images using the method described in Reinhard et al. (2002). Parameters controlling the overall luminance (key) and clipping of highlights (burn) in the image were set to the Mitsuba-default values of 0.18 and 0, respectively. The 720 × 720 pixel images subtend 26 degrees of visual angle at a viewing

distance of 50 cm, although viewing distance was not controlled in the online experiment. Instead, participants were instructed to place a credit card (or another card of equivalent size) on their display screen, and adjust the length and width of a rectangle to match the size of the card. This measurement was used to calibrate the size of the images such that they were approximately the same size for different displays (see https://pavlovia.org/Wake/screenscale).

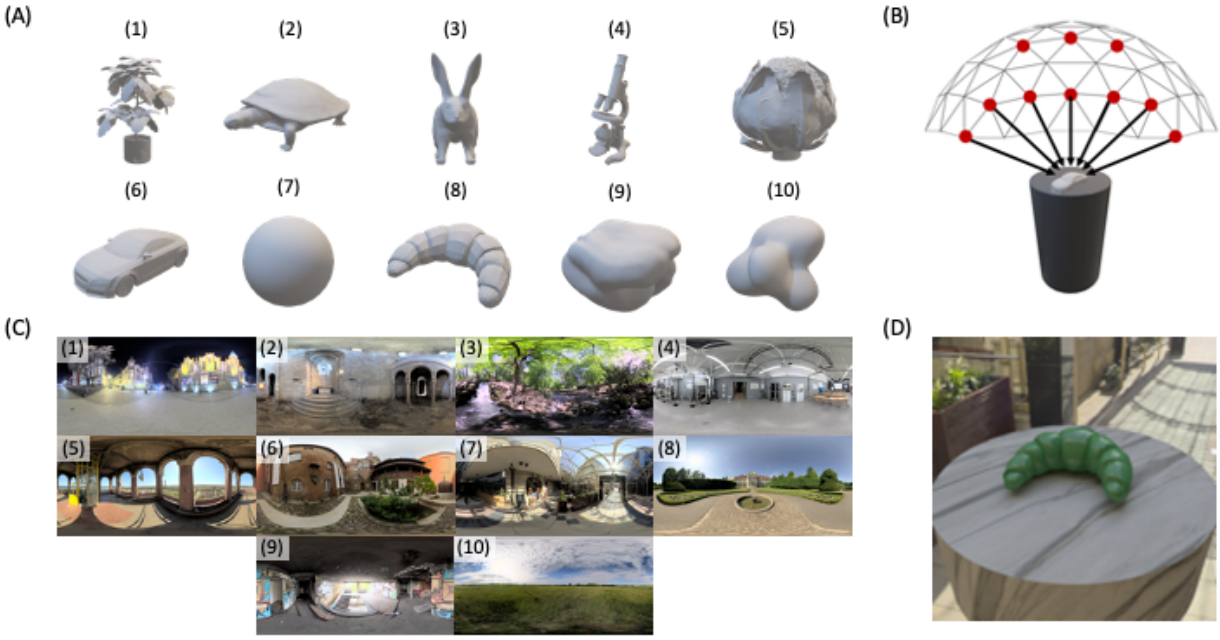

Figure 2. Scene variables that were used to create the stimulus images, including shapes (A), viewpoints (B), and illumination conditions (C). An example scene is shown in (D).

The basic scene (e.g., see Figure 2D) includes a central target object seated on a marble-textured pedestal under natural illumination. A set of 10 target objects (Figure 2A) was selected that span a variety of surface features that are more or less likely to accentuate gloss appearance. For example, some objects featured smoothly curved surfaces (e.g., car or turtle), while others featured rough or discontinuous surfaces (e.g., cabbage or plant). All objects in the scene were rendered with an improved version of the Ward BRDF model that obeys energy conservation and has better physical accuracy at grazing angles (Geisler-Moroder & Dür, 2010). The model has three parameters that control the specular reflectance (ρ_s), diffuse reflectance (ρ_d), and roughness (α) of a surface. The apparent glossiness of the target object was varied with two levels of surface

roughness (α = 0.01, 0.19) while specular and diffuse reflectance were fixed (ρ_s = 0.066, ρ_d = 0.1, 0.3, 0.1). A set of 10 high dynamic range environment maps (Figure 2C) was selected that featured a variety of indoor and outdoor illumination conditions. For example, environments with direct lighting can lead to bright, distinct highlights on a reflecting surface, whereas environments with diffuse lighting usually do not. Similarly, 10 evenly distributed viewpoints (Figure 2B) were sampled from the vertices of a hemi-icosphere positioned above the target object, thus providing a variety of high and low viewing angles for each target object. The combination of 10 shapes, 10 illuminations, and 10 viewpoints produced a set of 1000 scenes. A subset of 100 scenes was randomly sampled (without replacement) from this larger set for use in Experiment 1.

3.2.3 Procedure

The experiment was created in PsychoPy v2021.2.3 (Peirce & Macaskill, 2018) and run on the Pavlovia experiment hosting platform (pavlovia.org). To avoid requiring participants to pre-load a large set of images at the start of the experiment, 100 scenes were further divided into 10 subsets of 10 scenes. Separate groups of 15 participants were recruited to judge each subset of scenes; these can also be understood as 10 independent experiments with separate groups of participants and stimuli. The first stage of the experiment required participants to complete 20 practice trials in a simplified version of the task using luminance patches rather than rendered images. For each practice trial, participants were instructed to select the left or right *pair* of images that showed the larger difference in luminance, inspired by the Method of Quadruplets from Maximum Likelihood Difference Scaling (MLDS). Seven participants failed to correctly judge these exaggerated suprathreshold differences in luminance with at least 90% accuracy during the practice phase, and were excluded from the analysis, as it was assumed that they either did not understand the task instructions, or a technical problem impeded their performance. The task remained the same during the experimental trials, except that participants selected the left or right pair of images (rather than luminance patches) in which there is a larger difference in apparent gloss of the target object. Additionally, to account for outliers (e.g., due to random responding), participants who completed the experimental trials were excluded if their responses produced an extremely low correlation with other participants' judgments of the same set of images. Frequencies representing how often each scene is chosen were calculated for each participant, and compared across participants. If the average correlation between one participant's frequencies and those of the other participants

exceeds the Interquartile Range of these average correlations (multiplied by 1.5), this was considered an outlier, and the participant's data was excluded from the analysis. Twenty-two outliers were excluded in total. In summary, separate groups of 15 participants judged separate sets of 10 scenes. For each set of scenes, 45 unique scene pairs were presented in a random order across 6 repetitions. One-hundred-fifty participants collectively completed a total of 40,500 trials.

3.2.4 Image metrics

HDR-VDP-3 (Mantiuk et al., 2023) is a popular metric for predicting the visibility of image differences and assessing the impacts of compression or other image processing operations on image quality. In our analysis, HDR-VDP-3 was used to predict perceived differences between test and reference images. The metric was applied in a 'side-by-side' task mode, which is appropriate for comparing two images displayed adjacent to each other. Input images were encoded in 'sRGB-display' format to correspond with standard color images displayed on an sRGB monitor, with peak luminance calibrated to 100 cd/m^2 and a black level at 1 cd/m^2. The images were processed with a high angular resolution of 120 pixels per visual degree, appropriate for a close viewing distance or high-resolution display. For the modulation transfer function (MTF), which models the scattering of light in the eye's optics—referred to as glare—we chose to bypass this step by setting the 'mtf' option to 'none'. This decision was made because the glare effect, while significant for high-contrast HDR images, adds computational complexity that was not essential for our purposes. The output of HDR-VDP-3 provided us with a probability map of detection for each pixel (P_map), with values ranging from 0 to 1. We computed the mean of this probability map to represent the visibility metric for each stimulus condition, allowing us to assess the average detectability of image differences across the entire image. Although HDR-VDP-3 also provides a single valued probability of detection (P_det) for the whole image, we found that the average of the probability map (P_map) was a better predictor of the human data.

Unlike HDR-VDP-3, which can predict visible differences from full sRGB images, our measurements of contrast, coverage, sharpness, and skewness (similar to previous studies; see Marlow et al., 2012; Motoyoshi et al., 2007) were derived by first converting the sRGB images into luminance images (calibrated to cd/m^2). We then eliminated the diffuse component, thus ensuring the metrics were computed only from specular reflections. Subsequently, to remove reflections from within the object or from the pedestal, we thresholded the specular image. Pixels

exceeding a certain intensity threshold—determined as a percentage (k%) of the highest intensity, with k values set at 0, 1, 3, 5, 10, 20, 30, and 40—were retained. The k values were selected to evaluate a range of intensity thresholds. We then decomposed the thresholded highlight image into eight sub-band images through Gaussian band-pass filtering across a range of frequencies. This allowed us to capture the effects of spatial frequency modulation on the perception of gloss (Boyadzhiev et al., 2015). The contrast for each frequency band, as well as for the combined frequency image, was determined by calculating the root-mean-square-error (RMSE) of the pixel intensities. Alongside contrast, we also evaluated metrics for highlight coverage and sharpness, which were calculated from the thresholded highlight images. Coverage is the proportion of the object area that is covered by specular reflections, providing an indication of the extent of gloss across a surface. Sharpness is defined by the rate of change in luminance, measured using the slope of the local magnitude spectrum and local maximum total variation (TV) to emphasize areas of the object where intensity transitions are most pronounced (Vu et al., 2012). The values of k (for contrast, coverage, and sharpness) and spatial frequency bands (for contrast only) were varied to determine values that produced the best correlation with the human data. All of these image metrics were calculated with the scene background masked, leaving only the object region.

3.2.5 Results

The scene that most participants judged to depict the largest visible difference in apparent gloss is shown in Figure 3 (A) and (B), along with (C) RMSE difference across RGB channels between the two images and (D) predictions of one image metric (High Dynamic Range Visual Difference Predictor; Mantiuk et al., 2023) versus human judgments for all stimuli. Apparent differences in gloss caused by variations in lighting, shape and viewpoint were well predicted by this image metric, which produced a correlation of .81 with the behavioral data (i.e., how often each scene was selected for showing a larger gloss difference). Other image metrics were also evaluated for their ability to predict this data, including sub-band contrast, highlight coverage, highlight sharpness, and the skewness of the pixel intensity histogram. These metrics have all been shown to correlate with gloss appearance in experimental conditions (Anderson & Kim, 2009; Kim & Anderson, 2010; Morimoto et al., 2023b, 2023b; Motoyoshi et al., 2007; Schmid et al., 2023). A Generalized Linear Model (GLM) was used to assess the relationship between these metrics and human judgments. The model included HDR-VDP-3, contrast, coverage, sharpness, and skewness

as predictor variables. The analysis revealed that only HDR-VDP-3 (coefficient = 0.1215, SE = 0.013, $z = 9.603$, $p < 0.0001$) and Contrast (coefficient = 0.2645, $SE = 0.054$, $z = 4.942$, $p < 0.0001$) were significantly correlated with human judgments (see Figure 4). Although HDR-VDP-3 is optimized for the prediction of psychophysical data and is not meant to be a biologically plausible model of the human visual system, HDR-VDP-3 does explicitly model the optical and retinal transformations that occur in the first stages of human visual processing, as well as subsequent parsing of spatial frequency and orientation information in primary visual cortex. These features are used to model contrast masking and (neural) contrast sensitivity, and collectively influence the metric's assessment of image quality and visibility of image differences. Despite the complexity of HDR-VDP-3, its overall predictive power in our study appears to depend on simpler, more fundamental image attributes such as contrast.

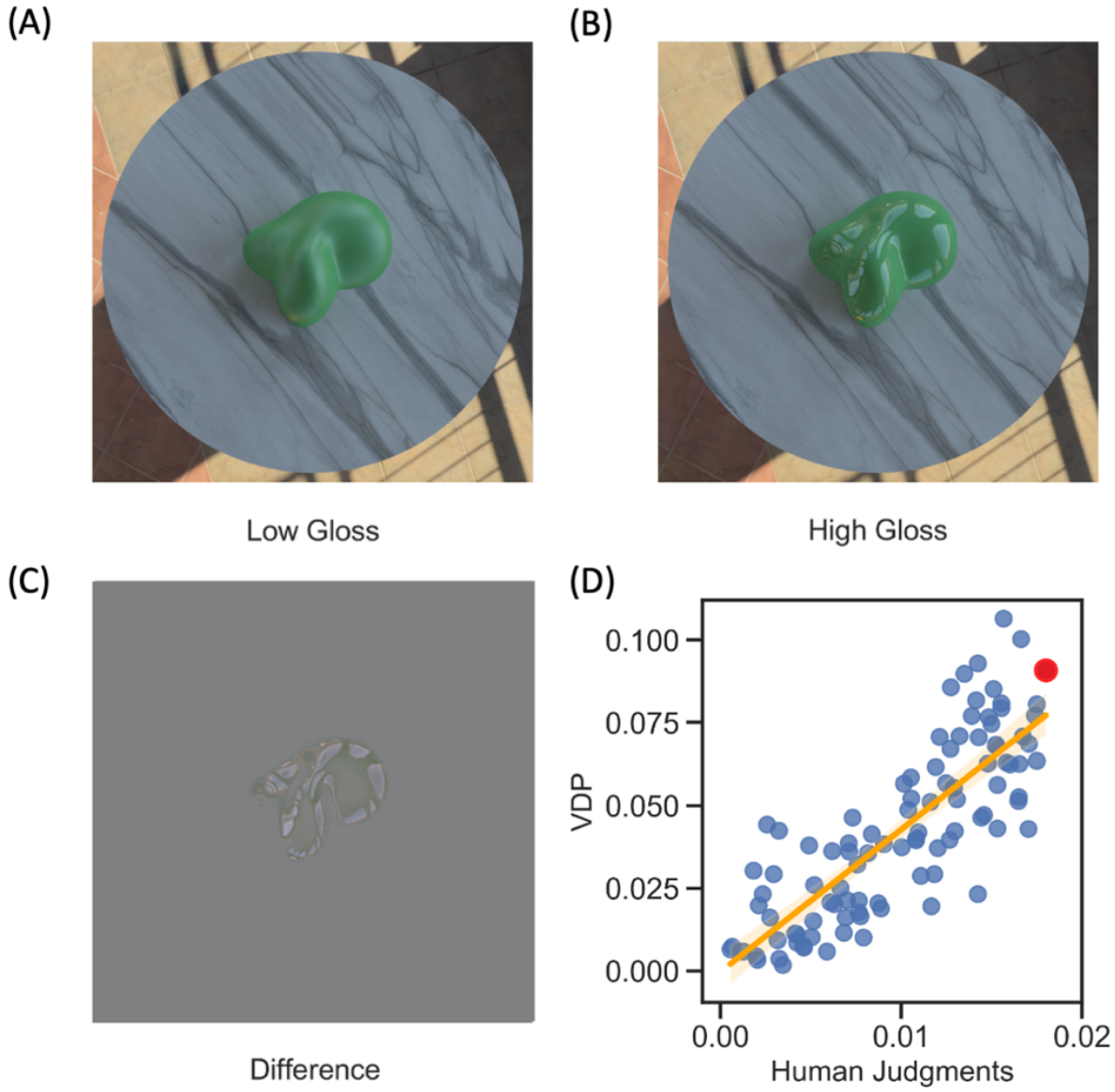

Figure 3. The scene with the largest visible difference (A-B), together with the luminance difference between the high and low gloss images (C). The scatterplot (D) shows a correlation of .81 between the VDP predictions (arbitrary units) and human judgments (proportion of trials each scene was chosen). Each datapoint in the scatter plot represents one of the hundred scenes; the datapoint highlighted in red corresponds to the scene with the largest visible difference (A-B).

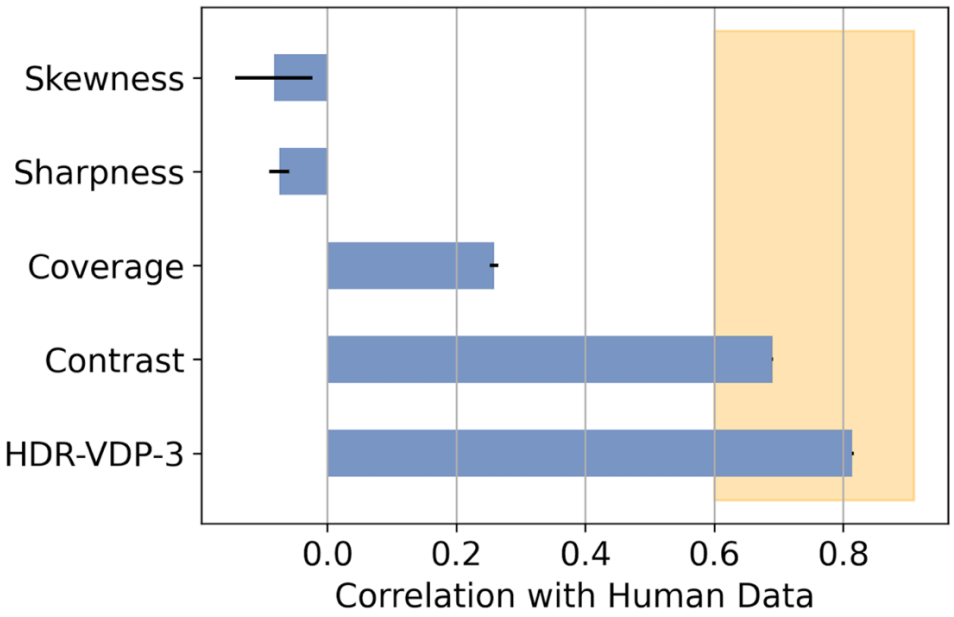

Figure 4. Correlations between image metrics and human data. Error bars represent SEM. The inter-participant correlations span the range highlighted in orange, with an average of .83.

3.3 Experiment 2: Image metric validation in a lab-based control experiment

We have a metric (HDR-VDP-3) that predicts human judgments of gloss differences across variations in lighting, shape and viewpoint. However, these judgments were collected from participants over the internet in uncontrolled viewing conditions with unknown display characteristics, viewing distance, and ambient illumination, which may have influenced our results (e.g., see Haghiri et al., 2019). The purpose of this second experiment was therefore to test whether the predictions of the model are also valid for controlled laboratory conditions using scenes selected from Experiment 1.

3.3.1 Participants

Twenty-two adults (7 males and 15 females; age range: 20 to 39 years; $M = 27$ years, $SD = 5$ years) with normal or corrected-to-normal visual acuity participated in the experiment and were paid €12 per hour. Participants were recruited from the university student population. All experimental procedures were approved by the Justus Liebig University Giessen Psychology Department Ethics Board and conformed with the guidelines of the American Psychological Association (Version

2017) and the Declaration of Helsinki (Version 2013, excluding pre-registration). Informed consent was obtained from all participants.

3.3.2 Stimuli

The distribution of VDP predictions for the full set of images is shown in Figure 5. As previously mentioned, we used a subset of 100 scenes for the first experiment (red bars). Now, for the second experiment, we selected two new scenes from the full set of 1,000 scenes (dark blue bars). Specifically, we chose two scenes where gloss sensitivity is predicted to be low or high, corresponding to the 10th and 90th percentile values of the distribution, respectively. For each scene, we rendered new images with finer differences in surface roughness using the log-spaced values illustrated in Figure 6 to validate the performance of the VDP across finer roughness levels. The standard roughness value was 0.1, and test images were rendered with the following values of roughness: 0.0702, 0.0824, 0.0901, 0.0950, 0.0981, 0.1019, 0.1050, 0.1099, 0.1176, and 0.1298. All other scene parameters remained identical to those used in Experiment 1.

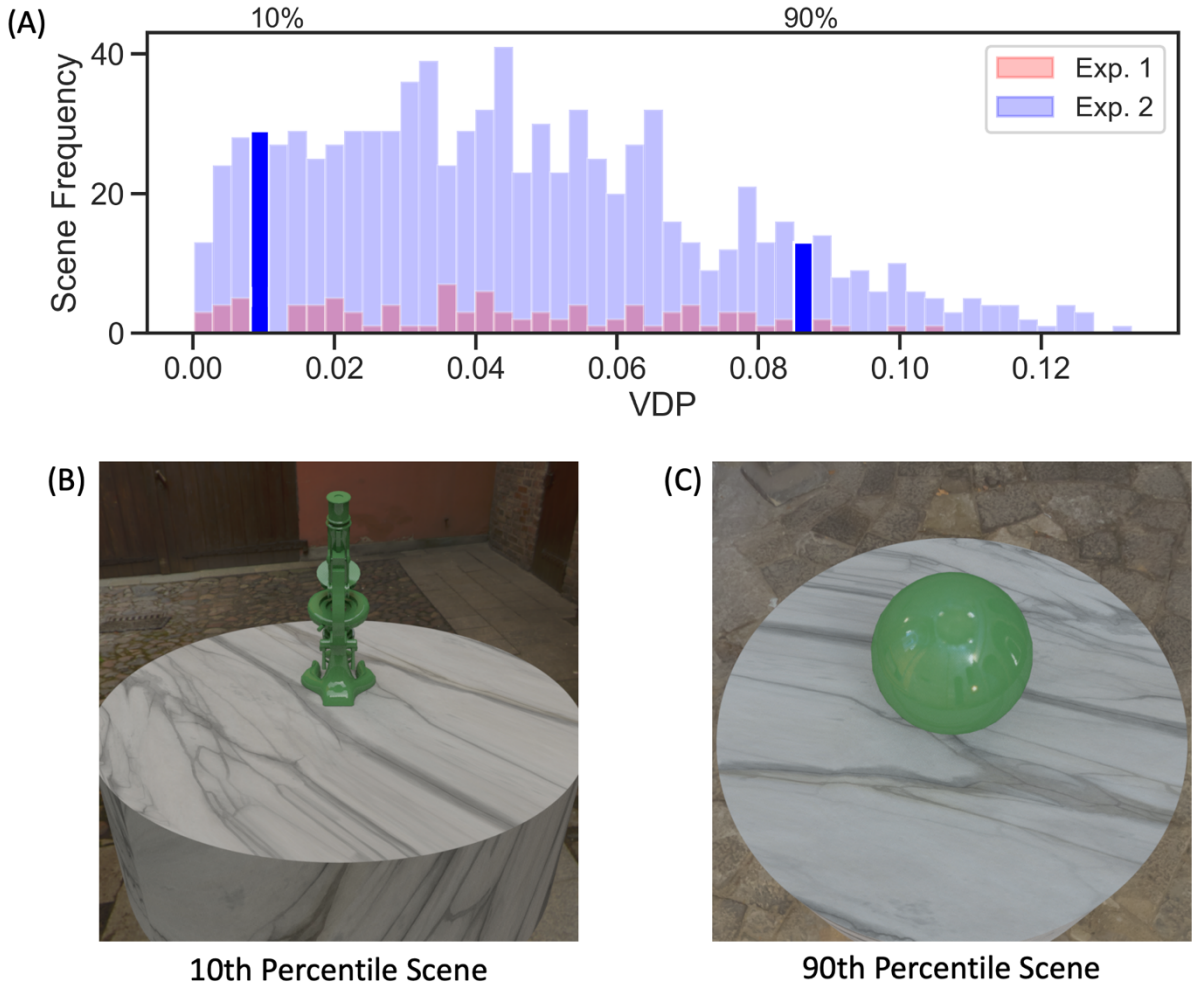

Figure 5. A histogram of the VDP predictions for the full set of images (A). Images used in Experiment 1 are highlighted in red. Images used for Experiment 2 (panels B and C) were selected from the 10th and 90th percentile bins of the histogram, shown highlighted in dark blue.

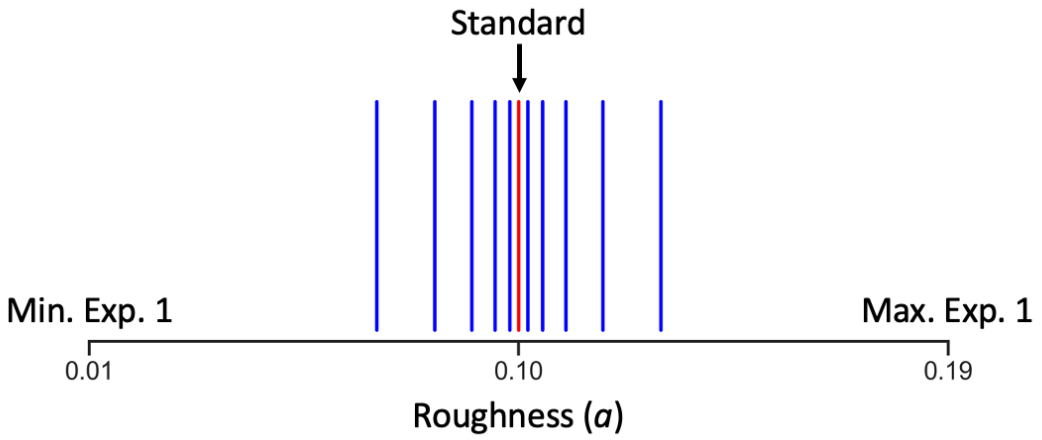

Figure 6. Roughness (α) parameter values of the Ward model used to create stimulus images for Experiment 2. The values of roughness used to create the low and high gloss images in Experiment 1 correspond to the minimum and maximum values of the range shown here.

3.3.3 Procedure

The experiment was created in PsychoPy v2021.2.3 (Peirce & Macaskill, 2018) and run on a Dell Precision T3500 desktop computer. The images were presented in a dark room on an Eizo ColorEdge CG277 LCD monitor, which features a 27-inch diagonal and a resolution of 2560 × 1440. Each image covered approximately 19 degrees of visual angle from a viewing distance of 50 cm. The display was calibrated to the sRGB color gamut with an 80 cd/m² D65 white point and a gamma setting of 2.2. During each trial, participants were presented with two images arranged side-by-side, and their task was to select the (left or right) image that showed a green target object with a higher degree of gloss. Other than this difference in the apparent gloss of the target object, all other scene variables were identical between the two images. This two-alternative forced-choice (2AFC) task replicates the procedure used in our previous work (Cheeseman et al., 2021). The first stage of the experiment required participants to complete 10 practice trials with a pair of images (selected from Experiment 1) showing a clearly visible gloss difference. Feedback was provided during the practice trials to indicate whether the object in the chosen image had a higher gloss (i.e., lower roughness). All participants were able to complete the practice trials without difficulty and were allowed to proceed to the next phase of the experiment. In this second phase, participants performed the same 2AFC task with images from two scenes that had not been shown in Experiment 1. Image pairs from each scene were presented repeatedly (40 times) in random order,

following the Method of Constant Stimuli; that is, for each scene, a standard image was presented alongside one of ten comparison images (see Stimuli section above). On average, the experiment lasted about 40 minutes, with a rest period at the halfway point. Collectively, our 22 participants completed 17,600 trials of this task.

3.3.4 Results

Figure 7 shows the minimum and maximum reflectance images for the 10th and 90th percentile scenes, along with their corresponding psychometric functions calculated from pooled observer data using *psignifit 4* (Schütt et al., 2016). HDR-VDP-3 predicts lower sensitivity for the 10th percentile scene, and higher sensitivity for the 90th percentile scene, respectively. Note that in Experiment 1, a fixed difference in roughness was predicted to be more visible with the combination of lighting, shape and viewpoint shown in the 90th percentile scene. In the current experiment, smaller differences in roughness were used to measure discrimination performance in controlled laboratory conditions. The significantly steeper slope of the psychometric function for the 90th percentile scene (mean slope = 12.65, *SE* = 6.46) compared to the 10th percentile scene (mean slope = 14.89, *SE* = 11.84) validates the prediction of HDR-VDP-3 – that participants would be more sensitive to the same reflectance differences when viewed with this combination of lighting, shape and viewpoint.

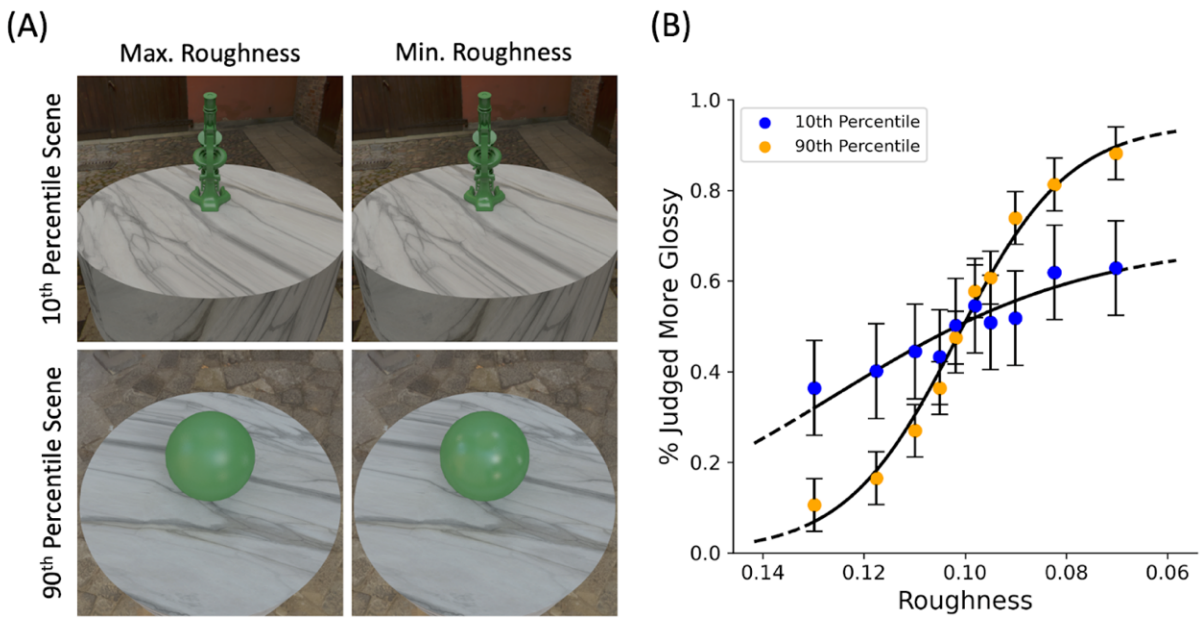

Figure 7. The minimum and maximum roughness images for the 10th and 90th percentile scenes (A), along with their corresponding psychometric functions calculated from pooled observer data (B). The significantly steeper slope of the psychometric function corresponding to the 90th percentile scene validates the prediction of HDR-VDP-3. Error bars signify the standard deviation of psychometric function fits across participants.

Given that HDR-VDP-3 has been validated in controlled laboratory conditions, the model predictions can be used to search for combinations of lighting, shape and viewpoint that should yield the highest sensitivity, shown in Figure 8. To evaluate the relative effect of lighting, shape and viewpoint on the model predictions, we used a random forest model (Breiman, 2001; Hastie et al., 2009). Random forests are particularly suited for this analysis because they do not assume linear relationships and are better able to predict continuous values using categorical variables. In this context, the target variable (mean P_map predicted by HDR-VDP-3) is estimated based on paths taken through a series of decision trees constructed from the categorical variables. The random forest algorithm considers every possible division of the VDP prediction values for levels of each categorical variable, and calculates which path will result in the largest decrease in variance of these values. Feature importance scores from a random forest model indicate which categorical variables contribute most to estimating the target variable across all the decision trees. Figure 9 shows importance scores for individual lighting maps, object shapes, and viewpoints, and the overall importance of each variable. Furthermore, if we calculate descriptive statistics on the

predictions of HDR-VDP-3, as shown in Figure 10, these can be used to estimate upper and lower bounds of predicted sensitivity for specific shapes (across lighting and viewpoint), or any other combination of these variables.

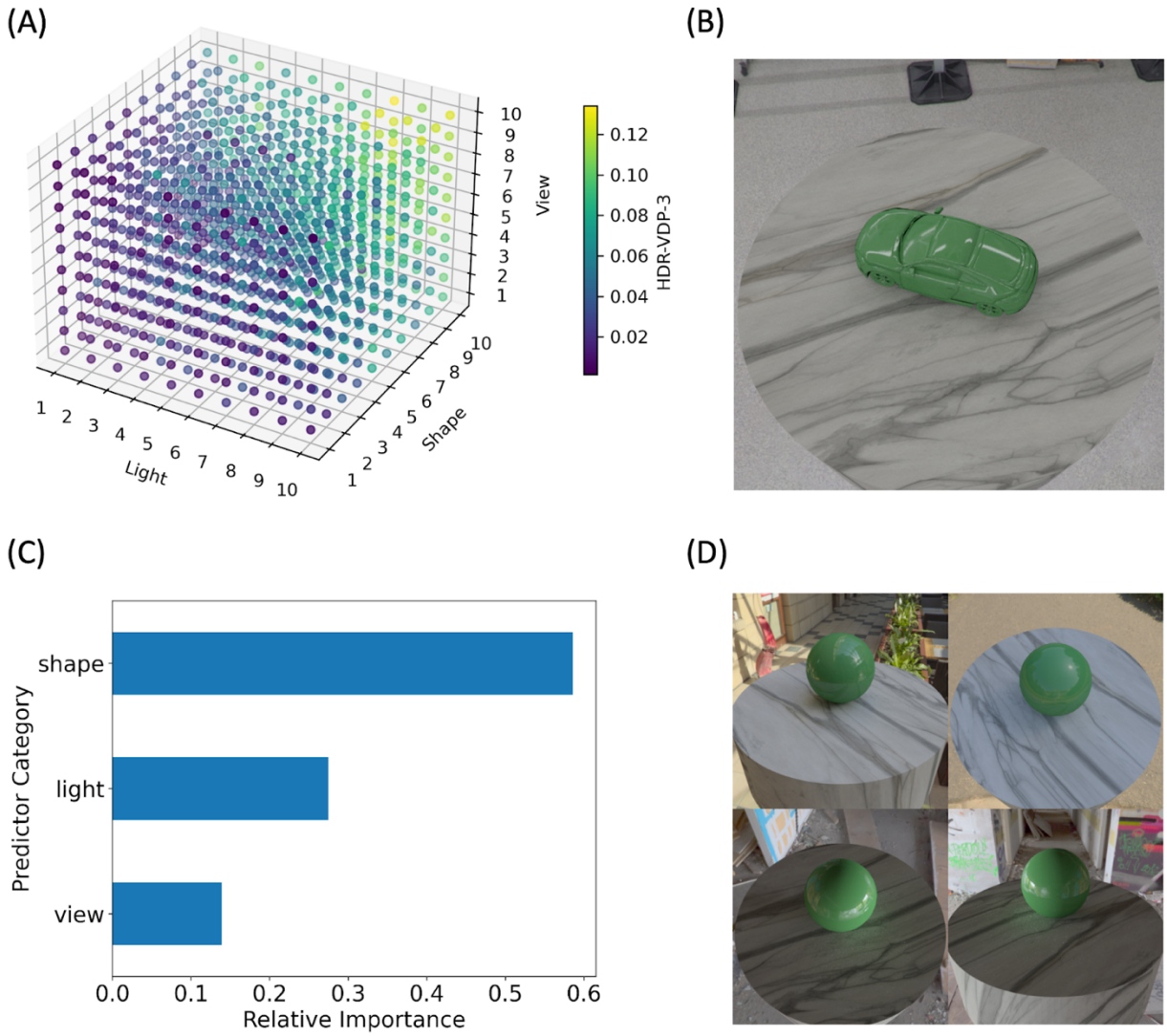

Figure 8. (A) Rank-ordered combinations of lighting, shape and viewpoint that should yield the highest sensitivity to differences in reflectance, according to an analysis of our full image set using HDR-VDP-3. (B) The scene with the combination of these scene variables that yielded the highest predicted discriminability. (C) A random forest model was used to determine the overall importance score of each variable, revealing that for our image set, object shape has the largest contribution to variance in discriminability. (D) According to this analysis, the sphere contributed the highest variation in discriminability across different lighting environments and viewpoints.

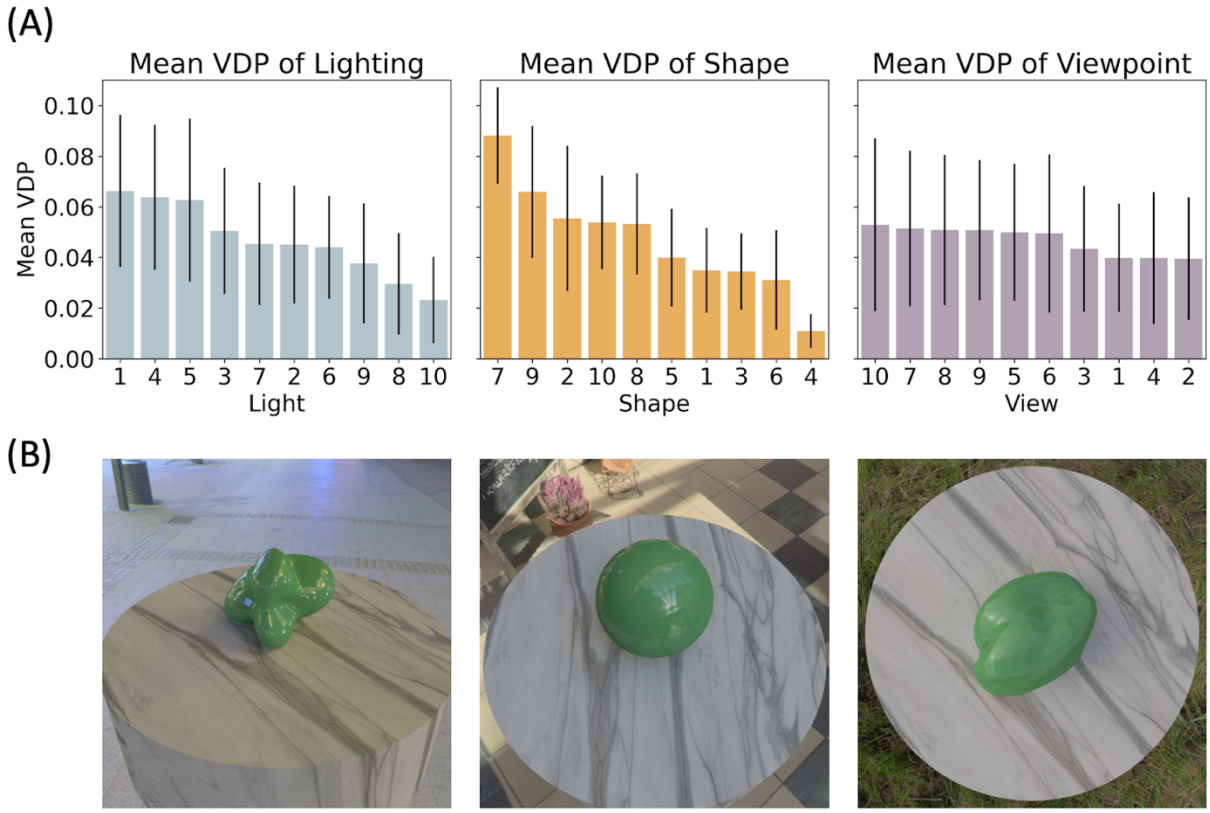

Figure 9. (A) Descriptive summaries of HDR-VDP-3 model predictions for individual lighting maps, object shapes, and viewpoints. Error bars represent standard deviation. Images corresponding to number labels for light and shape can be found in Figure 2. (B) Example scenes that included the lighting, shape, or viewpoint that produced the highest mean VDP.

3.4 Discussion

In Experiment 1, we rendered a set of images with a fixed difference in surface reflectance (roughness) for a variety of different lighting conditions, object shapes, and viewpoints. We collected human judgments of these images in an online experiment, finding that participants were highly consistent in their ranking of gloss differences across these viewing conditions. An existing model that predicts the visibility of image differences, HDR-VDP-3, was able to predict the human ranking of the gloss differences in our image set to a surprising degree – in fact, well within the range of inter-participant correlations (see Figure 4). This indicates that HDR-VDP-3 performed as well as any image-computable model could, given the variance in our data. Interestingly, in this case, similar performance can also be achieved by measuring the contrast of the specular term. In Experiment 2, HDR-VDP-3 was used to select two scenes from our full image set, representing

combinations of lighting, shape and viewpoint that represent opposing predictions of the model, leading to lower or higher sensitivity to the same difference in physical reflectance. The model predictions were validated in controlled laboratory conditions, evidenced by a significant difference in gloss sensitivity (see Figure 7). These model predictions were then used to estimate the relative contribution of specific viewing conditions to gloss sensitivity. This provides a first step towards characterizing the impact of viewing factors on gloss discrimination, so that 'reasonable bounds' on JNDs can be established.

3.4.1 Towards 'reasonable bounds' on JNDs for surface reflectance

Our study suggests that HDR-VDP-3 can predict how gloss discriminability varies across a range of viewing conditions—at least, about as well as individual participants predict one another. This lays the foundations for automatically establishing JNDs 'within reasonable bounds' for materials with particular appearances (e.g., coatings with particular formulations). Here we outline the approach and describe some of the additional open research questions that would need to be resolved to develop a working system.

The basic logic of the approach runs as follows. The JND for a given surface reflectance characteristic is defined as the smallest magnitude change in the physical BRDF of the surface that can be perceptually detected. As noted throughout this study, this value can vary due to extrinsic factors—lighting, object shape and viewpoint. Our goal is to predict the range of values the JND can take across 'reasonable' changes in viewing conditions.

As there are potentially an infinite number of possible changes to the BRDF, let us limit ourselves to the case where we wish to evaluate the JND for a *specific kind* of reflectance change. A simple case would be when varying a single parameter of an analytic BRDF model (as in our experiments), although any change that can be summarized with a single number suffices. For example, suppose a paint manufacturer wishes to determine tolerance bounds on a particular parameter of the paint formulation or manufacture process, such as the temperature at which a coating should be applied, or the duration of grinding of a particular ingredient in the paint. As long as the parameter leads to a *smooth and systematic* change in the BRDF—e.g., by changing the specular lobe—in a predictable way, then we can use images of samples with different parameter values viewed under constant conditions to estimate a JND with HDR-VDP-3 (or some other image-computable image-difference metric).

To be more precise, the assumption is that (small) changes in the manufacture parameter shift the BRDF along a specific vector in the high-dimensional space of all BRDFs (e.g., making the specular lobe broader in a particular way). The goal of determining the tolerance for the parameter then becomes the goal of determining the magnitude of that vector for which two samples can just be discriminated.

In the unusual circumstances that the BRDF will be seen exclusively under fixed viewing conditions (i.e., a single, specific shape, under fixed specific lighting, from a specific viewpoint), then it should be sufficient to image samples (e.g., render or photograph) with a few values of the parameter under those viewing conditions, and run the resulting images through the image difference predictor. The JND will be inversely proportional to the change in the image difference metric caused by a given change in the reflectance. Although in this study we used only a single pair of values of reflectance properties to estimate the impact on the image metric, in practice, using multiple samples with different values would give more robust estimates of the impact of the reflectance parameter on the image differences and therefore a more reliable estimate of the JND.

However, more typically the extrinsic view parameters (shape, lighting and viewpoint) are free to change. As extrinsic variables change, the impact of a given change in reflectance on the image also changes—making larger, more detectable image changes in some conditions, and smaller, less detectable ones in other conditions. As a result, there will be a distribution of values for the JND across extrinsic parameters. A route to estimating this distribution is to change the shape, lighting and viewpoint across a 'representative' range of conditions, and for each one, image the surface with a range of reflectance parameters. Again, by passing the resulting images through HDR-VDP-3 (or other image-difference metric), it should be possible to predict the JND for each particular combination of extrinsic parameters.

Given a distribution of JND values, an empirically informed decision can then be made about 'reasonable bounds' for the JND. For example, one might select the 95th percentile of the distribution of values, meaning that two materials within the tolerance for that reflectance-determining parameter would look indistinguishable in at least 95% of conditions.

In general, the greater the strictness of the tolerance requirements, the more different lighting, shape and viewpoint conditions would need to be evaluated to estimate the tail of the JND distribution. An alternative approach to sample the tail of the distribution more efficiently than

random sampling would be to seek out 'adversarial' combinations of lighting, shape and viewpoint that make the given differences in reflectance especially salient in the image (Bousseau et al., 2011). A particularly efficient way of achieving this in computer graphics contexts, would be to use differentiable rendering to optimize predicted visible difference between surfaces by varying lighting, shape and viewpoint. This would aid selecting tolerances based on 'worst case' scenarios. However, it is worth remembering that it is almost always possible to construct a particularly problematic combination of shape, lighting and viewpoint, and such non-generic worst-case conditions may essentially never be encountered in the real world (Freeman, 1994). Depending on the derivatives of the scene parameters that lead to very small JNDs, the 'worst case' may require extremely precise alignment of the viewpoint with the surface and light sources, for example, which are unlikely to occur except under carefully contrived circumstances.

3.4.2 Limitations and future directions

We have outlined a general approach to determining 'reasonable bounds' on JNDs for surface reflectance properties, however there are many open research topics, and additional steps to convert this outline into a working and validated system suitable for critical applications.

Here we illustrated the ability of HDR-VDP-3 to predict the relative discriminability of a change in roughness (distinctness-of-image gloss) across changes in lighting, shape and view angle. Future work should confirm that a similar approach is effective for other reflectance parameters. It would also be important to demonstrate that the approach generalizes beyond computer graphics to real-world conditions.

Previous studies have identified many factors that affect gloss constancy, including environmental factors such as illumination (Adams et al., 2018; Fleming et al., 2003; Ged et al., 2020; Ho et al., 2006; Morimoto et al., 2023b; Motoyoshi & Matoba, 2012; Olkkonen & Brainard, 2011; Pont & te Pas, 2006; Wendt & Faul, 2017), viewpoint (Ho et al., 2007), and intrinsic surface factors such as shape (Berzhanskaya et al., 2005; Morimoto et al., 2023b; Nishida & Shinya, 1998; Olkkonen & Brainard, 2011; Tiedemann, 2018), and diffuse reflectance (Morimoto et al., 2023b; Vladusich, 2013; Wendt et al., 2010; Wendt & Faul, 2018). Our approach to investigating these factors was, like most previous studies, to sample a rather arbitrary selection of shapes and illuminations and a limited range of view angles. A more thorough and systematic exploration of the impact of shape, lighting and viewpoint would be beneficial. This is challenging as the space

of possible shapes and illuminations is practically infinite. One approach would be to consider parametric spaces of lighting and shape, for example using spherical harmonics decompositions (Mazzarella et al., 2014; Mury et al., 2009; Norman et al., 2020; Ramamoorthi & Hanrahan, 2001).

Additionally, gloss constancy is affected by how surfaces are presented to participants; for example, studies have demonstrated that the presence of dynamic motion (Doerschner et al., 2011; Ferwerda & Padhye, 2021; Shiwen et al., 2023; Wendt et al., 2010; Wendt & Faul, 2018), Fresnel effects (Faul, 2019, 2021), disparity (Wendt et al., 2010), dynamic range (Doerschner et al., 2010), and the particular tone mapping operator used in the rendering process (Adams et al., 2018) are also important factors. These various factors also deserve consideration, and future work could explore the extent to which they impact upon the predictions of HDR-VDP-3, following a similar experimental framework.

Rather than assess how viewing conditions affect many differences in reflectance, as is typically done in studies of gloss constancy, here we assessed how viewing conditions affect a single, fixed difference in reflectance. This approach ensured that the observed differences in gloss perception were not confounded by variations in physical reflectance. Marlow and Anderson (2013) took a similar approach, manipulating surface geometry and the structure of the light field to assess their relative contributions to perceived gloss for a single value of physical reflectance. However, future work should test the assumption that changes in a reflectance parameter lead to proportional changes in the detectability of image differences (as predicted by HDR-VDP-3). It could be that for some reflectance characteristics, there is a nonlinear relationship between changes in the parameter value and changes in the image. The key assumption here is that for small changes, i.e., close to the JND, image changes are approximately linearly related to the reflectance characteristics. While this assumption seems reasonable, it should be tested. Moreover, while we have shown that there is a systematic relationship between HDR-VDP-3 and gloss discrimination, this falls short of explicitly estimating a specific value for the JND from the image difference metric. Additional work is necessary to identify quantitative mappings from HDR-VDP-3 to variations in reflectance parameters, so that the JND can be expressed in terms of units of change of the reflectance parameter. Another important limitation of our study is that it considers discrimination across distinct images of the same object. For many practical applications, however, the key question is whether two juxtaposed surfaces (e.g., two doors of a car), or two neighboring parts of the same surface have the same appearance. Future studies should investigate the

detectability of abrupt spatial transitions in reflectance (as in Figure 1), as the JNDs for these may be substantially lower than suggested by our findings.

Finally, in the long run, it will also be necessary to generalize our approach to asymmetric comparisons (perhaps involving dynamic scenes and physical surfaces), where the difference in reflectance is confounded by differences in viewing conditions or surface color. Under these conditions, it is clear that mere image difference metrics will not capture differences in surface appearance. An image-computable model that can evaluate visual equivalence would be a useful starting point for overcoming this limitation (e.g., see Ramanarayanan et al., 2007). Our recent work (Morimoto et al., 2023b) explored how object shape and lighting environment impacted the ability to make asymmetric comparisons of gloss across different lighting conditions and object shapes. Although substantial failures of gloss constancy were found in these experiments, participants were highly consistent in their deviations from physical ground truth. This finding agrees with the high inter-participant correlation obtained in Experiment 1 of the current study, which was also conducted online in uncontrolled viewing conditions. Apparently, whether the comparisons are symmetric or asymmetric, participants have little trouble consistently judging differences in gloss, or in making consistent adjustments to match gloss levels under different viewing conditions. The current study contributes to a growing body of literature on gloss perception, demonstrating the remarkable consistency of judgments across various viewing conditions. This means that there are good grounds for thinking that a quantitative, image-based approach can be used to predict discriminability of gloss and other surface reflectance characteristics.

3.5 Conclusions

Our study demonstrates the potential of using image metrics to predict gloss discrimination across a range of viewing conditions, challenging prior assumptions about the complexity of this task. While our findings show that judgments of gloss can vary under different viewing conditions, they also reveal a surprising degree of precision in how these judgments are made. These insights not only advance our understanding of material appearance but also point to potential practical applications in industrial quality control and computer graphics.

CHAPTER 4: SCALE AMBIGUITIES IN MATERIAL RECOGNITION

A similar version of this chapter has been published as:
Cheeseman, J. R., Fleming, R. W., & Schmidt, F. (2022). Scale ambiguities in material recognition. iScience, 25(3), 103970. https://doi.org/10.1016/j.isci.2022.103970

Many natural materials have complex, multi-scale structures. Consequently, the inferred identity of a surface can vary with the assumed spatial scale of the scene: a plowed field seen from afar can resemble corduroy seen up close. We investigated this 'material-scale ambiguity' using 87 photographs of diverse materials (e.g., water, sand, stone, metal, wood). Across two experiments, separate groups of participants (N = 72 adults) provided judgments of the material category depicted in each image, either with or without manipulations of apparent distance (by verbal instructions, or adding objects of familiar size). Our results demonstrate that these manipulations can cause identical images to be assigned to completely different material categories, depending on the assumed scale. Under challenging conditions, therefore, the categorization of materials is susceptible to simple manipulations of apparent distance, revealing a striking example of top-down effects in the interpretation of image features. This suggests that material categorization is governed not only by bottom-up interpretations of image features and the physical properties of reflecting surfaces, but also by top-down assumptions about the observer's distance to the surface. Much as familiar objects are associated with canonical sizes, materials are evidently associated with canonical scales, and are recognized through completely different cues at different distances.

4.1 Introduction

In most circumstances, observers can reliably and efficiently recognize the materials that surfaces are made of (Sharan et al., 2014). Here, we find, however, that under certain viewing conditions, surfaces made of entirely different materials can produce similar image features, leading to confusions. For example, the mottled surfaces in Figure 1 are photographs of (a) sea water and (b) marble. When asked to estimate the distances in these images, we are likely to guess that the sea water is much farther from the camera than the marble surface. If we instead imagine that (b) is sea water and (a) is marble, this reverses the difference in assumed distance; that is, sea water only appears this way from a relatively large distance. This 'material-scale ambiguity' suggests that our knowledge of the typical appearance of materials can be used to constrain the range of plausible viewing distances that could produce these images (Fleming, 2014), much like objects of familiar size can indicate depth and distance (Hubbard et al., 1989; Konkle & Oliva, 2011a, 2012b, 2012a). Conversely, if material identity and spatial scale are mutually constraining, this also raises the possibility that judgments of material identity may be influenced in a top-down way by the assumed viewing distance to objects and surfaces.

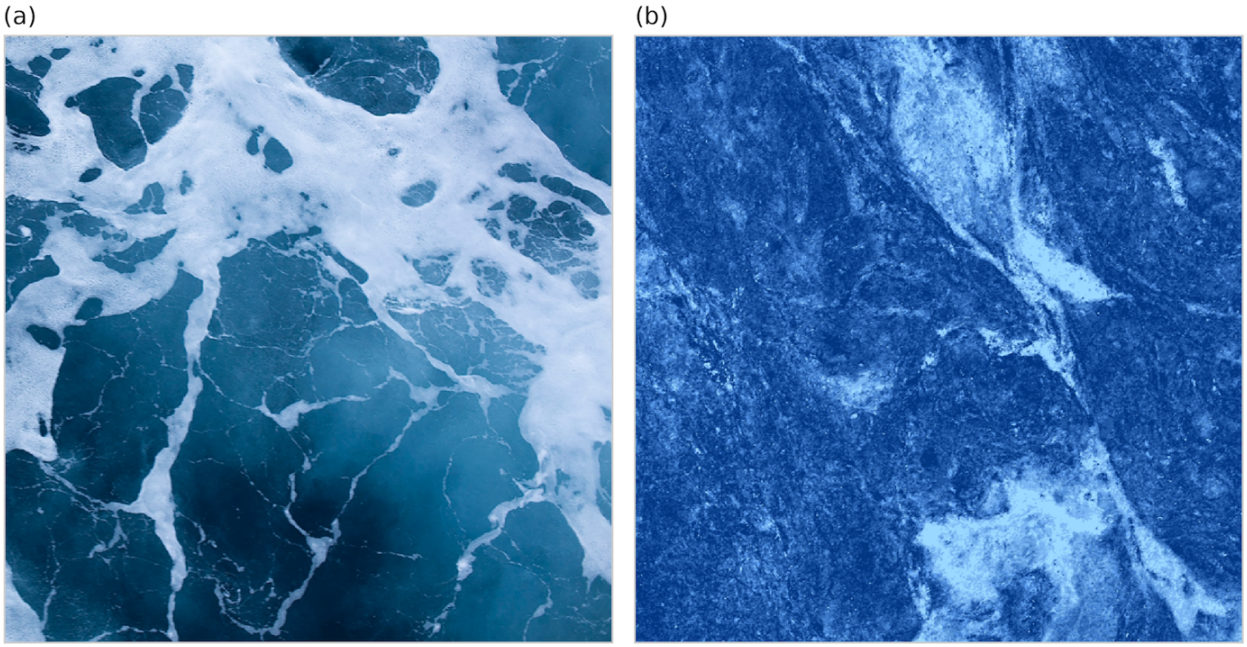

Figure 1. Demonstration of material and scale dependency. When viewing distance is ambiguous, the material properties depicted in these photographs of (a) sea water and (b) marble are confusable. Note that the relative difference in assumed viewing distance (i.e., the ocean water is farther than the marble surface) reverses if one imagines that the labels are switched.

Material appearance is intimately connected to the spatial scale of geometrical and optical processes at which light interacts with surfaces (Pont & Koenderink, 2002, 2005). A given material can change in appearance dramatically depending on the viewing distance. Individual water droplets seen close-up are visibly transparent with small punctate highlights, yet when seen *en masse* from afar, droplets blur together into a white haze. Tiny scratches on the surface of brushed metal appear crisp and mirror-like when viewed through a microscope, yet the ensemble appearance at normal view distances is of a silky, anisotropic gloss (Raymond et al., 2016). We reasoned that given the complex, multiscale nature of many materials, image features that emerge at a particular scale or viewing distance can sometimes be deceptive about material properties. For example, anyone who has touched the stems and spikelets of barley knows that they are harsh to touch, yet when seen from afar, a field of barley can have a distinctly 'soft' appearance. This opens the possibility of mis-categorizing materials as a function of the interpreted or assumed viewing distance.

Here, we investigated how contextual information about viewing distance influences material recognition. We did not seek to test the claim that *all* images of materials suffer from material-scale ambiguities, but rather the more restricted claim that there are certain categories of materials that are confusable when imaged from different viewing distances—and thus that certain images can exhibit this effect. We explicitly set out to find such images and investigate the patterns of confusions they cause. If the same image can be interpreted as completely different material categories depending on the viewer's assumptions about view distance, this has important implications for theories of visual categorization, which typically assume a mapping between particular image features and material categories (Bell et al., 2015; Fleming et al., 2013; Sharan et al., 2013).

In Experiment 1 we presented participants with (potentially) ambiguous photographs of different materials, and asked them to identify the surface material and estimate the distance between the camera and surface. We reasoned that depending on their interpretation of the surface material, they should judge the viewing distance as relatively small or large. In this way, we could discover not only the typical interpretations of these images, but also the assumed spatial scales associated with these interpretations. We find that indeed, the surfaces depicted in certain images can be identified as completely different materials, depending on whether the observer spontaneously interpreted the surface as seen from near or far. We then sought to test whether we

could directly manipulate material recognition by imposing different scale interpretations. In Experiment 2a separate groups of participants were asked to identify the depicted material when instructed to imagine the camera as either very close or very far from the surface plane, and in Experiment 2b participants judge modified sets of images that included familiar objects (insect or airplane) to provide contextual information about spatial scale. Our aim was to identify a subset of images for which these manipulations are effective rather than assessing the level of general ambiguity across all images. We hypothesized that if material recognition depends on assumptions about egocentric distance, the pattern of responses in each experiment would depend on the particular instructions or familiar objects provided to participants.

4.2 Experiment 1: Unbiased judgments of distance and material

We first sought to measure baseline judgments of distance and material for our stimulus set in the absence of explicit distance or scale information. Participants were presented with a series of potentially ambiguous photographs and asked to estimate the distance between the camera and depicted surface plane, and to identify the material category. The unbiased responses from this experiment served as a baseline of comparison with Experiment 2, in which judgments of distance and material were manipulated with contextual information.

4.2.1 Participants

A convenience sample of 24 students (14 women, 10 men; M_{age} = 24.3 years, SD_{age} = 3.3 years) participated in the experiment for 8€ per hour. All experimental procedures were approved by the Justus Liebig University Giessen Psychology Department Ethics Board and conformed with the guidelines of the American Psychological Association (Version 2017) and the Declaration of Helsinki (Version 2013). Only one previous study (Wiebel et al., 2013) has estimated material categorization accuracy under conditions similar to the current study, that is, with unlimited stimulus presentation time and no instruction to respond rapidly. All of the participants in that study achieved categorization accuracies (*Mdn* = 87%) that exceeded chance-level accuracy (25%). According to the 95% confidence interval (CI) obtained from a one sample binomial test, the probability that the observed median categorization accuracy exceeds chance-level is between 85% and 100%, $p < 0.001$, $g = 0.62$. Although this indicates that material categorization is quite

robust under normal conditions, given that our stimuli were selected for ambiguity, a larger sample (N = 24) was preferred for Experiment 1.

4.2.2 Stimuli

Potentially ambiguous photographs were collected from various internet sources. The selection criteria were as follows: (i) they did not contain objects, but clearly depicted a surface or scene, (ii) they were determined by the experimenter to be potentially ambiguous in the sense that at least two distinct material categories could describe the surface or scene, and (iii) these material categories could be associated with different spatial scales. Three observers (including two authors) screened hundreds of images on the internet and selected those that met the criteria, resulting in 87 images (Figure 2a). Ground truth category labels and copyright information for these images are listed in the Supplementary Material (Figure S1 and Table S1). Ground truth category labels for these images were determined by the original publisher. Images for which print copyright permissions could not be obtained are here replaced by synthetic textures derived from the originals (see Risser, 2020). All images were rescaled to 600 × 600 pixels.

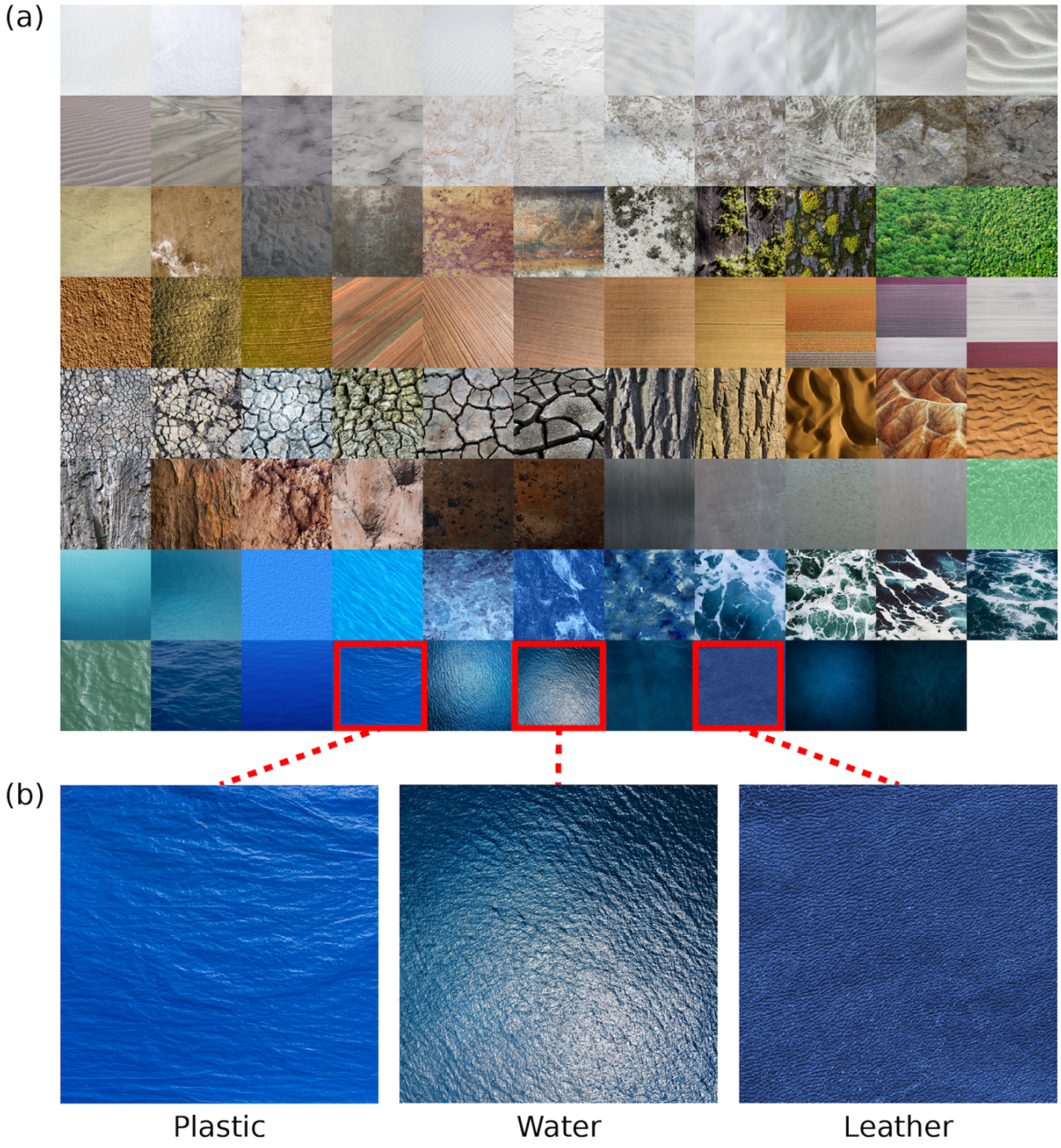

Figure 2. Stimuli for Experiments 1 and 2. The full set of 87 images (a) is arranged by approximate visual similarity. Similar image features (e.g., color, texture) can arise from different material categories (b), especially when spatial scale is ambiguous.

4.2.3 Procedure

The experiment was controlled by a Dell Precision T3500 desktop computer running Windows 10 and Qualtrics software (v02.2019). Stimulus images were displayed using an Eizo ColorEdge CG277 self-calibrating LCD monitor (68.4 cm diagonal; 2560 × 1440 resolution). Participants were seated in a dark room approximately 50 cm from the monitor, which displayed the images against a uniform grey background. At this viewing distance, the 600 × 600 pixel images subtended approximately 16 degrees of visual angle. On each trial, participants viewed a single, randomly-selected image and judged the depicted material, followed by the distance between the camera and surface plane. Different kinds of responses were collected in separate blocks of trials. In the first block (*free-response*), participants described the surface material as precisely as possible by typing a written response. In the second block (*multiple-choice*), participants identified surface materials by selecting one category from 26 options (see Experiment 1 Multiple Choice Instructions in the Supplementary Material). Participants also provided a confidence rating (from 1 to 7) for judgments of material. In the third block (*distance estimation*), participants estimated the distance between the camera and surface plane by selecting an appropriate unit of measurement (micrometers, millimeters, centimeters, meters, or kilometers) and metric value (e.g., 8 cm, or 1 km). English translations of the original German instructions for all experiments are provided in the Supplementary Material.

4.2.4 Results

Our first main finding is that overall, participants were quite good at identifying the correct categories of materials. All participants achieved categorization accuracies (*Mdn* = 33%) that exceeded chance-level accuracy (4%). According to the 95% confidence interval (CI) obtained from a one sample binomial test, the probability that the observed median categorization accuracy exceeds chance-level is between 96% and 100%, $p < 0.001$, $g = 0.29$. Additionally, there was a strong rank correlation between the frequency of common terms represented in the free-response and multiple-choice data ($r_s = 0.79$, $p < 0.001$; see Free-Response Transformations in the Supplementary Material for details). These findings show that although the images were selected for their ambiguity, participants' responses were systematic and consistent across tasks. Thus, deviations from ground truth were likely purposeful. We next sought to answer whether such errors were systematically related to the assumed scale (or distance) of the scene.

To assess this, we divided the multiple-choice responses for each image into two *distance groups*: (i) responses from participants who selected distance units indicating a relatively small scale (*near group*: micrometer, millimeter, or centimeter), and (ii) responses from participants who selected distance units that indicated a relatively large scale (*far group*: meter or kilometer). Material category confusions can occur in two directions (near → far vs. far → near). Some confusions between categories are likely to be distance-independent simply because the materials are similar in appearance (e.g., confusing stone and concrete irrespective of distance), and will therefore tend to be symmetrical in each direction. Instead, we focus on *asymmetrical* patterns of confusion that reflect a *distance-dependent* material ambiguity. That is, are there images that are assigned to one material (e.g., bark) when seen as close-up, but which are assigned to another material (e.g., stone) when seen as far away, *but not vice versa*?

The relative frequency of directional category confusions is indicated by shaded lines in Figure 3a. The observed asymmetry between the pattern of blue lines (near → far confusions) and red lines (far → near confusions) illustrates that material category confusions vary systematically with assumed viewing distance. For clarity, cases where the two distance groups selected the same category are omitted, and only categories corresponding to ground truth image labels are shown (for details, see Relative Frequencies in the Supplementary Material).

Figure 3b displays the same relative frequencies (here normalized between 0 and 1) plotted as cells in a matrix, with responses of the near and far groups represented on separate axes. Category confusions that occur in the near → far direction are located in the lower triangle of the matrix (blue-shaded cells), and confusions in the far → near direction are located in the upper triangle (red-shaded cells). Instances where the two distance groups selected the same category are plotted on the diagonal of the matrix (purple-shaded cells). Directional asymmetries can be compared by examining category pairs on opposite sides of the diagonal. For example, we find that surfaces identified as *bark* at smaller apparent distances tend to be confused with *stone* at larger apparent distances (frequency = 0.4; blue square in Fig 3b). However, the reverse is not true: when a surface is seen as *bark* at larger apparent distances, it is less often confused with *stone* at smaller apparent distances (frequency = 0.1; yellow square in Fig 3b). Fig 3c shows one such image, and for comparison Fig 3d shows an image with a distance-dependent ambiguity in the opposite direction (water seen from far looks like plastic seen from near).

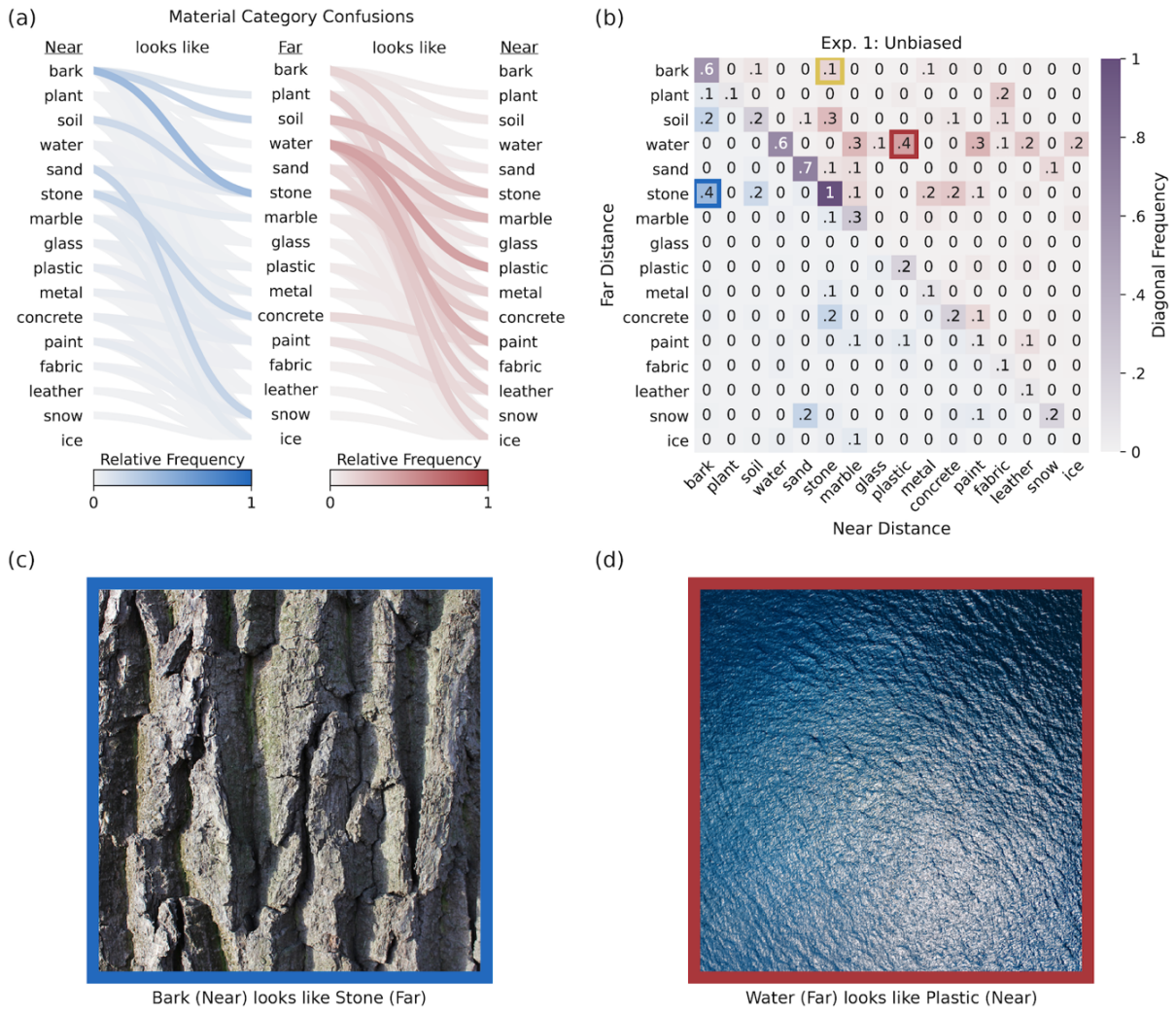

Figure 3. Material category confusions exhibiting material-scale ambiguity. The relative frequency of confusions between pairs of categories is indicated by shaded connecting lines in (a). Categories selected with smaller distance units (micrometer, millimeter, or centimeter) are defined as near, while categories selected with larger distance units (meter or kilometer) are defined as far. Confusions in the near → far direction (blue-shaded lines) differ systematically from confusions in the far → near direction (red-shaded lines). The same asymmetry can be seen when (normalized) relative frequencies are plotted in a matrix format (b), with distance-dependent responses on separate axes. Category pairs above and below the negative diagonal indicate the relative frequency of distance-dependent confusions; distance-independent responses are shown on the diagonal. The top-ranked images for the most frequent near → far confusion (c) and the most frequent far → near confusion (d) illustrate different material-scale ambiguities.

The matrix representation of category confusions also provided a simple way to measure the overall magnitude of material-scale ambiguity for this set of images. This is accomplished by calculating the Root Mean Squared Error (RMSE) between corresponding frequencies in the lower and upper triangles of the matrix (e.g., blue cells vs. red cells in Fig 3b, excluding the diagonal). This metric, which we call the *Material-Scale Ambiguity* (MSA), represents the amount of directional asymmetry in category confusions. MSA can also be calculated for individual images, which allows us to identify the subset of images that were particularly effective in producing distance-dependent category confusions. MSA for individual images was then compared with asymmetries that occur by chance, by calculating the mean MSA across 1000 random permutations of the distance units associated with each response (i.e., category confusions are shuffled in each direction). Figure 4a shows images ranked by MSA, with the corresponding chance-level MSA drawn in lilac. Sixty-one (out of 87) images produced MSA values above chance. Figure 4b shows that the mean MSA across all images (teal; $M = 1.4$, 95% CI [1.3, 1.51]) was significantly greater than chance-level MSA (lilac; $M = 0.95$, 95% CI [0.88, 1.02]; Wilcoxon $T = 850$, $p < 0.001$, $d = 0.52$). This result indicates that, as a whole, the image set produces systematic distance-dependent category confusions, but also that there is a subset of images that are particularly ambiguous in this way. For example, the image shown in Figure 4c was reliably seen as *bark* regardless of apparent distance, whereas the image shown in Figure 4d produced confusions between *plant* (seen from a relatively large distance) and *carpet* (seen from a relatively small distance). The material categories associated with relatively small or large assumed distances are shown in word clouds below each image, where category frequency is indicated by word size.

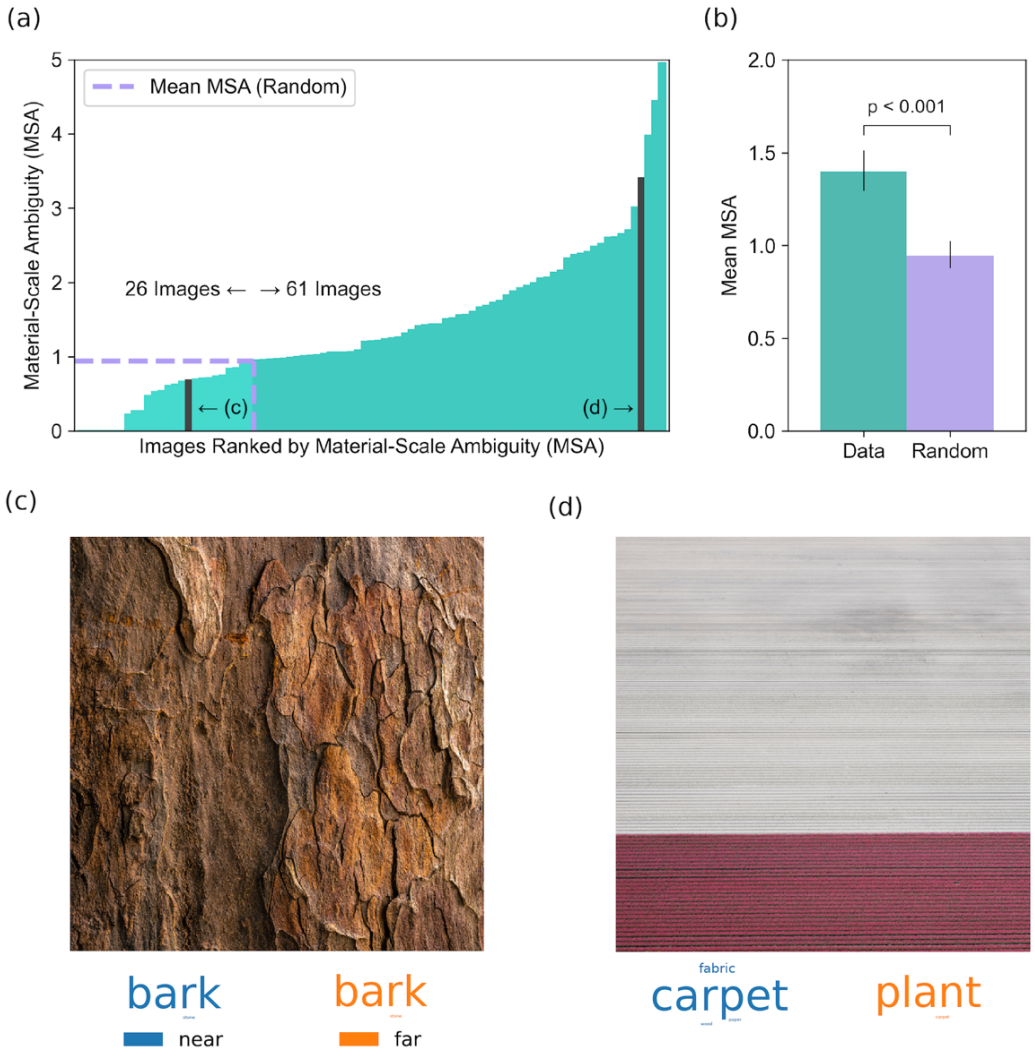

Figure 4. Material-Scale Ambiguity (MSA) calculated for individual images. Images can be ranked by MSA (a) along with the mean MSA resulting from random permutation of distance units drawn in lilac. The mean MSA across images (b) for the original responses (teal) and random permutation of distance units (lilac). A Wilcoxon signed-rank test indicated that the observed difference in MSA is significant. Error bars represent 95% confidence intervals. An image with low MSA (c) is reliably seen as one material (bark) regardless of assumed distance. Word clouds of responses (size weighted by frequency) associated with different distance assumptions are shown below. An image with high MSA (d) is reliably seen as different materials (carpet vs. plant) depending on the assumed distance (far vs. near).

Taken together, the results of this experiment show that, without contextual information to specify viewing distance, certain images exhibit a striking distance-dependent ambiguity about the category of material. The material category labels assigned to these images covary with whether the surfaces are interpreted as near to or far from the camera. To test whether the assumed viewing distance plays a *causal* role in driving this effect, we next sought to directly manipulate assumed viewing distance.

4.3 Experiment 2: Biased judgments of distance and material

To test whether manipulating assumed viewing distance influences judgments of material, we presented the same stimuli to four additional groups of naïve participants. Assumed viewing distance was manipulated by instructing each group to imagine a small or large distance to the surface plane (Experiment 2A), or by presenting modified images that featured two different familiar objects to indicate different spatial scales (Experiment 2B).

4.3.1 Participants

Four groups of 12 students (48 total) participated in the experiment. Group 1 (9 women, 3 men; M_{age} = 24.5 years, SD_{age} = 3.01 years) and Group 2 (9 women, 3 men; M_{age} = 22.5 years, SD_{age} = 3.64 years) participated in Experiment 2A. Group 3 (4 males and 8 females; M_{age} = 24.5 years, SD_{age} = 4.01 years) and Group 4 (4 males and 8 females; M_{age} = 22.83 years, SD_{age} = 2.37 years) participated in Experiment 2B.

4.3.2 Stimuli

The same set of images presented in Experiment 1 was shown to each group of participants in Experiment 2A. For Experiment 2B, we created two sets of images from the original set, each of which featured a large or small familiar object (3D models of an airplane or a hornet) that provided contextual information about the spatial scale of the scene (for examples, see Figure S2 in the Supplementary Material). The objects were digitally inserted into the images with 3D modelling software that simulated lighting direction and cast shadows. The 3D models ("Boeing 787 8(1)" by turbosquid.com/companion_3d, published under Editorial Use license; "Hornet" by blendswap.com/Misfit410, published under CC-BY license), were inserted into the original images using Blender 2.79 (Stichting Blender Foundation, Amsterdam, NL). In these virtual

scenes, the original images were arranged parallel to the ground plane, while the objects were positioned slightly above. The camera was positioned perpendicular to the ground plane with a point light source slightly offset from the center of the ground plane, casting a shadow of the object onto the ground plane image.

4.3.3 Procedure

Randomly selected images from the corresponding set were presented to participants in the same viewing conditions as in Experiment 1. For each image, participants identified the depicted surface material (free-response and multiple-choice, in separate blocks of trials) and indicated their confidence (1-7). In Experiment 2A, Group 1 was instructed to imagine that the camera was "very far" from the surface plane, and Group 2 was instructed to imagine that the camera was "very near" to the surface plane. These instructions were intended to ensure that the imagined distance range could vary with the interpretations for each image. In Experiment 2B, Group 2 was presented with images that featured a large familiar object (airplane), and Group 3 was presented with images that featured a small familiar object (hornet).

4.3.4 Results

Material-Scale Ambiguity (MSA) was calculated separately for Experiment 2A and Experiment 2B. To test whether the observed asymmetries in category confusions were statistically significant, we compared them to chance-level MSA calculated from 1000 random permutations of group membership in each experiment. Images ranked by MSA for all experiments are shown in Figure 5a, along with chance-level MSA in lilac. Figure 5b shows the mean MSA across images for the original responses (teal) and chance-level MSA (lilac) separately for all experiments. In Experiment 2A, values of MSA for the original responses (M = 1.48, 95% CI [1.37, 1.6]) were significantly greater than chance-level MSA (M = 1.15, 95% CI [1.07, 1.25]; Wilcoxon T = 799, $p < 0.001$, d = 0.49). Magnitudes of MSA produced by manipulation of assumed viewing distance with instructions (Experiment 2A) are therefore similar to those produced by unbiased judgments of material and distance (Experiment 1). Although familiar objects produced above-chance MSA in 46 images (Experiment 2B), this manipulation was less effective overall (Wilcoxon T = 1812, p = 0.80, d = 0.07). Interestingly, confidence ratings in Experiment 2B (M = 4.8, SD = 0.71) were significantly lower than those in Experiment 2A (M = 4.98, SD = 0.5; Mann–Whitney U = 192.5,

$p = 0.025$, $d = 0.25$). It is also noteworthy that in general, the same images produced high MSA across experiments. Values of MSA in Experiment 1 are significantly correlated in rank with those obtained in Experiment 2A ($r_s = .43$, $p < .001$) and Experiment 2B ($r_s = .28$, $p = .009$); the rank correlation between Experiment 2A and 2B was not significant. These results suggest that, compared to instructions, the effectiveness of familiar objects is more dependent on whether the objects appear plausible within a given scene.

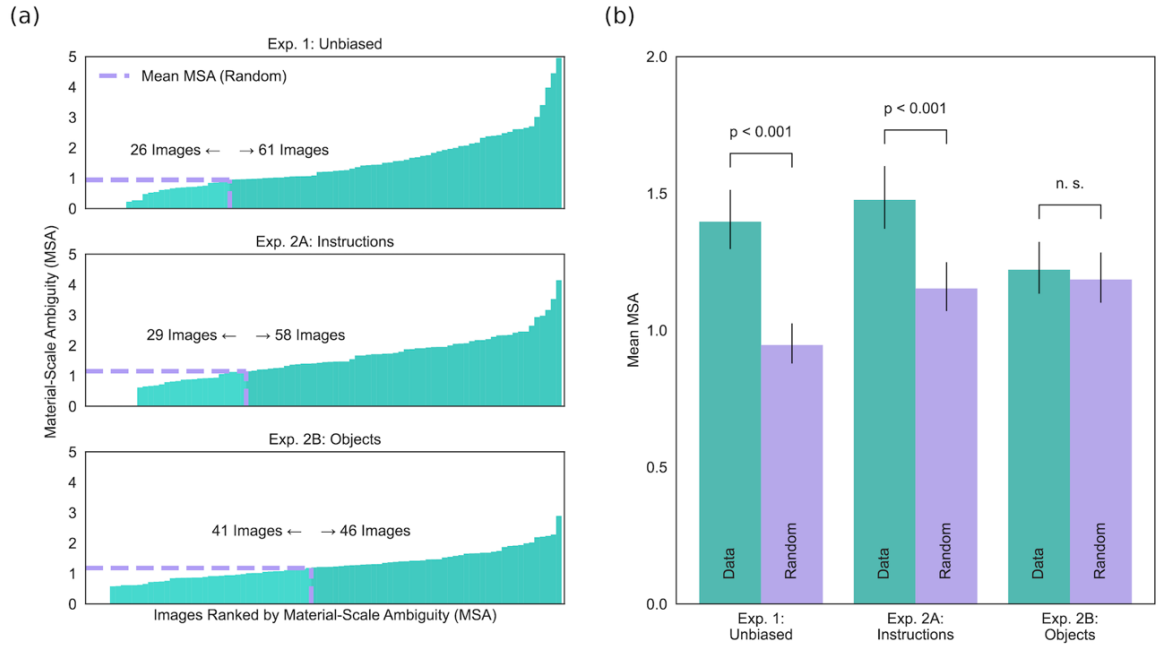

Figure 5. Material-Scale Ambiguity (MSA) calculated separately for each experiment. (a) Images ranked by MSA. The mean MSA resulting from random permutation of distance units is drawn in orange. (b) Mean MSA across images for the original responses is shown in blue, while mean MSA obtained from random permutation of distance units (Experiment 1) or group membership (Experiment 2) is shown in orange. The observed difference in MSA is significant for Experiments 1 (unbiased) and 2A (instructions). Error bars represent 95% confidence intervals.

Ranking our images by the mean MSA across experiments provides the clearest summary of our findings (Figure 6). Images with low MSA (bottom row) are immune to manipulations of viewing distance. Yet, when MSA is high (top row), the assumed distance between the camera and the surface plane determines which categories are selected.

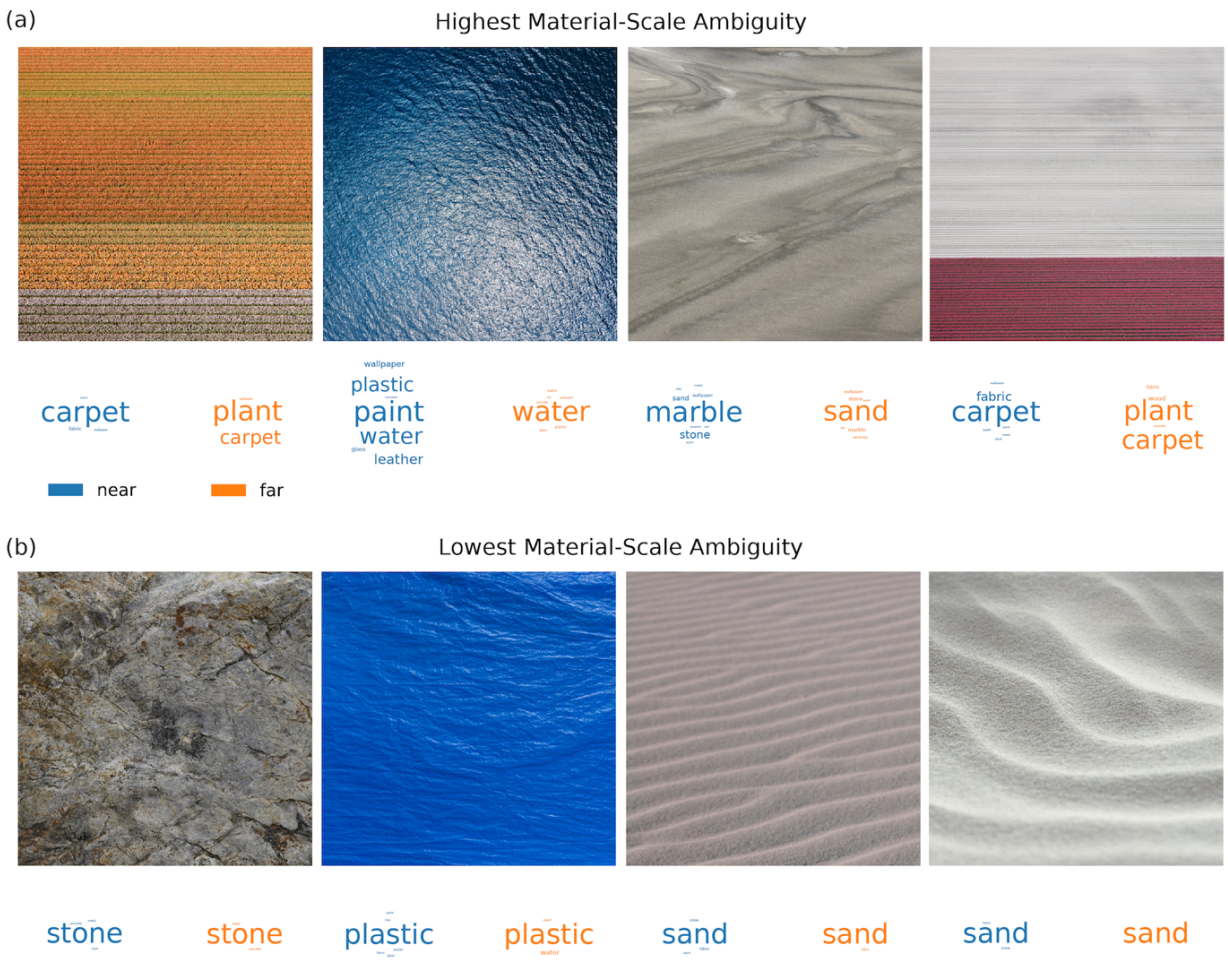

Figure 6. Images from our set with the highest (a) and lowest (b) MSA across experiments. Word clouds below each image show the distinct material categories (size weighted by frequency) selected by participants who assume relatively small or large distances to the surface plane (blue and orange data, respectively).

4.4 Discussion

Surface material appearance is characterized by complex physical and optical structures at multiple spatial scales. As materials can have strikingly different appearances at different viewing distances, we reasoned that for certain images, the appearance of a material at one distance could potentially resemble the appearance of another material at a different distance, leading to different category assignments. Here, we have established that such images exist and shown how the assumed view distance can radically alter how they are interpreted.

In Experiment 1, we found spontaneous confusions about material category co-occurring with substantial differences in the assumed viewing distance. In Experiment 2, we showed that both instructions and the insertion of familiar objects indicating different physical scales can alter material classification in a subset of images. An important caveat is that our conclusions are necessarily dependent on the images in our stimulus set. Our goal was to identify specific images that are susceptible to such manipulations, rather than to draw conclusions about the frequency of such ambiguities across all natural images. Although we rarely make such confusions about material categories in everyday life, the very existence of such images demonstrates a substantial top-down effect in visual recognition, in which context can radically alter the interpretation of identical images.

Our finding that distance assumptions can bias material categorization supports the notion that materials have *canonical scales* that constrain diagnostic image cues. Much as certain viewpoints and sizes of objects can be considered canonical (Konkle & Oliva, 2011b; Palmer et al., 1981), image cues associated with particular material categories may be more likely to arise at *typical* viewing distances (Fleming, 2014). Conversely, when a material is viewed from an *atypical* distance (e.g., in macro or aerial photography; see De Giuli, 2018), the cues may resemble other materials, resulting in material-scale ambiguities. To date, theories of material perception and appearance have tended to ignore their scale-specificity; future work on material recognition should consider how material appearance and relevant image features vary with distance.

4.4.1 Image cues and material-scale ambiguity

Which image characteristics cause material-scale ambiguity, rather than the general kind of ambiguity that is unrelated to apparent scale? Are there any features that predict the ambiguity across classes, or is the phenomenon driven by different cues for different materials? The images in Figure 6 hint at some compositional elements that might contribute. When the camera is (approximately) perpendicular to the surface plane, this limits depth cues (e.g., texture and optical blur gradients; Cutting & Vishton, 1995; Hafri et al., 2021; Sedgwick, 2005; Vishwanath, 2010) presumably increasing their ambiguity. The absence of visible object boundaries likely also plays an important role, although many such unbounded texture images are unambiguous (Fleming et al., 2013; Wiebel et al., 2013). Moreover, if an image contains scale-invariant structure (e.g., fractal or self-similar texture; see Brachmann and Redies, 2017), such image features provide little

information about viewing distance, whereas image features that do vary with viewing distance may be more diagnostic of certain material categories.

If there are certain cues whose presence or absence tends to increase material-scale ambiguity, then it should be possible to predict the degree of ambiguity across material classes based on these features. In a first attempt to test this, we identified five appearance characteristics that could plausibly be relevant to material-scale ambiguity (atmospheric blue tint, blurriness, direct lighting, surface gloss, and surface slant) and obtained human ratings for all images in our dataset, along with one image metric (self-similarity). However, a classifier based on these feature values failed to accurately classify high-MSA vs. low-MSA images, although we did find that slanted surfaces and natural materials were associated with larger apparent distances (for details, see Exploratory Analyses in the Supplementary Information). These results suggest that the image statistics that would predict MSA for a given pair of categories (e.g., bark and stone) likely differ from those needed for another pair of categories (e.g., water and plastic).

Considering the great diversity of appearances associated with different material classes, it is unlikely that a single image characteristic exists that can determine whether a given image exhibits material-scale ambiguity across all possible category confusions. Yet, for a given material category, there may be certain cues that are particularly important. For example, Sawayama et al. (2017) have shown that participants can judge the fineness of fibrous textures (like human hair), even when individual texture elements are smaller than the resolution of the imaging system. The main cue driving the super-resolution judgments in their displays was contrast: lower contrast patterns appeared to contain finer elements because of averaging of texture elements within each pixel. If the visual system relies on such inferences beyond the image data to infer specific material properties, it is necessarily open to scale-related ambiguities. Our results demonstrate that material recognition can depend on the 'beholder's share' rather than image statistics per se (Gombrich, 1961).

4.4.2 Limitations and future directions

The conclusions we can draw from this study are necessarily limited to specific images that exhibit material-scale ambiguity. Further investigation of this effect—and how it relates to material perception in general—will require innovations that lower the difficulty of creating or discovering such images. One approach to probing the characteristics of such images is to blend the statistics

of different images that produce the same kind of material-scale ambiguity. For example, Figure 7a displays an image of *bark* that was often confused with *stone*, together with two synthetic 'lookalikes' generated by a deep convolutional neural network used for image style transfer (Gatys et al., 2016). In this example, the 'styles' are other images of *bark* that were often confused with *stone* (i.e., the ground truth category is the same in the style and transfer images). A variation of this approach could involve mixing images that differ in ground truth category, but which produce the same kind of material-scale ambiguity (e.g., the images of water and leather in Figure 7b). An intriguing alternative approach would be to ask artists to create ambiguous images and analyze the techniques they use to depict material appearance (Di Cicco et al., 2021; Lin et al., 2021). Such methods may make it possible to identify key features that alter perceived scale or material appearance.

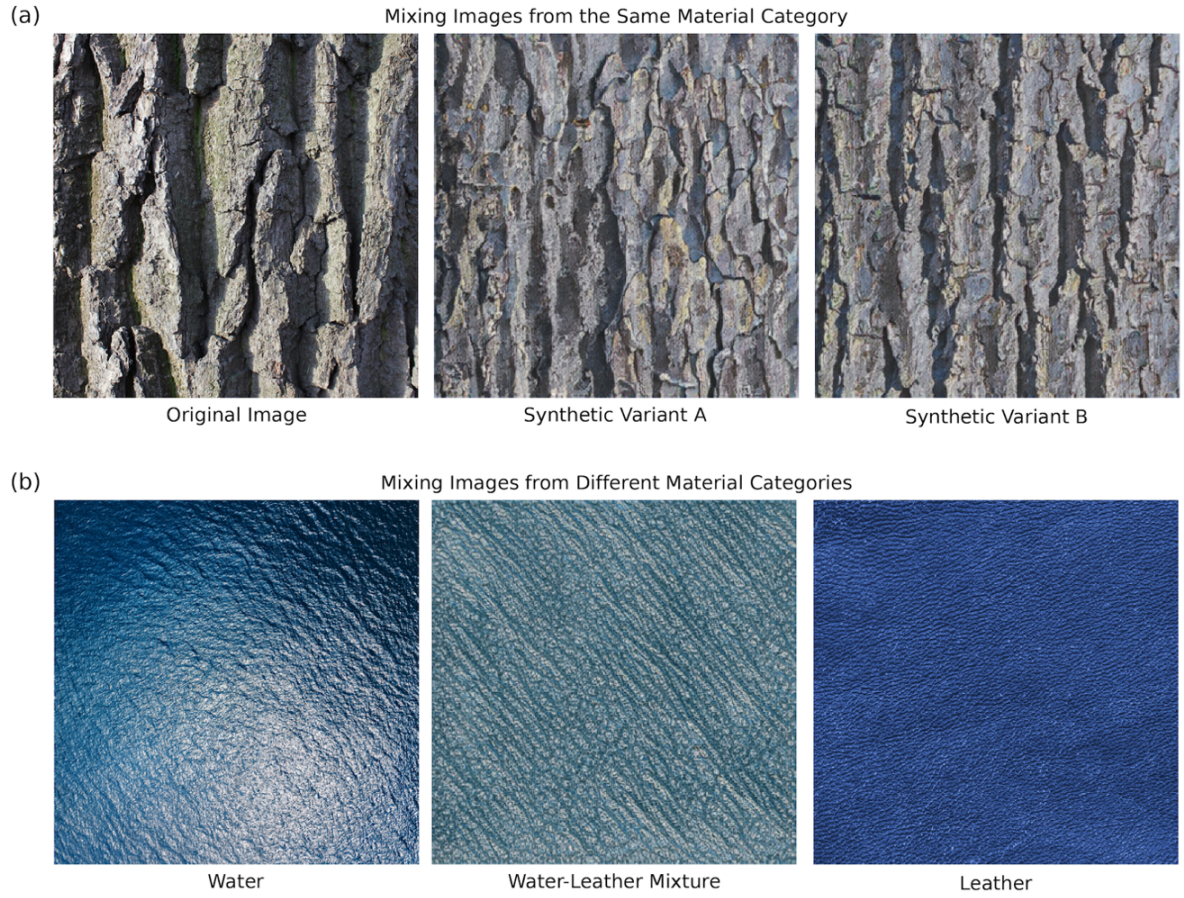

Figure 7. Material ambiguity might be created with the aid of deep neural networks. (a) An original image from our set can be modified to mimic the image statistics of other exemplars from the same material category. (b) Original images from different material categories (water and leather) can be used to create a synthetic image that inherits the statistics of both images. These techniques could be used to aid the discovery or invention of images that possess specific kinds of material ambiguity.

4.5 Conclusions

Our findings reveal a novel top-down effect in visual categorization. We find that certain images can appear to depict two or more completely different categories of material, depending on the apparent or implied viewing distance—a material-scale ambiguity. Future research focusing on automatically finding and synthesizing images with scale-dependent material ambiguity could inform theories of material perception, and more broadly, how image statistics relate to image interpretation.

CHAPTER 5: GENERAL DISCUSSION

In summary, Chapters 2 and 3 both explored the variability in human perception of gloss, but from different angles. While Chapter 2 focused on how discriminability of gloss changes with specular reflectance magnitude, Chapter 3 extended this by examining how external factors like lighting, shape and viewpoint influence the perception of gloss differences. Both chapters illustrated how gloss perception is influenced by the properties of the surface and the context in which it is viewed. The findings in Chapter 3, particularly the consistency in observer rankings and the success of HDR-VDP-3 in predicting these rankings, point the way toward developing a perceptual standard for gloss. This links to the practical aspects highlighted in Chapter 2, where the efficiency and applicability of methods like MLDS are discussed, indicating the potential for these methods in real-world settings like industrial manufacturing and quality control. Chapter 4 introduced a complementary perspective by focusing on material-scale ambiguity and the role of top-down effects, such as assumed viewing distance, in material categorization. This relates to the themes of the first two chapters by illustrating how perceptual judgments, whether they are about gloss or material category, are susceptible to a range of external influences, leading to varying interpretations of the same physical properties. Collectively, these chapters present a unified narrative: human material perception depends on both the physical properties of surfaces, and the diverse conditions under which they are observed.

5.1 Gloss and material perception

Material perception has traditionally been understood as a hierarchical process in which low-level retinal image measurements evolve into mid-level surface estimates, which then project into a high-dimensional feature space used for the recognition of material classes (Anderson, 2011; Fleming, 2017; Komatsu & Goda, 2018). According to this standard feedforward view, material properties or classes are represented as trajectories or regions within this feature space, facilitating the retrieval of semantic knowledge about surface qualities. As a prime example of this line of thinking in the case of surface gloss, it has been argued that local contrast measurements are assembled into summary statistics of reflectance that describe specular highlights, and these determine our overall perception of glossiness (Motoyoshi et al., 2007).

However, the research presented in this thesis suggests that this traditional view may be overly simplistic, and it is increasingly clear that mid- and high-level aspects of material perception are far more interconnected than previously thought. For example, the results of Chapter 2 (Cheeseman et al., 2021) demonstrated that there is no single JND for gloss, even when varying only one dimension (specular reflectance) in symmetric viewing conditions. Additionally, as documented in Chapter 3, a fixed difference in reflectance can be easily discriminable in some viewing conditions, yet almost completely indiscriminable in others (Cheeseman et al., 2024). These results agree with other recent work in our lab (Morimoto et al., 2023b, 2023a), showing that although judgments of gloss frequently deviate from ground truth values of surface reflectance, they are nevertheless highly consistent across observers.

In the realm of material recognition, Chapter 4 showed that identical images can be interpreted as completely different materials based on the assumed distance to a surface (Cheeseman et al., 2022), a previously undocumented phenomenon that aligns with other top-down effects in color and material perception (Alley et al., 2019; Gegenfurtner et al., 2015; Hansen et al., 2006). How could our brains possibly implement such a flexible representation of material properties? Recent research, reviewed by Schmid and Doerschner (2019), suggests that the neural processing of material properties is not confined to a single pathway (e.g., the ventral stream) but rather is coupled with the computations of other object and scene attributes, indicating a more distributed neural representation of materials than previously understood. In a similar vein, Schmid et al. (2023) demonstrate that the recognition of material categories can be directly inferred from specular image structure. Their findings reveal that these image features not only influence the perception of gloss, but also influence the broader categorization of materials, which challenges the notion of a purely feedforward neural processing mechanism in material perception.

Despite the complexity of this emerging view of material perception, for centuries visual artists have wielded the ‘mind’s eye’ to convincingly depict subtle details of surfaces to indicate their material composition (Di Cicco et al., 2019, 2019; van Zuijlen et al., 2020). This mastery of the essential visual cues for material depiction, acquired through careful observation and countless hours of practice, represents high-level knowledge about how material properties of surfaces correlate with proximal image features. This suggests that the visual perception of materials involves a spectrum of cognitive abilities and is not restricted to bottom-up processing of visual

information; rather, the specific image cues utilized can vary drastically based on the context, whether in recognizing known materials, assessing unfamiliar ones, or for artistic representation.

5.2 Signal and noise

Although there are well-founded reasons to doubt the feedforward view of material perception, this still leaves us with the problem of how to characterize the high-dimensional feature space that maps functional knowledge of material classes to estimates of their specific properties. The core of this problem is in identifying the invariant sources of information that our brains draw upon in each context. That is, the physical world—which includes the surfaces that we interact with and the nervous systems that somehow represent these surfaces—is highly variable. Even in highly controlled laboratory conditions, a given surface does not reflect photons in exactly the same way in every instance (Corrêa & Saldanha, 2016), and even if this were possible, a given neuron could not respond in exactly the same way to this input (Tinsley et al., 2016). This means that visual perception must discount internal and external sources of noise, while simultaneously tracking relevant signals that also change over time.

The research in Chapter 2 presented a psychophysical method of inferring how internal noise—due to random fluctuations in neuronal activity—could vary with the magnitude of surface reflectance. The logic of this inference proceeds as follows: (i) assume that internal noise is identical at different physical magnitudes of the stimulus, (ii) implement a quantitative model of perceived magnitude based on this assumption, and (iii) compare predictions of the model with a pattern of human judgments made in controlled laboratory conditions. If a model with this assumption can fit the behavioral data reasonably well, this suggests that internal noise in the brain obeys the same rule. Critically, we found that both near-threshold and suprathreshold judgments of apparent gloss were well predicted by the same model (based on Maximum Likelihood Difference Scaling; see Maloney & Knoblauch, 2020; Maloney & Yang, 2003), which assumes that internal noise is additive, normally distributed, and with fixed variance. If the model had failed to predict either near-threshold or suprathreshold judgments, the validity of these assumptions could be doubted, as others have done (Aguilar et al., 2017; Hillis & Brainard, 2007; Protonotarios et al., 2016; Shooner & Mullen, 2022); however, our results align with studies supporting the fixed noise assumption (Devinck & Knoblauch, 2012a; Kingdom, 2016).

The fundamental issue underlying these mixed results is that small differences in appearance do not necessarily sum to large differences in appearance. For example, in a recent study of achromatic luminance discrimination, it was found that large luminance differences are perceived as less than the sum of smaller ones, challenging traditional models of perceptual color space (Brainard, 2022; Bujack et al., 2022). This basic problem, however, is not new, and can be traced back to Fechner's reformulation of Weber's Law (Fechner, 1860/1966), which proposed that perceived stimulus intensity is proportional to the logarithm of physical stimulus magnitude. This logarithmic perceptual scale results from integrating JNDs—these are known today as discrimination scales, or alternatively, Fechner scales (Baird & Noma, 1978). However, without knowing how internal noise grows with stimulus magnitude, discrimination scales can only describe near-threshold sensitivity, not suprathreshold similarity (Kingdom & Prins, 2016). This confusion eventually contributed to theoretical debates about the quantitative and qualitative relationship between judgments of small and large stimulus magnitudes, initially led by Franz Brentano (1874/2015), then the Gestalt school (Koffka, 1922; Wertheimer, 1912/2012), and later by Stanley S. Stevens (1961, 1975) and others (e.g., Krantz, 1971; Luce & Edwards, 1958).

To measure perceived distances between stimuli, it is necessary to estimate these distances not along any arbitrary path in the representational space, but along the *shortest* (geodesic) path that covaries with suprathreshold similarity (Borg & Groenen, 2005, p. 362; Shepard, 1957). As these authors and others have noted, measurements of sensitivity and similarity will depend on the range of stimulus intensities chosen by the experimenter (Baird & Noma, 1978). For example, Shooner and Mullen (2022) demonstrated that correspondence between perceptual scaling (MLDS) and discrimination (2AFC) exists only for middle ranges of chromatic and achromatic contrast—where Weber's Law typically applies. At the extremes of stimulus intensity, however, MLDS had consistently higher estimates of internal noise, leading to underestimates of sensitivity. This discrepancy was attributed to the increased complexity of the MLDS task, which requires a comparison of stimulus *differences*, rather than simple discrimination. Task complexity can be conceptualized as any source of external, task-irrelevant variation whose effect on the perceptual representation matches that of internal noise, with more complex tasks leading to underestimates of sensitivity (Singh et al., 2022). The nature of task complexity also likely differs across physical continua and stimulus conditions; for example, departures from Weber's Law have been documented for the perception of pitch (Gulick, 1971) and temporal intervals (Michon, 1964).

Potential analytic solutions to this problem have been proposed. For example, the underlying scaling function for internal noise can be directly modeled without assumptions about its shape, whether additive or non-additive (Teti et al., 2022). Methods without assumptions about internal noise, such as non-metric multidimensional scaling (Noorlander & Koenderink, 1983) or ordinal embedding (Haghiri et al., 2020), could also be helpful. Recently, a novel mathematical framework has been introduced which accounts for both near-threshold sensitivity and suprathreshold similarity for gustatory, auditory, and visual continua (Zhou, 2023; Zhou et al., 2024). This analysis has suggested that suprathreshold scaling methods like MLDS essentially produce a discrimination scale, which would explain why the shape of the perceptual scale produced by MLDS—and predicted JNDs derived from the scale—remain stable under different model assumptions about internal noise (Maloney & Knoblauch, 2020; Maloney & Yang, 2003), yet seem to depend on task complexity and stimulus conditions. In other words, according to this unified framework, MLDS measures near-threshold sensitivity, not suprathreshold similarity. If true, the prediction of JNDs using suprathreshold scaling methods like MLDS does not provide conclusive evidence that internal noise grows *additively* with stimulus magnitude, as the same JNDs can be predicted with a model assuming *multiplicative* internal noise. This points to an important limitation of the work presented in Chapter 2, and implicates previous studies that have used suprathreshold scaling to draw inferences about internal noise (e.g., Kingdom, 2016).

Other potential solutions have emerged from applying deep learning methods to problems of artificial and biological vision. In this vein, recent work from our lab has shown that a high-dimensional feature space of material properties and classes is learnable from sufficiently large and diverse image sets (Fleming & Storrs, 2019; Prokott & Fleming, 2022; Prokott et al., 2021; Storrs et al., 2021; Tamura et al., 2022; van Assen et al., 2020). Data-driven deep learning methods have also recently enabled targeted parametric control of complex appearance attributes, including material properties of objects (Liao et al., 2023; Sharma, Jampani, et al., 2023; Sharma, Philip, et al., 2023). These methods, if combined, could be used to probe neural responses related to material property estimation and material recognition. For example, Vacher et al. (2020) used MLDS to constrain image perturbations generated by convolutional neural networks (CNNs) to the shortest (geodesic) path between two images, allowing for smooth interpolation between textures. They then applied this method to assess how neurons in different visual cortical areas respond to these image perturbations, finding that interpolating between a naturalistic texture and a spectrally-

matched Gaussian texture increases stimulus-related information linearly in macaque V4 neurons, but not in V1 neurons. Konkle and Alvarez (2022) have also demonstrated that CNNs trained to represent individual images can predict human ventral stream responses, and that these response patterns form interpretable classes of objects and scenes. Interestingly, the same models struggled to distinguish fine-grained differences in shape for surfaces with similar textures (perhaps due a bias toward texture-based representations; see Geirhos et al., 2018), and yet, this categorical organization of the model's latent space emerged regardless of the distribution of visual features in the model's training data.

While this result would suggest that a stable, coherent space of category representations, primarily organized by texture, also exists in the ventral stream of the human visual system, artificial neural networks and humans alike are vulnerable to *adversarial examples*—small, targeted changes to images that lead to drastically different category perceptions (Yuan et al., 2019). Notably, a recent study by Gaziv et al. (2023) using artificial neural networks has revealed 'wormholes' in perceptual space where adversarial image manipulations can succeed in biasing human category perception, even when the models that generate these image changes are robust to adversarial examples. As a general statement, artificial and biological visual systems are more vulnerable to mis-classification when sensory inputs are ambiguous—in other words, when a stimulus does not sufficiently constrain the brain's predictions, internal biases have greater influence on perception (Shepard, 1984).

The results of Chapter 4 illustrated examples of images that can provoke categorically different interpretations that depend on assumptions about viewing distance. Although we identified photographs of real scenes with this kind of ambiguity, such images could also be synthesized by artificial neural networks, and we are currently exploring how these tools can be used to better understand categorical representation in human brain networks. That is, research on the neural representation of categorical visual perception typically uses images that are relatively unambiguous in the sense that they are recognizable members of specific classes of objects. By using ambiguous images that define category boundaries, and mapping perception of these images to brain activity, we may be able to characterize which image features and neural populations lead to 'wormholes' in material perception, and categorical perception more generally.

5.3 Conclusions

This thesis set out to address a series of questions that bridge the gap between theoretical understanding and practical applications in gloss perception and material recognition:

1. How does gloss sensitivity vary across different magnitudes of specular reflectance, and what might this reveal about its underlying perceptual mechanisms?
2. Can observer rankings of perceived gloss under diverse conditions guide the development of a perceptual measurement standard for gloss appearance?
3. In what ways does apparent spatial scale influence material categorization, and what does this reveal about top-down effects in material recognition?

Regarding the first open question about visual sensitivity to specular reflectance, the research presented in Chapter 2 demonstrated that this sensitivity varies significantly with the magnitude of specular reflectance. This investigation, which utilized Maximum Likelihood Difference Scaling (MLDS) and the Method of Constant Stimuli, reveals that discriminability in gloss perception diminishes at high levels of specular reflectance. This suggests that gloss sensitivity depends on the magnitude of image changes due to varying reflectance values and is influenced by internal sensory noise. The consistency of sensory representations across suprathreshold and near-threshold intervals, as noted in the studies by Kingdom (2016) and Devinck and Knoblauch (2012), supports the effectiveness of MLDS in evaluating gloss sensitivity, especially when gloss is tightly correlated with image contrast.

Addressing the second question about developing a perceptual standard for gloss discrimination, the findings of Chapter 3 showed the potential of using image metrics, such as HDR-VDP-3, to predict gloss perception across changes in lighting, shape and viewpoint. The consistency in observer rankings, even in these diverse conditions, challenges prior assumptions about the inherent complexity of gloss perception and points to the practicality of simpler image contrast measurements in predicting perceived gloss differences. This consistency in observer judgments, which aligns with other research by Morimoto et al. (2023a, 2023b) and Schmid et al. (2023), could guide the development of a more comprehensive industrial standard for gloss measurement and communication.

In response to the third question about top-down effects in material recognition, Chapter 4 demonstrated how assumptions about viewing distance can drastically alter material categorization. This material-scale ambiguity, as shown in experiments measuring apparent distance, reveals a substantial and hitherto undocumented top-down effect in visual recognition, where context can radically change the interpretation of identical images. This aligns with Fleming's (2014) view that materials are represented by 'statistical appearance models' influencing diagnostic image cues, and supports the idea that material perception theories should account for scale-specificity.

Finally, concerning the challenges and opportunities presented by AI-generated imagery in understanding material categorization, as explored in ongoing research, these tools can potentially be used to target different assumptions about the causal origin of image features. By blending the statistics of different images or utilizing artistic input, as suggested by Gatys et al. (2016) and van Zuijlen et al. (2020), AI-generated imagery could become a powerful tool in probing the characteristics of images that lead to material ambiguities. This approach could significantly inform theories of material perception and the broader relationship between image statistics and interpretation.

REFERENCES

Adams, W. J., Elder, J. H., Graf, E. W., Leyland, J., Lugtigheid, A. J., & Muryy, A. (2016). The Southampton-York Natural Scenes (SYNS) dataset: Statistics of surface attitude. *Scientific Reports*, *6*, 35805. https://doi.org/10.1038/srep35805

Adams, W. J., Kucukoglu, G., Landy, M. S., & Mantiuk, R. K. (2018). Naturally glossy: Gloss perception, illumination statistics and tone mapping. *Journal of Vision*, *18*(9).

Aguilar, G., & Maertens, M. (2020). Toward reliable measurements of perceptual scales in multiple contexts. *Journal of Vision*, *20*(4), 19. https://doi.org/10.1167/jov.20.4.19

Aguilar, G., Wichmann, F. A., & Maertens, M. (2017). Comparing sensitivity estimates from MLDS and forced-choice methods in a slant-from-texture experiment. *Journal of Vision*, *17*(1), 37. https://doi.org/10.1167/17.1.37

Alley, L. M., Schmid, A. C., & Doerschner, K. (2019). Visual perception of surprising materials in dynamic scenes. *bioRxiv*, 744458. https://doi.org/10.1101/744458

Anderson, B. L. (2011). Visual perception of materials and surfaces. *Current Biology*, *21*(24), R978–R983. https://doi.org/10.1016/j.cub.2011.11.022

Anderson, B. L., & Kim, J. (2009). Image statistics do not explain the perception of gloss and lightness. *Journal of Vision*, *9*(11), 10. https://doi.org/10.1167/9.11.10

Appelle, S. (1972). Perception and discrimination as a function of stimulus orientation: The "oblique effect" in man and animals. *Psychological Bulletin*, *78*(4), 266–278. https://doi.org/10.1037/h0033117

Baird, J., & Noma, E. (1978). *Fundaments of Scaling and Psychophysics*. John Wiley & Sons.

Bell, S., Upchurch, P., Snavely, N., & Bala, K. (2015, June). Material recognition in the wild with the materials in context database. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Berzhanskaya, J., Swaminathan, G., Beck, J., & Mingolla, E. (2005). Remote effects of highlights on gloss perception. *Perception*, *34*(5), 565–575. https://doi.org/10.1068/p5401

Borg, I., & Groenen, P. J. F. (2005). *Modern multidimensional scaling: Theory and applications*. Springer-Verlag. https://doi.org/10.1007/0-387-28981-X

Bousseau, A., Chapoulie, E., Ramamoorthi, R., & Agrawala, M. (2011). Optimizing Environment Maps for Material Depiction. *Computer Graphics Forum*, *30*(4), 1171–1180. https://doi.org/10.1111/j.1467-8659.2011.01975.x

Boyadzhiev, I., Bala, K., Paris, S., & Adelson, E. (2015). Band-sifting decomposition for image-based material editing. *ACM Transactions on Graphics*, *34*(5), 163:1--163:16. https://doi.org/10.1145/2809796

Brachmann, A., & Redies, C. (2017a). Computational and Experimental Approaches to Visual Aesthetics. *Frontiers in Computational Neuroscience*, *11*. https://doi.org/10.3389/fncom.2017.00102

Brachmann, A., & Redies, C. (2017b). Defining self-similarity of images using features learned by convolutional neural networks. *Electronic Imaging*, *2017*(14), 188–194. https://doi.org/10.2352/ISSN.2470-1173.2017.14.HVEI-142

Brainard, D. H. (2022). Proximity matters. *Proceedings of the National Academy of Sciences*, *119*(27), e2206437119. https://doi.org/10.1073/pnas.2206437119

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brentano, F. (2015). *Psychology from an Empirical Standpoint*. Routledge.

Bujack, R., Teti, E., Miller, J., Caffrey, E., & Turton, T. L. (2022). The non-Riemannian nature of perceptual color space. *Proceedings of the National Academy of Sciences*, *119*(18), e2119753119. https://doi.org/10.1073/pnas.2119753119

Campbell, F. W., Nachmias, J., & Jukes, J. (1970). Spatial-frequency discrimination in human vision. *Journal of the Optical Society of America*, *60*(4), 555–559. https://doi.org/10.1364/JOSA.60.000555

Campbell, F. W., & Robson, J. G. (1968). Application of Fourier analysis to the visibility of gratings. *The Journal of Physiology*, *197*(3), 551–566. https://doi.org/10.1113/jphysiol.1968.sp008574

Chadwick, A. C., & Kentridge, R. W. (2015). The perception of gloss: A review. *Vision Research*, *109*, 221–235. https://doi.org/10.1016/j.visres.2014.10.026

Chapiro, A., Hanji, P., Ashraf, M., Asano, Y., & Mantiuk, R. (2024). Visible Difference Predictors: A Class of Perception-Based Metrics. *SID Symposium Digest of Technical Papers*. https://youtu.be/RNsMh-lhLcA

Charrier, C., Knoblauch, K., Maloney, L. T., Bovik, A. C., & Moorthy, A. K. (2012). Optimizing multiscale SSIM for compression via MLDS. *IEEE Transactions on Image Processing*, *21*(12), 4682–4694. https://doi.org/10.1109/TIP.2012.2210723

Cheeseman, J. R., Ferwerda, J. A., Maile, F. J., & Fleming, R. W. (2020). *Supplemental materials: Scaling and discriminability of perceived gloss*. https://doi.org/10.17605/OSF.IO/9H75A

Cheeseman, J. R., Ferwerda, J. A., Maile, F. J., & Fleming, R. W. (2021). Scaling and discriminability of perceived gloss. *Journal of the Optical Society of America A*, *38*(2), 203–210. https://doi.org/10.1364/JOSAA.409454

Cheeseman, J. R., Ferwerda, J. A., Morimoto, T., & Fleming, R. W. (2024). *Gloss discrimination: Towards an image-based perceptual model*.

Cheeseman, J. R., Fleming, R. W., & Schmidt, F. (2022). Scale ambiguities in material recognition. *iScience*, *25*(3), 103970. https://doi.org/10.1016/j.isci.2022.103970

Corrêa, R., & Saldanha, P. L. (2016). Photon reflection by a quantum mirror: A wave-function approach. *Phys. Rev. A*, *93*(2), 023803. https://doi.org/10.1103/PhysRevA.93.023803

Cutting, J. E., & Vishton, P. M. (1995). Chapter 3—Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In W. Epstein & S. B. T. Rogers (Eds.), *Handbook of Perception and Cognition* (pp. 69–117). Academic Press. https://doi.org/10.1016/B978-012240530-3/50005-5

Daly, S. J. (1992). Visible differences predictor: An algorithm for the assessment of image fidelity. *SPIE 1666, Human Vision, Visual Processing, and Digital Display III*, 1614–1666. https://doi.org/10.1117/12.135952

De Giuli, R. (2018). *ICN T1*. https://www.terracollage.com/icn-t1

Devinck, F., & Knoblauch, K. (2012a). A common signal detection model accounts for both perception and discrimination of the watercolor effect. *Journal of Vision*, *12*(3), 19. https://doi.org/10.1167/12.3.19

Devinck, F., & Knoblauch, K. (2012b). A common signal detection model accounts for both perception and discrimination of the watercolor effect. *Journal of Vision*, *12*(3), 19. https://doi.org/10.1167/12.3.19

Di Cicco, F., van Zuijlen, M. J. P., Wijntjes, M. W. A., & Pont, S. C. (2021). Soft like velvet and shiny like satin: Perceptual material signatures of fabrics depicted in 17th century paintings. *Journal of Vision*, *21*(5), 10. https://doi.org/10.1167/jov.21.5.10

Di Cicco, F., 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. https://doi.org/10.1167/19.3.7

Doerschner, K., Fleming, R. W., Yilmaz, O., Schrater, P. R., Hartung, B., & Kersten, D. (2011). Visual motion and the perception of surface material. *Current Biology*, *21*(23), 2010–2016. https://doi.org/10.1016/j.cub.2011.10.036

Doerschner, K., Maloney, L. T., & Boyaci, H. (2010). Perceived glossiness in high dynamic range scenes. *Journal of Vision*, *10*(9), 11. https://doi.org/10.1167/10.9.11

European coatings dossier on testing and measuring. (2019). In *European Coatings Journal*. Vincentz. http://european-coatings-promotions.com/downloads/ec-dossier-testing-measuring/ec_dossier_2019_testing_and_measuring

Fairchild, M. D. (2005). *Color appearance models*. John Wiley & Sons.

Faul, F. (2019). The influence of Fresnel effects on gloss perception. *Journal of Vision*, *19*(13), 1. https://doi.org/10.1167/19.13.1

Faul, F. (2021). Perceived roughness of glossy objects: The influence of Fresnel effects and correlated image statistics. *Journal of Vision*, *21*(8), 1. https://doi.org/10.1167/jov.21.8.1

Fechner, G. T. (1966). *Elements of Psychophysics Vol. I (Translated by Adler, H. E.)* (D. H. Howes & E. G. Boring, Eds.). Holt, Rinehart and Winston.

Ferwerda, J. A., & Padhye, S. A. (2021). Visual perception of surface properties through manipulation. *Color and Imaging Conference*, *2021*(29). https://doi.org/10.2352/issn.2169-2629.2021.29.66

Ferwerda, J. A., Pellacini, F., & Greenberg, D. P. (2001). Psychophysically based model of surface gloss perception. *Human Vision and Electronic Imaging VI*, *4299*, 4211–4299.

Filip, J., & Kolafová, M. (2019). Perceptual attributes analysis of real-world materials. *ACM Transactions on Applied Perception*, *16*(1). https://doi.org/10.1145/3301412

Fleming, R. W. (2014). Visual perception of materials and their properties. *Vision Research*, *94*, 62–75. https://doi.org/10.1016/j.visres.2013.11.004

Fleming, R. W. (2017). Material perception. *Annual Review of Vision Science*, *3*(1), 365–388. https://doi.org/10.1146/annurev-vision-102016-061429

Fleming, R. W., Dror, R. O., & Adelson, E. H. (2003). Real-world illumination and the perception of surface reflectance properties. *Journal of Vision*, *3*(5), 3. https://doi.org/10.1167/3.5.3

Fleming, R. W., & Storrs, K. R. (2019). Learning to see stuff. *Current Opinion in Behavioral Sciences*, *30*, 100–108. https://doi.org/10.1016/j.cobeha.2019.07.004

Fleming, R. W., Wiebel, C., & Gegenfurtner, K. (2013). Perceptual qualities and material classes. *Journal of Vision*, *13*(8), 9. https://doi.org/10.1167/13.8.9

Fores, A., Fairchild, M. D., & Tastl, I. (2014). Improving the perceptual uniformity of a gloss space. *Color and Imaging Conference*, 7–13.

Freeman, W. T. (1994). The generic viewpoint assumption in a framework for visual perception. *Nature*, *368*(6471), 542–545. https://doi.org/10.1038/368542a0

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gaziv, G., Lee, M. J., & DiCarlo, J. J. (2023). *Robustified ANNs Reveal Wormholes Between Human Category Percepts*. https://doi.org/10.48550/arXiv.2308.06887

Ged, G., Rabal-Almazor, A. M., Himbert, M. E., & Obein, G. (2020). Assessing gloss under diffuse and specular lighting. *Color Research & Application*. https://doi.org/10.1002/col.22510

Gegenfurtner, K. R., Bloj, M., & Toscani, M. (2015). The many colours of 'the dress.' *Current Biology*, *25*(13), R543–R544. https://doi.org/10.1016/j.cub.2015.04.043

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). *ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness*. https://doi.org/10.48550/arXiv.1811.12231

Geisler-Moroder, D., & Dür, A. (2010). A new Ward BRDF model with bounded albedo. *Computer Graphics Forum*, *29*(4), 1391–1398. https://doi.org/10.1111/j.1467-8659.2010.01735.x

Gilchrist, A., Kossyfidis, C., Agostini, T., Li, X., Bonato, F., Cataliotti, J., Spehar, B., Annan, V., & Economou, E. (1999). An anchoring theory of lightness perception. *Psychological Review*, *106*(4), 795–834. https://doi.org/10.1037/0033-295X.106.4.795

Gombrich, E. H. (1961). *Art and illusion: A study in the psychology of pictorial representation*. Princeton University Press.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). Wiley.

Greenberg, D. P., Torrance, K. E., Shirley, P., Arvo, J., Lafortune, E., Ferwerda, J. A., Walter, B., Trumbore, B., Pattanaik, S., & Foo, S. C. (1997). A framework for realistic image synthesis. *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, 477–494. https://doi.org/10.1145/258734.258914

Gulick, W. L. (1971). *Hearing: Physiology and psychophysics.* (pp. xi, 258). Oxford U. Press.

Hafri, A., Wadhwa, S., & Bonner, M. F. (2021). *"Honey, I shrunk the scene": Decoupling the impact of distance from content on memory for scene boundaries*. https://doi.org/10.31234/osf.io/hy3qs

Haghiri, S., Rubisch, P., Geirhos, R., Wichmann, F., & von Luxburg, U. (2019). *Comparison-based framework for psychophysics: Lab versus crowdsourcing*.

Haghiri, S., Wichmann, F. A., & von Luxburg, U. (2020). Estimation of perceptual scales using ordinal embedding. *Journal of Vision*, *20*(9), 14. https://doi.org/10.1167/jov.20.9.14

Hansen, T., Olkkonen, M., Walter, S., & Gegenfurtner, K. R. (2006). Memory modulates color appearance. *Nature Neuroscience*, *9*(11), 1367–1368. https://doi.org/10.1038/nn1794

Harrison, V. G. W. (1945). *Definition and measurement of gloss: A survey of the published literature*. W. Heffer & Sons Ltd.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition* (2nd ed.). Springer New York, NY. https://doi.org/10.1007/978-0-387-84858-7

Hillis, J. M., & Brainard, D. H. (2007). Distinct mechanisms mediate visual detection and identification. *Current Biology*, *17*(19), 1714–1719. https://doi.org/10.1016/j.cub.2007.09.012

Ho, Y. X., Landy, M. S., & Maloney, L. T. (2006). How direction of illumination affects visually perceived surface roughness. *Journal of Vision*, *6*(5), 8. https://doi.org/10.1167/6.5.8

Ho, Y. X., Landy, M. S., & Maloney, L. T. (2008). Conjoint measurement of gloss and surface texture. *Psychological Science*, *19*(2), 196–204. https://doi.org/10.1111/j.1467-9280.2008.02067.x

Ho, Y. X., Maloney, L. T., & Landy, M. S. (2007). The effect of viewpoint on perceived visual roughness. *Journal of Vision*, *7*(1), 1. https://doi.org/10.1167/7.1.1

Hubbard, T. L., Kall, D., & Baird, J. C. (1989). Imagery, memory, and size-distance invariance. *Memory & Cognition*, *17*(1), 87–94. https://doi.org/10.3758/BF03199560

Hunter, R. S. (1937). Methods of determining gloss. *NBS Research Paper RP*, *958*.

Hunter, R. S., & Harold, R. W. (1987). *The measurement of appearance*. John Wiley & Sons.

Jakob, W. (2010). *Mitsuba renderer*. https://www.mitsuba-renderer.org/

Jensen, H. W. (1996). Global illumination using photon maps. In X. Pueyo & P. Schröder (Eds.), *Proceedings of the eurographics workshop on Rendering techniques '96* (pp. 21–30). Springer Vienna. https://doi.org/10.1007/978-3-7091-7484-5_3

Kildau, J. (2016). *Perceptual dimensions of high gloss materials*. Justus-Liebig-Universität Gießen.

Kim, J., & Anderson, B. L. (2010). Image statistics and the perception of surface gloss and lightness. *Journal of Vision*, *10*(9), 3. https://doi.org/10.1167/10.9.3

Kingdom, F. A. A. (2016). Fixed versus variable internal noise in contrast transduction: The significance of Whittle's data. *Vision Research*, *128*, 1–5. https://doi.org/10.1016/j.visres.2016.09.004

Kingdom, F. A. A., & Prins, N. (2016). *Psychophysics: A practical introduction*. Academic Press.

Koffka, K. (1922). Perception: An introduction to the Gestalt-Theorie. *Psychological Bulletin*, *19*(10), 531–585. https://doi.org/10.1037/h0072422

Komatsu, H., & Goda, N. (2018). Neural mechanisms of material perception: Quest on Shitsukan. *Neuroscience*. https://doi.org/10.1016/j.neuroscience.2018.09.001

Konkle, T., & Alvarez, G. A. (2022). A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communications*, *13*(1), 491. https://doi.org/10.1038/s41467-022-28091-4

Konkle, T., & Oliva, A. (2011a). Canonical visual size for real-world objects. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 23–37. https://doi.org/10.1037/a0020413

Konkle, T., & Oliva, A. (2011b). Canonical visual size for real-world objects. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 23–37. https://doi.org/10.1037/a0020413

Konkle, T., & Oliva, A. (2012a). A familiar-size Stroop effect: Real-world size is an automatic property of object representation. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(3), 561–569. https://doi.org/10.1037/a0028294

Konkle, T., & Oliva, A. (2012b). A real-world size organization of object responses in occipitotemporal cortex. *Neuron*, *74*(6), 1114–1124. https://doi.org/10.1016/j.neuron.2012.04.036

Kontsevich, L. L., & Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, *39*(16), 2729–2737. https://doi.org/10.1016/S0042-6989(98)00285-5

Krantz, D. H. (1971). Integration of just-noticeable differences. *Journal of Mathematical Psychology*, *8*(4), 591–599. https://doi.org/10.1016/0022-2496(71)90008-3

Leloup, F. B., Pointer, M. R., Dutré, P., & Hanselaer, P. (2012). Overall gloss evaluation in the presence of multiple cues to surface glossiness. *Journal of the Optical Society of America A*, *29*(6), 1105–1114. https://doi.org/10.1364/JOSAA.29.001105

Liao, C., Sawayama, M., & Xiao, B. (2023). Unsupervised learning reveals interpretable latent representations for translucency perception. *PLOS Computational Biology*, *19*(2), e1010878. https://doi.org/10.1371/journal.pcbi.1010878

Lieberman, H. R., & Pentland, A. P. (1982). Microcomputer-based estimation of psychophysical thresholds: The Best PEST. *Behavior Research Methods & Instrumentation*, *14*(1), 21–25. https://doi.org/10.3758/BF03202110

Lin, H., Van Zuijlen, M., Wijntjes, M. W. A., Pont, S. C., & Bala, K. (2021). Insights from a large-scale database of material depictions in paintings. In A. Del Bimbo, R. Cucchiara, S. Sclaroff, G. M. Farinella, T. Mei, M. Bertini, H. J. Escalante, & R. Vezzani (Eds.), *International Conference on Pattern Recognition* (pp. 531–545). Springer International Publishing. https://doi.org/10.1007/978-3-030-68796-0_38

Linke, B., & Das, J. (2016). Aesthetics and gloss of ground surfaces: A review on measurement and generation. *Journal of Manufacturing Science and Engineering*, *138*(6), 64501–64505. https://doi.org/10.1115/1.4032587

Luce, R. D., & Edwards, W. (1958). The derivation of subjective scales from just noticeable differences. *Psychological Review*, *65*(4), 222–237. https://doi.org/10.1037/h0039821

MacAdam, D. L. (1942). Visual sensitivities to color differences in daylight. *Journal of the Optical Society of America*, *32*(5), 247–274. https://doi.org/10.1364/JOSA.32.000247

MacAdam, D. L. (1943). Specification of small chromaticity differences. *Journal of the Optical Society of America*, *33*(1), 18–26. https://doi.org/10.1364/JOSA.33.000018

Maloney, L. T., & Knoblauch, K. (2020). Measuring and modeling visual appearance. *Annual Review of Vision Science*. https://doi.org/10.1146/annurev-vision-030320-041152

Maloney, L. T., & Yang, J. N. (2003). Maximum likelihood difference scaling. *Journal of Vision*, *3*(8), 5. https://doi.org/10.1167/3.8.5

Mantiuk, R. K., Hammou, D., & Hanji, P. (2023). *HDR-VDP-3: A multi-metric for predicting image differences, quality and contrast distortions in high dynamic range and regular content*.

Mantiuk, R., Kim, K. J., Rempel, A. G., & Heidrich, W. (2011). HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM SIGGRAPH 2011 Papers*, 40:1--40:14. https://doi.org/10.1145/1964921.1964935

Marlow, P. J., & Anderson, B. L. (2013). Generative constraints on image cues for perceived gloss. *Journal of Vision*, *13*(14), 2. https://doi.org/10.1167/13.14.2

Marlow, P. J., & Anderson, B. L. (2024). Interactions Between 3D Surface Shape and Material Perception. *Annual Review of Vision Science*. https://doi.org/10.1146/annurev-vision-102122-094213

Marlow, P. J., Kim, J., & Anderson, B. L. (2012). The perception and misperception of specular surface reflectance. *Current Biology*, *22*(20), 1909–1913. https://doi.org/10.1016/j.cub.2012.08.009

Mazzarella, J., Cholewiak, S., Phillips, F., & Fleming, R. (2014). Limits on the estimation of shape from specular surfaces. *Journal of Vision*, *14*(10), 721–721. https://doi.org/10.1167/14.10.721

Michon, J. A. (1964). Temporal Structure of Letter Groups and Span of Perception. *Quarterly Journal of Experimental Psychology*, *16*(3), 232–240. https://doi.org/10.1080/17470216408416373

Morimoto, T., Akbarinia, A., Storrs, K., Cheeseman, J. R., Smithson, H. E., Gegenfurtner, K. R., & Fleming, R. W. (2023a). A large-scale measurement of human gloss judgments revealed highly consistent and systematic failures of gloss constancy. *Journal of Vision*, *23*(9), 4781–4781. https://doi.org/10.1167/jov.23.9.4781

Morimoto, T., Akbarinia, A., Storrs, K., Cheeseman, J. R., Smithson, H. E., Gegenfurtner, K. R., & Fleming, R. W. (2023b). Color and gloss constancy under diverse lighting environments. *Journal of Vision*, *23*(7), 8–8. https://doi.org/10.1167/jov.23.7.8

Motoyoshi, I., & Matoba, H. (2012). Variability in constancy of the perceived surface reflectance across different illumination statistics. *Vision Research*, *53*(1), 30–39. https://doi.org/10.1016/j.visres.2011.11.010

Motoyoshi, I., Nishida, S., Sharan, L., & Adelson, E. H. (2007). Image statistics and the perception of surface qualities. *Nature*, *447*, 206. https://doi.org/10.1038/nature05724

Mury, A. A., Pont, S. C., & Koenderink, J. J. (2009). Structure of light fields in natural scenes. *Appl. Opt.*, *48*(28), 5386–5395. https://doi.org/10.1364/AO.48.005386

Nachmias, J., & Kocher, E. C. (1970). Visual detection and discrimination of luminance increments. *Journal of the Optical Society of America*, *60*(3), 382–389. https://doi.org/10.1364/JOSA.60.000382

Ngan, A., Durand, F., & Matusik, W. (2005). Experimental Analysis of BRDF Models. *Proceedings of the Sixteenth Eurographics Conference on Rendering Techniques*, 117–126. https://doi.org/10.2312/EGWR/EGSR05/117-126

Nicodemus, F. E., Richmond, J. C., Hsia, J. J., Ginsberg, I. W., Limperis, T., Galloway, K. ~F., & Roitman, P. (1977). Geometrical considerations and nomenclature for reflectance. In *Final Report National Bureau of Standards, Washington, DC. Inst. For Basic Standards.*

Nishida, S., & Shinya, M. (1998). Use of image-based information in judgments of surface-reflectance properties. *Journal of the Optical Society of America A*, *15*(12), 2951–2965. https://doi.org/10.1364/JOSAA.15.002951

Noorlander, C., & Koenderink, J. J. (1983). Spatial and temporal discrimination ellipsoids in color space. *J. Opt. Soc. Am.*, *73*(11), 1533–1543. https://doi.org/10.1364/JOSA.73.001533

Norman, J. F., & Phillips, F. (2016). *Bell Peppers (v1.1) [3D Object Files]*. http://www.skidmore.edu/~flip

Norman, J. F., Phillips, F., Cheeseman, J. R., Thomason, K. E., Ronning, C., Behari, K., Kleinman, K., Calloway, A. B., & Lamirande, D. (2016). Perceiving object shape from specular highlight deformation, boundary contour deformation, and active haptic manipulation. *PLOS ONE*, *11*(2), 1–15. https://doi.org/10.1371/journal.pone.0149058

Norman, J. F., Todd, J. T., & Phillips, F. (2020). Effects of illumination on the categorization of shiny materials. *Journal of Vision*, *20*(2). https://doi.org/10.1167/jov.20.5.2

Obein, G., Knoblauch, K., & Viéot, F. (2004). Difference scaling of gloss: Nonlinearity, binocularity, and constancy. *Journal of Vision*, *4*(9), 4.

Olkkonen, M., & Brainard, D. H. (2011). Joint effects of illumination geometry and object shape in the perception of surface reflectance. *I-Perception*, *2*(9), 1014–1034. https://doi.org/10.1068/i0480

Palmer, S. E., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In J. Long & A. Baddeley (Eds.), *Attention and performance IX* (pp. 135–151). Lawrence Erlbaum Associates.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Peirce, J., & Macaskill, M. (2018). *Building experiments in PsychoPy*. SAGE Publications Ltd.

Pellacini, F., Ferwerda, J. A., & Greenberg, D. P. (2000). Toward a psychophysically-based light reflection model for image synthesis. *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, 55–64. https://doi.org/10.1145/344779.344812

Phillips, J. B., Ferwerda, J. A., & Luka, S. (2009). Effects of image dynamic range on apparent surface gloss. *Color and Imaging Conference*, 193–197.

Phillips, J. B., Ferwerda, J. A., & Nunziata, A. (2010). Gloss discrimination and eye movements. *Proc.SPIE*, *7527*. https://doi.org/10.1117/12.845399

Pokorny, J., & Smith, V. C. (1970). Wavelength Discrimination in the Presence of Added Chromatic Fields. *J. Opt. Soc. Am.*, *60*(4), 562–569. https://doi.org/10.1364/JOSA.60.000562

Pont, S. C., & Koenderink, J. J. (2002). Bidirectional reflectance distribution function of specular surfaces with hemispherical pits. *Journal of the Optical Society of America A*, *19*(12), 2456–2466. https://doi.org/10.1364/JOSAA.19.002456

Pont, S. C., & Koenderink, J. J. (2005). Bidirectional texture contrast function. *International Journal of Computer Vision*, *62*(1), 17–34. https://doi.org/10.1023/B:VISI.0000046587.42611.2c

Pont, S. C., & te Pas, S. F. (2006). Material—Illumination ambiguities and the perception of solid objects. *Perception*, *35*(10), 1331–1350. https://doi.org/10.1068/p5440

Prokott, E., & Fleming, R. W. (2022). Identifying specular highlights: Insights from deep learning. *Journal of Vision*, *22*(7), 6–6. https://doi.org/10.1167/jov.22.7.6

Prokott, K. E. (2016). *Perception of high gloss materials*. Justus-Liebig-Universität Gießen.

Prokott, K. E., Tamura, H., & Fleming, R. W. (2021). Gloss perception: Searching for a deep neural network that behaves like humans. *Journal of Vision*, *21*(12), 14–14. https://doi.org/10.1167/jov.21.12.14

Protonotarios, E. D., Johnston, A., & Griffin, L. D. (2016). Difference magnitude is not measured by discrimination steps for order of point patterns. *Journal of Vision*, *16*(9), 2. https://doi.org/10.1167/16.9.2

Ramamoorthi, R., & Hanrahan, P. (2001). An efficient representation for irradiance environment maps. *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, 497–500. https://doi.org/10.1145/383259.383317

Ramanarayanan, G., Ferwerda, J., Walter, B., & Bala, K. (2007). Visual equivalence: Towards a new standard for image fidelity. *ACM Transactions on Graphics (TOG)*, *26*(3), 11. https://doi.org/10.1145/1276377.1276472

Raymond, B., Guennebaud, G., & Barla, P. (2016). Multi-scale rendering of scratched materials using a structured SV-BRDF model. *ACM Transactions on Graphics*, *35*(4). https://doi.org/10.1145/2897824.2925945

Reinhard, E., Stark, M., Shirley, P., & Ferwerda, J. (2002). Photographic tone reproduction for digital images. *ACM Trans. Graph.*, *21*(3), 267–276. https://doi.org/10.1145/566654.566575

Risser, E. (2020). Optimal textures: Fast and robust texture synthesis and style transfer through optimal transport. *arXiv*.

Sawayama, M., Nishida, S., & Shinya, M. (2017). Human perception of subresolution fineness of dense textures based on image intensity statistics. *Journal of Vision*, *17*(4), 8. https://doi.org/10.1167/17.4.8

Schmid, A. C., Barla, P., & Doerschner, K. (2023). Material category of visual objects computed from specular image structure. *Nature Human Behaviour*, *7*(7), 1152–1169. https://doi.org/10.1038/s41562-023-01601-0

Schmid, A. C., & Doerschner, K. (2019). Representing stuff in the human brain. *Current Opinion in Behavioral Sciences*, *30*, 178–185. https://doi.org/0.1016/j.cobeha.2019.10.007

Schütt, H. H., Harmeling, S., Macke, J. H., & Wichmann, F. A. (2016). Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data. *Vision Research*, *122*, 105–123. https://doi.org/10.1016/j.visres.2016.02.002

Sedgwick, H. A. (2005). Visual space perception. In E. B. Goldstein, G. Humphreys, M. Shiffrar, & W. Yost (Eds.), *Blackwell handbook of sensation and perception* (pp. 138–139). Blackwell Publishing.

Sharan, L., Liu, C., Rosenholtz, R., & Adelson, E. H. (2013). Recognizing materials using perceptually inspired features. *International Journal of Computer Vision*, *103*(3), 348–371. https://doi.org/10.1007/s11263-013-0609-0

Sharan, L., Rosenholtz, R., & Adelson, E. H. (2014). Accuracy and speed of material categorization in real-world images. *Journal of Vision*, *14*(9), 12.

Sharma, P., Jampani, V., Li, Y., Jia, X., Lagun, D., Durand, F., Freeman, W. T., & Matthews, M. (2023). *Alchemist: Parametric Control of Material Properties with Diffusion Models*. https://arxiv.org/abs/2312.02970

Sharma, P., Philip, J., Gharbi, M., Freeman, W. T., Durand, F., & Deschaintre, V. (2023). *Materialistic: Selecting Similar Materials in Images*.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*(4), 325–345. https://doi.org/10.1007/BF02288967

Shepard, R. N. (1984). Ecological constraints on internal representation: Resonant kinematics of perceiving, imagining, thinking, and dreaming. *Psychological Review*, *91*(4), 417–447. https://doi.org/10.1037/0033-295X.91.4.417

Shiwen, L., Morimoto, T., Harris, J. M., & Smithson, H. E. (2023). Task-dependent extraction of information from videos of iridescent and glossy samples. *J. Opt. Soc. Am. A*, *40*(3), A160–A168. https://doi.org/10.1364/JOSAA.479795

Shooner, C., & Mullen, K. T. (2022). Linking perceived to physical contrast: Comparing results from discrimination and difference-scaling experiments. *Journal of Vision*, *22*(1), 13. https://doi.org/10.1167/jov.22.1.13

Singh, V., Burge, J., & Brainard, D. H. (2022). Equivalent noise characterization of human lightness constancy. *Journal of Vision*, *22*(5), 2. https://doi.org/10.1167/jov.22.5.2

Smith, T., & Guild, J. (1931). The C.I.E. colorimetric standards and their use. In *Transactions of the Optical Society*. https://doi.org/10.1088/1475-4878/33/3/301

Stevens, S. S. (1961). To honor Fechner and repeal his law. *Science*, *133*(3446), 80–86.

Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural and social prospects*. John Wiley & Sons, Inc.

Storrs, K. R., Anderson, B. L., & Fleming, R. W. (2021). Unsupervised learning predicts human perception and misperception of gloss. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-021-01097-6

Tamura, H., Prokott, K. E., & Fleming, R. W. (2022). Distinguishing mirror from glass: A “big data” approach to material perception. *Journal of Vision*, *22*(4), 4–4. https://doi.org/10.1167/jov.22.4.4

te Pas, S. F., & Pont, S. C. (2005). A comparison of material and illumination discrimination performance for real rough, real smooth and computer generated smooth spheres. *Proceedings of the 2Nd Symposium on Applied Perception in Graphics and Visualization*, 75–81. https://doi.org/10.1145/1080402.1080415

Teti, E. S., Turton, T. L., Miller, J. M., & Bujack, R. (2022). Maximum likelihood estimation of difference scaling functions for suprathreshold judgments. *Journal of Vision*, *22*(10), 9–9. https://doi.org/10.1167/jov.22.10.9

Tiedemann, H. (2018). *The influence of shape complexity on gloss constancy*. Christian-Albrechts-Universität Kiel.

Tinsley, J. N., Molodtsov, M. I., Prevedel, R., Wartmann, D., Espigulé-Pons, J., Lauwers, M., & Vaziri, A. (2016). Direct detection of a single photon by humans. *Nature Communications*, *7*, 12172. https://doi.org/10.1038/ncomms12172

Toscani, M., Guarnera, D., Guarnera, G. C., Hardeberg, J. Y., & Gegenfurtner, K. R. (2020). Three perceptual dimensions for specular and diffuse reflection. *ACM Transactions on Applied Perception*, *1*(1), 27. https://doi.org/10.1145/3380741

Vacher, J., Davila, A., Kohn, A., & Coen-Cagli, R. (2020). Texture Interpolation for Probing Visual Perception. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 22146–22157). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/fba9d88164f3e2d9109ee770223212a0-Paper.pdf

van Assen, J. J. R., Nishida, S., & Fleming, R. W. (2020). Visual perception of liquids: Insights from deep neural networks. *PLOS Computational Biology*, *16*(8), e1008018. https://doi.org/10.1371/journal.pcbi.1008018

van Zuijlen, M. J. P., Pont, S. C., & Wijntjes, M. W. A. (2020). Painterly depiction of material properties. *Journal of Vision*, *20*(7), 7. https://doi.org/10.1167/jov.20.7.7

Vangorp, P., Laurijssen, J., & Dutré, P. (2007). The influence of shape on the perception of material reflectance. *ACM Transactions on Graphics (TOG)*, *26*(3). https://doi.org/10.1145/1276377.1276473

Vishwanath, D. (2010). Visual information in surface and depth perception: Reconciling pictures and reality. In L. Albertazzi, G. van Tonder, & D. Vishwanath (Eds.), *Perception beyond inference: The information content of visual processes* (pp. 201–240). MIT Press. https://doi.org/10.7551/mitpress/8594.003.0012

Vladusich, T. (2013). A unified account of gloss and lightness perception in terms of gamut relativity. *Journal of the Optical Society of America A*, *30*(8), 1568–1579. https://doi.org/10.1364/JOSAA.30.001568

Vu, C. T., Phan, T. D., & Chandler, D. M. (2012). \bf S_3: A Spectral and Spatial Measure of Local Perceived Sharpness in Natural Images. *IEEE Transactions on Image Processing*, *21*(3), 934–945. https://doi.org/10.1109/TIP.2011.2169974

Watson, A. B. (2017). QUEST+: A general multidimensional Bayesian adaptive psychometric method. *Journal of Vision*, *17*(3), 10. https://doi.org/10.1167/17.3.10

Wendt, G., & Faul, F. (2017). Increasing the complexity of the illumination may reduce gloss constancy. *I-Perception*, *8*(6), 2041669517740369. https://doi.org/10.1177/2041669517740369

Wendt, G., & Faul, F. (2018). Can color and motion information be used to disentangle the influence of multiple light sources on gloss perception? *I-Perception*, *9*(5), 2041669518803964. https://doi.org/10.1177/2041669518803964

Wendt, G., Faul, F., Ekroll, V., & Mausfeld, R. (2010). Disparity, motion, and color information improve gloss constancy performance. *Journal of Vision*, *10*(9), 7. https://doi.org/10.1167/10.9.7

Wertheimer, M. (2012). *On perceived motion and figural organization* (L. Spillmann, Ed.). MIT Press.

Whittle, P. (1986). Increments and decrements: Luminance discrimination. *Vision Research*, *26*(10), 1677–1691. https://doi.org/10.1016/0042-6989(86)90055-6

Whittle, P. (1992). Brightness, discriminability and the "Crispening Effect." *Vision Research*, *32*(8), 1493–1507. https://doi.org/10.1016/0042-6989(92)90205-W

Wiebel, C. B., Valsecchi, M., & Gegenfurtner, K. R. (2013). The speed and accuracy of material recognition in natural images. *Attention, Perception, & Psychophysics*, *75*(5), 954–966. https://doi.org/10.3758/s13414-013-0436-y

Wu, M., Xu, H., Wang, Z., & Li, H. (2016). Towards a practical metric of surface gloss for metallic coatings from automotive industry. *Journal of Coatings Technology and Research*, *13*(3), 469–477. https://doi.org/10.1007/s11998-015-9771-3

X. Yuan, P. He, Q. Zhu, & X. Li. (2019). Adversarial Examples: Attacks and Defenses for Deep Learning. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(9), 2805–2824. https://doi.org/10.1109/TNNLS.2018.2886017

Xia, L., Pont, S. C., & Heynderickx, I. (2014). The visual light field in real scenes. *I-Perception*, *5*(7), 613–629. https://doi.org/10.1068/i0654

Zhang, F., de Ridder, H., Barla, P., & Pont, S. (2020). Effects of light map orientation and shape on the visual perception of canonical materials. *Journal of Vision*, *20*(4), 13. https://doi.org/10.1167/jov.20.4.13

Zhou, J. (2023). Quantifying and predicting chromatic thresholds. *bioRxiv*, 2023.06.06.543898. https://doi.org/10.1101/2023.06.06.543898

Zhou, J., Duong, L. R., & Simoncelli, E. P. (2024). A unified framework for perceived magnitude and discriminability of sensory stimuli. *Proceedings of the National Academy of Sciences*, *121*(25), e2312293121. https://doi.org/10.1073/pnas.2312293121

APPENDIX

Supplementary material for Chapter 4

A similar version of this appendix has been published as supplementary material for:
Cheeseman, J. R., Fleming, R. W., & Schmidt, F. (2022). Scale ambiguities in material recognition. iScience, 25(3), 103970. https://doi.org/10.1016/j.isci.2022.103970

Table S1. Sources, copyright information and ground truth labels.

No.	Copyright Information	Label
1	by unsplash.com/Lysander Yuen [reprinted under Unsplash License]	Bark
2	Print image substituted with synthesized texture [https://github.com/JCBrouwer/OptimalTextures]	Marble
3	by 123rf.com/Arina Zaiachin [reprinted with permission]	Sand
4	by 123rf.com/Juergen Schonnop [reprinted with permission]	Sand
5	by 123rf.com/Mikhail Kokhanchikov [reprinted with permission]	Sand
6	Print image substituted with synthesized texture [https://github.com/JCBrouwer/OptimalTextures]	Concrete
7	by freepick.com/LuqueStock [reprinted under Freepik-Lizenz]	Metal
8	by freepik.com [reprinted under Freepik-Lizenz]	Metal
9	by unsplash.com/Aidan Brown [reprinted under Unsplash License]	Water
10	by unsplash.com/Anchor Lee [reprinted under Unsplash License]	Sand
11	by freeImages.com/Brandon Blinkenberg [reprinted under freeimages content license]	Stone
12	Print image substituted with synthesized texture [https://github.com/JCBrouwer/OptimalTextures]	Stone
13	by pixabay.com/AnnaAr [reprinted under CC0 1.0]	Bark
14	Print image substituted with synthesized texture [https://github.com/JCBrouwer/OptimalTextures]	Concrete
15	by Wagner Treppenbau, wagner-treppenbau.de [reprinted with permission]	Concrete
16	by Wagner Treppenbau, wagner-treppenbau.de [reprinted with permission]	Concrete
17	by slon.pics [reprinted under slon.pics free license]	Water
18	by unsplash.com/Carlo Verso [reprinted under Unsplash License]	Plant
19	by unsplash.com/Carlo Verso [reprinted under Unsplash License]	Plant
20	by unsplash.com/Carlo Verso [reprinted under Unsplash License]	Plant

21	by freeImages.com/dlritter [reprinted under freeimages content license]	Concrete
22	by unsplash.com/Das Sasha [reprinted under Unsplash License]	Water
23	by depositphotos.com/Petkov [reprinted with permission]	Paint
24	by depositphotos.com/studioDG [reprinted with permission]	Leather
25	by depositphotos.com/Lunamarina [reprinted with permission]	Paint
26	by depositphotos.com/Sergieiev [reprinted with permission]	Soil
27	by depositphotos.com/mario7 [reprinted with permission]	Leather
28	by depositphotos.com/Homydesign [reprinted with permission]	Leather
29	by depositphotos.com/Mankukuku [reprinted with permission]	Paint
30	by depositphotos.com/Alexis84 [reprinted with permission]	Plant
31	by depositphotos.com/tuja66 [reprinted with permission]	Metal
32	by depositphotos.com/kues [reprinted with permission]	Paint
33	by depositphotos.com/Watman [reprinted with permission]	Plant
34	by depositphotos.com/alxbaev@gmail.com [reprinted with permission]	Soil
35	by depositphotos.com/wayne0216 [reprinted with permission]	Soil
36	by depositphotos.com/Watman [reprinted with permission]	Soil
37	by depositphotos.com/cristi180884 [reprinted with permission]	Snow
38	by depositphotos.com/Natalt [reprinted with permission]	Leather
39	by depositphotos.com/ekina1 [reprinted with permission]	Soil
40	by depositphotos.com/VitaliyPliushc [reprinted with permission]	Metal
41	by depositphotos.com/VitaliyPliushc [reprinted with permission]	Metal
42	by depositphotos.com/ViktoriaSapata [reprinted with permission]	Soil
43	by unsplash.com/Evelyn Fjord [reprinted under Unsplash License]	Water
44	by freeImages.com/Krzysztof Isbrandt [reprinted under freeimages content license]	Water
45	by freeImages.com/Jose Mora [reprinted under freeimages content license]	Soil
46	by pixabay.com/TeroVesalainen [reprinted under CC0 1.0]	Ice
47	by unsplash.com/Ivan Bandura [reprinted under Unsplash License]	Plant
48	by unsplash.com/Jason Leem [reprinted under Unsplash License]	Water
49	by unsplash.com/Plenio [reprinted under Unsplash License]	Stone
50	by unsplash.com/Jude Infantini [reprinted under Unsplash License]	Bark

51	by bgfons.com [reprinted under CC BY-NC 4.0]	Marble
52	by 123freevectors.com[reprinted under 123freevectors License]	Marble
53	by unsplash.com/Matt Seymour [reprinted under Unsplash License]	Plant
54	by freeImages.com/Bjarne Henning Kvaale [reprinted under freeimages content license]	Metal
55	by photos-public-domain.com [reprinted under CC0 1.0]	Glass
56	by unsplash.com/Nate Bell [reprinted under Unsplash License]	Bark
57	by freepick.com/bedneyimages [reprinted under Freepik-Lizenz]	Glass
58	by unsplash.com/Rainer Basten [reprinted under Unsplash License]	Stone
59	“Veld 2” 2016 by David Burdeny, courtesy of Kostuik Gallery [reprinted with permission]	Plant
60	“Veld 6” 2016 by David Burdeny, courtesy of Kostuik Gallery [reprinted with permission]	Plant
61	“Veld 1” 2016 by David Burdeny, courtesy of Kostuik Gallery [reprinted with permission]	Plant
62	by jooinn.com [reprinted under CC0 1.0]	Water
63	Print image substituted with synthesized texture [https://github.com/JCBrouwer/OptimalTextures]	Stone
64	by shutterstock.com/rattiya lamrod [reprinted with permission]	Snow
65	by shutterstock.com/SAHACHATZ [reprinted with permission]	Sand
66	by shutterstock.com/Fred Mastison [reprinted with permission]	Glass
67	by shutterstock.com /Quality Stock Arts [reprinted with permission]	Plastic
68	by shutterstock.com/Dudarev Mikhail [reprinted with permission]	Water
69	by shutterstock.com/Shulevskyy Volodymyr [reprinted with permission]	Fabric
70	by shutterstock.com/Dudarev Mikhail [reprinted with permission]	Water
71	by shutterstock.com/SoulQuess [reprinted with permission]	Glass
72	by freeImages.com/meral akbulut [reprinted under freeimages content license]	Snow
73	by freeImages.com/Michal Zacharzeweski [reprinted under freeimages content license]	Snow
74	by unsplash.com/Steinar Engeland [reprinted under Unsplash License]	Bark
75	by pixabay.com/PellissierJP [reprinted under CC0 1.0]	Paint
76	by freeImages.com/Leo Celso [reprinted under freeimages content license]	Paint
77	by freeImages.com/Florian Florea [reprinted under freeimages content license]	Paint
78	by unsplash.com/Tim Johnson [reprinted under Unsplash License]	Water
79	by tonytextures.de [reprinted under tonytextures license]	Bark
80	by tonytextures.de [reprinted under tonytextures license]	Bark

81	by tonytextures.de [reprinted under tonytextures license]	Soil
82	by tonytextures.de [reprinted under tonytextures license]	Metal
83	by tonytextures.de [reprinted under tonytextures license]	Stone
84	by tonytextures.de [reprinted under tonytextures license]	Stone
85	by freeImages.com/Terry V. Haslett [reprinted under freeimages content license]	Bark
86	by freeImages.com/Rene Cerney [reprinted under freeimages content license]	Bark
87	by unsplash.com/Vanda Teixeira [reprinted under Unsplash License]	Sand

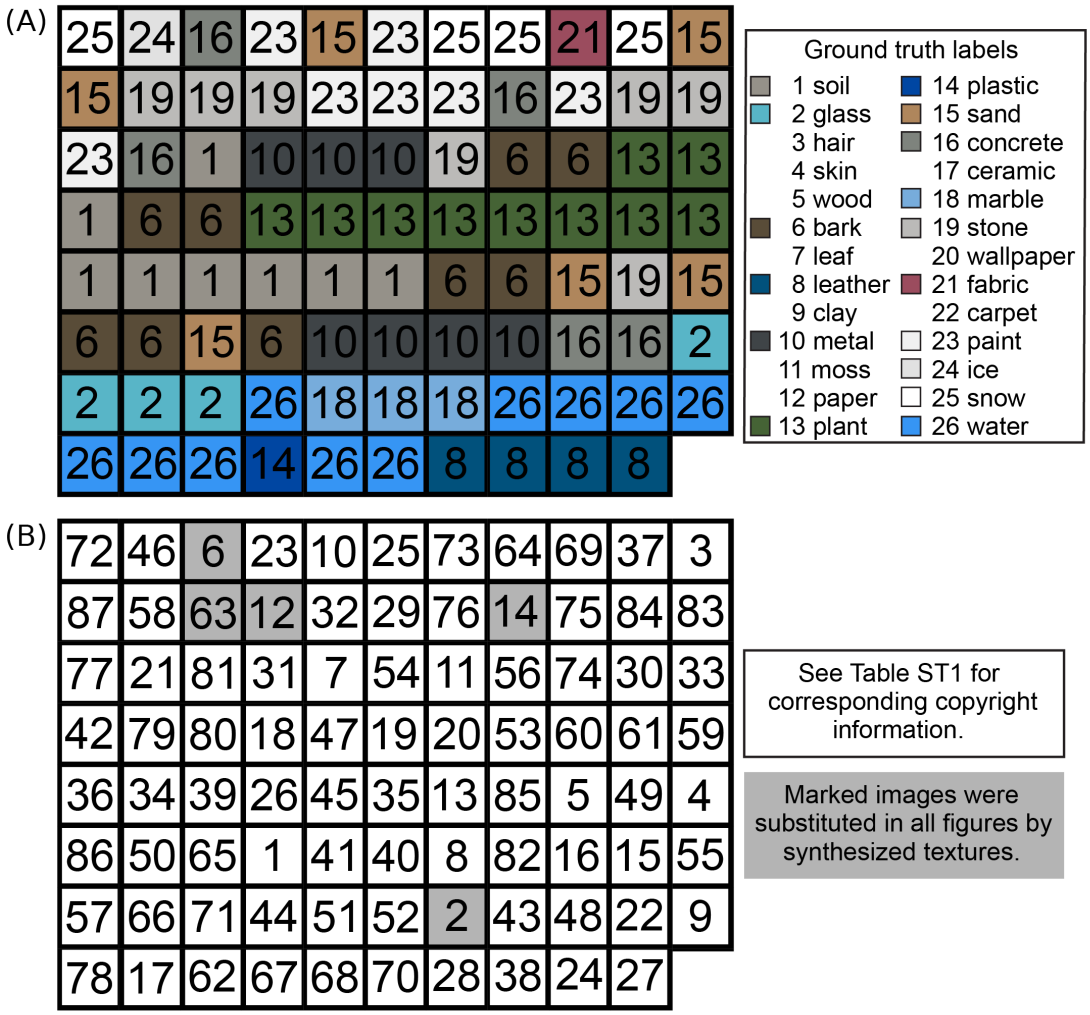

Figure S1. Ground truth material category labels (A) and image file numbers (B).

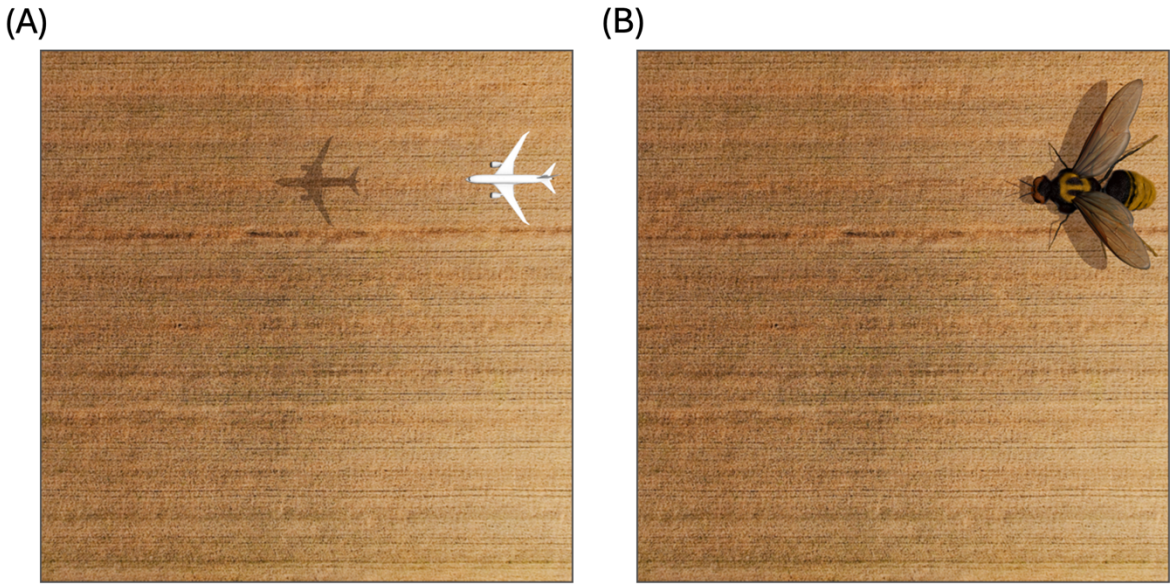

Figure S2. Examples of Experiment 2 stimuli for far (A) and near (B) conditions.

Distance estimate analysis

Distance estimates involved selecting a unit of measurement (micrometers, millimeters, centimeters, meters, or kilometers) before assigning a metric value (e.g., 8 cm, or 1 km). Although this task was designed to allow participants to quickly estimate very small or very large distances, the variance associated with each unit of measurement is unequal (e.g., distance estimates in kilometers will have a much larger variance than estimates in micrometers). When converted to a common unit (centimeters), therefore, the distribution of these values is extremely skewed (see Figure S3) and spans several orders of magnitude (SD = 106 × 106 cm). For this reason, analyses of distance estimates focused on the selected units of measurement, which represent judgements of viewing distance on an ordinal scale.

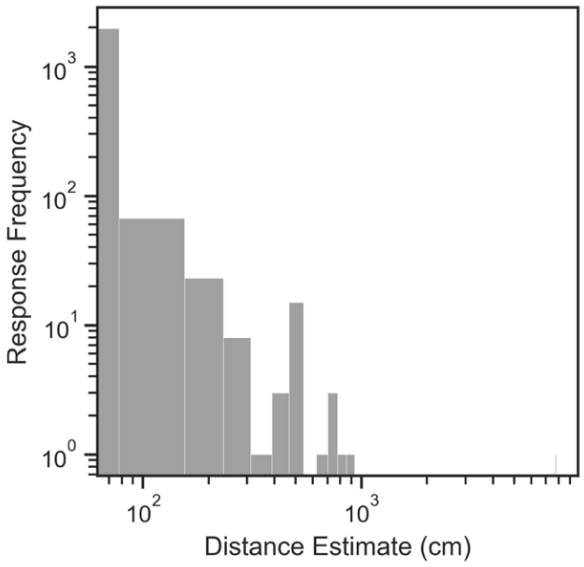

Figure S3. Log-scaled distribution of distance estimates (converted to centimeters).

Relative frequencies

Multiple-choice responses for each image were first divided into two groups based on the distance unit selected with the material: responses selected with micrometer, millimeter, or centimeter defined the near group, while those selected with meter or kilometer defined the far group. Responses of the near group were then paired with responses of the far group, such that all pairwise permutations in both directions (near → far vs. far → near) were represented. This was done separately for each image (i.e., responses associated with different images were never paired). The frequency of each unique pair of responses represents the relative frequency of directional category confusions. The relative frequencies shown in Figure 3 are calculated from response pairs for all images. Material-Scale Ambiguity (MSA) is calculated from response pairs for individual images.

Exploratory analyses

In a separate pilot experiment, naive participants (4 women, 1 man; M_{age} = 24.4 years, SD_{age} = 4.1 years) rated five appearance attributes (atmospheric blue tint, blurriness, direct lighting, surface gloss, and surface slant) on a continuous scale for the original set of 87 images. A self-similarity (i.e., scale-invariance) metric was also calculated for each image (Brachmann & Redies, 2017b). In order to assess whether this data might predict Material-Scale Ambiguity (MSA), a Support

Vector Machine (SVM) classifier (Pedregosa et al., 2011) was trained on 1000 random splits of the data into training and test sets. The target variable for classification was defined as a median split of MSA calculated from data in Experiment 1. A separate SVM classifier (using a Radial Basis Function kernel) was trained for each set of training and test data. Optimal parameters (C and Gamma) for this kernel function were obtained separately for each model using a grid search method. The mean cross-validated classification accuracy across these 1000 models is 56%, which is only marginally better than the accuracy that would be expected from a model that only learns the distribution of the target variable. That is, if exactly half of the images produce MSA that is greater than the median MSA produced by the whole set of images, a classifier that guesses randomly would be expected to have an accuracy of 50%. In order to understand this failure, the mean rating of each appearance attribute was calculated separately for each image. The most frequent distance unit selected for each image was then identified and grouped according to the near and far distance groups defined in Experiment 1. Finally, the mean rating of each appearance attribute was calculated separately for each distance group. Figure S4 plots these distance-dependent attribute ratings, which reveals that four of the five attributes were not significantly different between distance groups. The one exception is surface slant, which was much more highly rated for images that depicted relatively large distances (Mann–Whitney $U = 297$, $p < 0.001$, $d = 1.16$).

Additionally, these participants rated the 26 material categories used in Experiment 1 along a continuous scale from “inorganic” (man-made) to “organic” (natural). We calculated the mean rating for each material category and created two groups from a median split of these values; “organic” categories had a mean rating above the median and “inorganic” categories had a mean rating below the median. We then calculated the frequency of these two category groups within the near and far distance groups defined in Experiment 1. Notably, “organic” material categories made up 76% of responses when participants estimated surfaces to be far from the camera, but only 45% when surfaces were estimated to be near to the camera (see Figure S5).

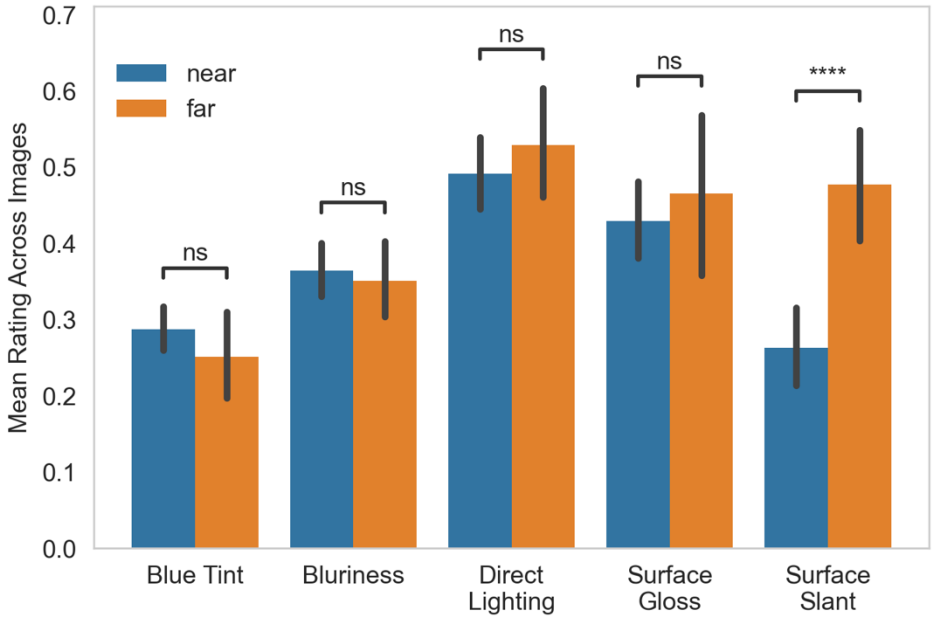

Figure S4. Mean ratings (across images) for each appearance attribute.

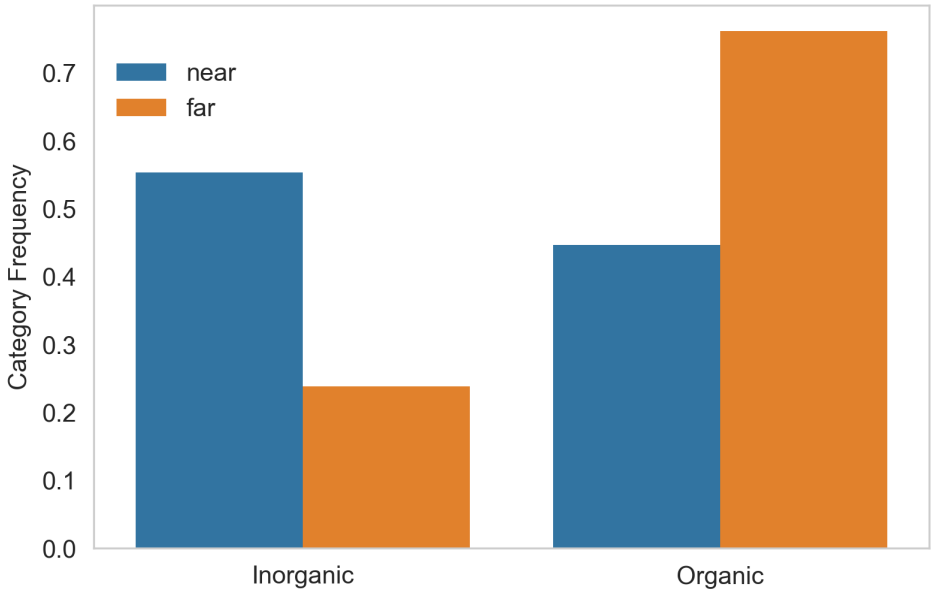

Figure S5. Distance-dependent frequency of categories rated as organic vs. inorganic.

Free-response transformation

Participants initially judged each image by providing a written description of the material, after which they selected a single material category from a list of options. To determine the extent of agreement between judgements of material in multiple-choice and free-response trials, the free-response data was manually reduced to single-word descriptions, which included all but one of the material categories available during the multiple-choice trials ("hair"), as well as the following 101 additional descriptors: aluminum, asphalt, basalt, bast, bone, brick, bronze, cake, canvas, cardboard, cellulose, cement, chalk, chalk stone, chocolate, cloud, coal, copper, coral, cord, cork, cotton, cotton candy, crystal, denim, detergent, dirt, dough, dust, screed, felt, fleece, flour, foam, foil, fungus, gel, gem, glue, gold, granite, graphite, grass, hay, honeycomb, iron, jade, jelly, jute, laminate, lead, lime, limestone, linen, meat, mirror, moss, mold, mud, none, nylon, oil, pearl, peat, photopaper, plaster, plexiglass, powder, pumice, putty, PVC, quartz, resin, rock, root, rubber, rust, salt, sandpaper, silicone, silk, silver, slate, slime, slom, smoke, soap, soil, sponge, steel, straw, styrofoam, sugar, sulfur, talcum, tar, velvet, vinyl, wheat, wool, hair

The free-response data was manually transformed using the following rules:

1. Corrected typos and spelling errors
2. Nouns only
    a. No adjectives (e.g., "scratched stone" = "stone")
    b. No noun adjuncts (e.g., "desert sand" = "sand")
3. Specific materials/stuff only (e.g., "landscape", "material", "don't know" = "none")
4. Resolve synonyms (e.g., "frozen water" = "ice")
5. Translate to English

The strong rank correlation between the frequency of common terms ($r_s = 0.74$, $p < 0.001$) indicated that multiple-choice responses capture the free-response ranking of material categories to a significant degree. Subsequent analyses of material judgements therefore focused on multiple-choice responses.

Experiment 1 general instructions

Original German: „Lieber Teilnehmer, in diesem Experiment siehst Du Bilder verschiedener Materialien. Bitte schau Dir jedes Bild genau an und gib an, welches Material Du siehst. Insgesamt gibt es 87 Bilder. In einem ersten Durchgang sollst Du zu jedem Bild die Materialnamen, die Dir zutreffend erscheinen in ein Textfeld schreiben. In einem späteren, zweiten Durchgang werden Dir die selben Bilder noch einmal gezeigt und Du sollst den zutreffendsten Materialnamen aus einer Liste auswählen und anschließend angeben, wie zufrieden Du mit dieser Auswahl bist. In einem dritten Durchgang werden Dir die selben Bilder noch einmal gezeigt und Du sollst angeben wie weit Du glaubst, dass die Kamera im Moment der Aufnahme von dem Material entfernt war (z.B. 9 cm, 400 m, 12 km). Die Dauer des Experiments beträgt etwa 1,5 Stunden."

English translation: Dear participant, in this experiment you see pictures of different materials. Please take a close look at each picture and indicate which material you see. There are 87 images in total. In the first step you should write the material names for each picture that appear to be appropriate in a text field. In a later, second round, you will be shown the same pictures again and you should select the most appropriate material name from a list and then state how satisfied you are with this selection. In a third run you will be shown the same pictures again and you should indicate how far you think the camera was away from the material at the moment of the recording (e.g., 9 cm, 400 m, 12 km). The duration of the experiment is about 1.5 hours.

Experiment 1 free-response instructions

Original German: „Aus welchem Material besteht obiges Bild? Gib alle Materialnamen an, die Dir einfallen und das Material möglichst genau beschreiben. Wie überzeugt bist Du davon, dass das Bild wirklich dieses Material zeigt? (Wobei 1 bedeutet: sehr wenig überzeugt; 7 bedeutet: sehr überzeugt)"

English translation: What material is the above picture made of? Enter all material names that you can think of and that describe the material as precisely as possible. How convinced are you that the picture really shows this material? (Where 1 means: very little convinced; 7 means: very convinced)

Experiment 1 multiple-choice instructions

Original German: „Aus welchem Material besteht obiges Bild? Wähle aus dem Dropdown-Menü. ['Erde', 'Glas', 'Haar', 'Haut', 'Holz (Holz)', 'Holz (Rinde)', 'Laub', 'Leder', 'Lehm', 'Metall', 'Moos', 'Papier', 'Pflanzen', 'Plastik', 'Sand', 'Stein (Beton)', 'Stein (Keramik)', 'Stein (Marmor)',

'Stein (Stein)', 'Tapete', 'Textil (Stoff)', 'Textil (Teppich)', 'Wandfarbe', 'Wasser (Eis)', 'Wasser (Schnee)', 'Wasser (Wasser)'] Wie überzeugt bist Du davon, dass das Bild wirklich dieses Material zeigt? (Wobei 1 bedeutet: sehr wenig überzeugt; 7 bedeutet: sehr überzeugt)“

English translation: What material is the above picture made of? Select from the drop-down menu. ['soil', 'glass', 'hair', 'skin', 'wood', 'bark', 'leaves', 'leather', 'clay', 'metal', 'moss', 'paper', 'plants', 'plastic', 'sand', 'concrete', 'ceramic', 'marble', 'stone', 'wallpaper', 'fabric', 'carpet', 'paint', 'ice', 'snow', 'water'] How convinced are you that the picture really shows this material? (Where 1 means: very little convinced; 7 means: very convinced)

Experiment 1 distance-estimate instructions

Original German: „Was ist der Abstand zwischen der Kamera und der Oberfläche? Maßeinheit auswählen. ['Mikrometer', 'Millimeter', 'Zentimeter', 'Meter', 'Kilometer'] Welcher Wert in dieser Maßeinheit (z. B. 1,2 Zentimeter oder 2,5 Kilometer)?”

English translation: What is the distance between the camera and the surface? Select unit of measure. ['micrometer', 'millimeter', 'centimeter', 'meter', 'kilometer'] Which value in this unit of measurement (e.g. 1.2 centimeters or 2.5 kilometers)?

Experiment 2A instructions (far group)

Original German: „Lieber Teilnehmer, in diesem Experiment siehst Du Bilder verschiedener Materialien. Bitte schau Dir jedes Bild genau an und gib an, welches Material Du siehst. Insgesamt gibt es 87 Bilder. In einem ersten Durchgang sollst Du zu jedem Bild die Materialnamen, die Dir zutreffend erscheinen in ein Textfeld schreiben. In einem späteren, zweiten Durchgang werden Dir die selben Bilder noch einmal gezeigt und Du sollst den zutreffendsten Materialnamen aus einer Liste auswählen und anschliessend angeben, wie zufrieden Du mit dieser Auswahl bist. Berücksichtige, dass die Kamera im Moment der Aufnahme SEHR WEIT von dem Objekt/der Materialoberfläche entfernt war. Die Dauer des Experiments beträgt etwa 1 Stunde.”

English translation: Dear participant, in this experiment you see pictures of different materials. Please take a close look at each picture and indicate which material you see. There are 87 images in total. In the first step you should write the material names for each picture that appear to be appropriate in a text field. In a later, second round, you will be shown the same pictures again and you should select the most appropriate material name from a list and then state how satisfied you are with this selection. Take into account that the camera was VERY FAR away from the object / material surface at the moment of recording. The duration of the experiment is about 1 hour.

Experiment 2A instructions (near group)

Original German: „Lieber Teilnehmer, in diesem Experiment siehst Du Bilder verschiedener Materialien. Bitte schau Dir jedes Bild genau an und gib an, welches Material Du siehst. Insgesamt gibt es 87 Bilder. In einem ersten Durchgang sollst Du zu jedem Bild die Materialnamen, die Dir zutreffend erscheinen in ein Textfeld schreiben. In einem späteren, zweiten Durchgang werden Dir die selben Bilder noch einmal gezeigt und Du sollst den zutreffendsten Materialnamen aus einer Liste auswählen und anschliessend angeben, wie zufrieden Du mit dieser Auswahl bist. Berücksichtige, dass die Kamera im Moment der Aufnahme SEHR NAH von dem Objekt/der Materialoberfläche entfernt war. Die Dauer des Experiments beträgt etwa 1 Stunde."

English translation: Dear participant, in this experiment you see pictures of different materials. Please take a close look at each picture and indicate which material you see. There are 87 images in total. In the first step you should write the material names for each picture that appear to be appropriate in a text field. In a later, second round, you will be shown the same pictures again and you should select the most appropriate material name from a list and then state how satisfied you are with this selection. Take into account that the camera was VERY NEAR to the object / material surface at the moment of recording. The duration of the experiment is about 1 hour.

Experiment 2B instructions

Original German: „Lieber Teilnehmer, in diesem Experiment siehst Du Bilder verschiedener Materialien. Bitte schau Dir jedes Bild genau an und gib an, welches Material Du siehst. Insgesamt gibt es 87 Bilder. In einem ersten Durchgang sollst Du zu jedem Bild die Materialnamen, die Dir zutreffend erscheinen in ein Textfeld schreiben. In einem späteren, zweiten Durchgang werden Dir die selben Bilder noch einmal gezeigt und Du sollst den zutreffendsten Materialnamen aus einer Liste auswählen und anschliessend angeben, wie zufrieden Du mit dieser Auswahl bist. Die Dauer des Experiments beträgt etwa 1 Stunde."

English translation: Dear participant, in this experiment you see pictures of different materials. Please take a close look at each picture and indicate which material you see. There are 87 images in total. In the first step you should write the material names for each picture that appear to be appropriate in a text field. In a later, second round, you will be shown the same pictures again and you should select the most appropriate material name from a list and then state how satisfied you are with this selection. The duration of the experiment is about 1 hour.

List of Publications

Cheeseman, J. R., Ferwerda, J. A., Maile, F. J., & Fleming, R. W. (2021). Scaling and discriminability of perceived gloss. Journal of the Optical Society of America A, 38(2), 203–210. https://doi.org/10.1364/JOSAA.409454

Cheeseman, J. R., Ferwerda, J. A, Morimoto, T., & Fleming, R. W. (2024). Gloss discrimination: Towards an image-based perceptual model. PsyArXiv Preprint. https://doi.org/10.31234/osf.io/anx4q

Cheeseman, J. R., Fleming, R. W., & Schmidt, F. (2022). Scale ambiguities in material recognition. iScience, 25(3), 103970. https://doi.org/10.1016/j.isci.2022.103970

Declaration

"I hereby declare that I have prepared the thesis at hand independently and without undue aid or the use of any resources other than indicated within the thesis. All parts of my thesis taken either verbatim or analogously from the published or unpublished works of or based on oral communications with others are indicated as such. Regarding all aspects of my scientific enquiries as they appear in my thesis, I have upheld the tenets of good scientific practice as laid out in the "Satzung der Justus-Liebig-Universität Gießen zur Sicherung guter wissenschatlicher Praxis" and complied with the precept of ethics, data protection and animal welfare. I declare that I have neither directly nor indirectly given monetary or any other valuable considerations to others in connection with the thesis at hand. I declare that I have not presented the thesis at hand, either in an identical or similar form, to an examination office or agency in Germany or any other country as part of any examination or degree. All materials from other sources as well as all works performed by others used or directly referenced within the thesis at hand have been indicated as such. In particular, all persons involved directly or indirectly in the development of the thesis at hand have been named. I agree with the screening of my thesis for plagiarism via offline or online detection-software."