# Interaction improves perception of gloss

**Matthias Scheller Lichtenauer**
Empa, Laboratory for Media Technology, Swiss Federal Institute for Materials Science and Technology, Dübendorf, Switzerland

**Philipp Schuetz**
Empa, Laboratory for Electronics, Metrology and Reliability, Swiss Federal Institute for Materials Science and Technology, Dübendorf, Switzerland

**Peter Zolliker**
Empa, Laboratory for Electronics, Metrology and Reliability, Swiss Federal Institute for Materials Science and Technology, Dübendorf, Switzerland

**Rendering materials on displays becomes ubiquitous in industrial design, architecture, and visualization. Yet the experience of the material from other modes of perception is missing in that representation. This forces observers to rely on visual cues only while judging material properties. In the present study, we compare judgments of rough and glossy surfaces by interacting and passive observers. We investigate whether observers actively exploring rendered stimuli judge properties differently than observers passively watching renderings. Resulting interobserver agreement is significantly higher for interacting observers.**

## Introduction

Visually perceivable material properties beyond lightness and color have gained attention lately. A special issue of this journal was dedicated to that subject; see Maloney and Brainard (2010). On the one hand, researchers study cues that contribute to the perception of material properties. On the other hand, researchers also investigate how the human brain integrates cues to notions like glossiness, roughness, or other percepts. Furthermore, differences in perception of physical stimuli and their representations rendered on displays become subjects of research not only due to increasing ubiquity of renderings but also because of a combination of rendered and physical objects within the same field of view in augmented reality applications. In the present work, we investigate the role of interaction in the perception of gloss on rough, moving surfaces.

Hunter (1975) provides an often-referenced typology of glossiness. The notion of glossiness describes the perception of spatial distributions of illumination interacting with distributions of reflection properties of a material across the surface, angles of incidence, and viewing angles. Roughness describes the variation of surface orientation and elevation as a function of distance orthogonal to a mean surface normal in a neighborhood. Perception of gloss and roughness therefore depend on the spatial resolution of the sensory systems involved.

In the literature, glossiness and roughness of surfaces are often described as perceptually linked phenomena. Ho, Landy, and Maloney (2008) investigated joint judgments of glossiness and bumpiness and found a positive correlation that could be fitted by a linear model. Their experiments were extended by Marlow, Kim, and Anderson (2012), who varied both illumination and bumpiness. The positive correlation found by Ho et al. appeared with some combinations in Marlow et al.'s (2012) data although there were negative correlations or nonmonotonic relationships with other combinations. Marlow et al. (2012) could explain most of the variance of their glossiness data by a linear weighting of other perceptual cues independently judged with the same stimuli. The stimuli in both contributions were static images rendered from combinations of ellipsoids; hence, all visible surfaces were convex.

Methven and Chantler (2012) hypothesized that the effect of binocular disparity to enhance perceived glossiness would break down with increasing roughness, but they designated their results as not yet conclusive. In contrast to Marlow et al. (2012), the

Citation: Scheller Lichtenauer, M., Schuetz, P., & Zolliker, P. (2013). Interaction improves perception of gloss. *Journal of Vision*, *13*(14):14, 1–13, http://www.journalofvision.org/content/13/14/14, doi:10.1167/13.14.14.

doi: 10.1167/13.14.14 Received June 17, 2013; published December 18, 2013 ISSN 1534-7362 

algorithm they used to create surfaces included concavities and discontinuities of the surface normal at ridges.

Gloss was also investigated with regard to distinction between reflectance and illumination, for matte surfaces an aspect of color constancy. Obein, Knoblauch, and Viéot (2004, p. 719) proposed the term "gloss constancy" for a postulated human ability "to compensate for luminous flux variations due to a change in angle of illumination and to maintain an invariable gloss percept, typical for the sample itself." Fleming, Dror, and Adelson (2003) investigated natural distributions of illumination across angles and concluded that human judgment of surface reflectance properties improved when apparent illumination coincided with patterns learned during one's lifetime. This finding is supported by recent results of Faisman and Langer (2013b), who let observers estimate ridge and valley positions in a relief rendered using different types of illumination. Marlow, Kim, and Anderson (2011) wondered how the visual system could distinguish gloss from surface variations due to pigmentation or illumination. As an answer, they provided evidence that congruence of highlight orientation and brightness gradient with the shading due to diffuse reflections is needed for a glossy appearance. Anderson and Kim (2009) had previously shown that histograms would not be sufficient to describe glossiness as histograms ignore exactly these neighborhood relationships.

Because gloss depends on viewing and illumination angle, cues provided by motion and binocular disparity (stereo viewing) are of interest. Nishida and Shinya (1998) let observers match diffuse and specular reflectance between achromatic, moving surfaces of different shapes rendered on a display. They found that surface properties are easier to match if shapes are identical up to the depth of the relief. Wendt, Faul, Ekroll, and Mausfeld (2010) replicated and extended the findings of Nishida and Shinya with chromatic stimuli and further established that the hue difference between highlight and diffuse reflection would enhance constancy of gloss perception. Sakano and Ando (2010) compared effects of monocular view, including perspective change due to head motion, with binocular disparity. They found the isolated effects of head motion and binocular disparity cues on perceived gloss to be of equal magnitude. Adding head motion to binocular view led to higher perceived glossiness. Sakano and Ando remarked that not only disparity of highlight position (Blake & Bülthoff, 1990) but, additionally, the difference in intensity between the two eyes and between viewing positions would be informative. Marlow et al. (2012) let their observers judge glossiness of both monocularly and binocularly rendered stimuli. Highly bumpy surface reliefs appeared glossier when rendered monocularly, but their stimuli were static. To our knowledge, the combined effects of motion, binocular disparity, and depth of relief have not been systematically studied yet.

Doerschner et al. (2011) attached reflected patterns as a matte texture to a surface. While objects were moving, observers were able to disambiguate between such painted reflection and a shiny reflecting object, but they were at chance performance with static stimuli. Analyzing the underlying image sequences, Doerschner et al. (2011) formulate a model based on optic flow that is able to explain most of their observations. Discontinuities of this flow due to ridges and troughs of the surface are attributed a crucial role when judging moving stimuli. The rendered surfaces in the studies including motion were mostly curved on a megascale but smooth on a mesoscale (Ho et al., 2008). A rough surface at the mesoscale should disrupt the optical flow in a different way.

Constancy was also studied with regard to the perceptual combination of cues. Brenner, Granzier, and Smeets (2011) used object motion in combination with a colored surface texture to ensure that observers could reliably judge whether a change in color would be due to a change in illumination or due to a change in surface reflectance. This indicates that observers should also be able to attribute hue changes near moving highlights to illumination.

With regard to perception of renderings, not only visual cues, but also cross modal relationships are of interest. Muller, Brenner, and Smeets (2007) found that conflicting cues for slant and surface structure did not result in information loss. They concluded that, in matching two properties simultaneously, observers use the information relevant for one property even if it conflicts with perception of the other property. This is consistent with the findings of van Beers, van Mierlo, Smeets, and Brenner (2011) that haptic feedback of slant can influence weighting in a model of conflicting visual cues.

As a summary, researchers agree that binocular disparity, motion parallax, and congruence of specular and diffuse reflection as well as intensity and hue differences between diffuse and specular reflection provide cues to the perception of gloss.

Representations on displays have limitations relative to physical reality. Dynamic, time, and spatial resolution of an imaging system are mostly lower. Interactive frame rates set further limits on the accuracy of the representation. Surface variations at the microscale may, for instance, be approximated by random variations of surface normals across a planar facet; cast shadows may be missing, or intensities beyond the maximal luminance of the display, which are crucial in perception of gloss, may be clipped.

With the notable exception of Faisman and Langer (2013b), no study discusses the effects of occluding

Figure 1. Judgment of samples in natural viewing.

contours or cast shadows, which help to perceive the shape of the objects as further cues to surface orientation, position, and direction of the light sources. Doerschner, Maloney, and Boyaci (2010) studied perception of gloss on physical spheres. This enabled dynamic ranges that are not available on contemporary displays. Observers viewed stimuli monocularly and using a chin rest, so binocular disparity and motion cues were eliminated. They reported a higher perceived glossiness in front of a black background, an effect that has not been reported in studies on displays.

Cues on weight, temperature, hardness, roughness, shape, position, or movement of an object resulting from touch and perception of body position are altered or missing on displays relative to physical reality (Doerschner et al., 2011; Lederman & Klatzky, 2009). Users mostly interact by a proxy (mouse, joystick) that alters all these experiences relative to ecological conditions. With the exception of Sakano and Ando (2010), observers passively experienced object motion in the studies cited so far although users can interactively move the rendered surfaces in most design applications. In recent studies on photographic reproduction quality of glossy, rough surfaces embedded in high dynamic range scenes, we could not make out perceptual differences between judgments of videos and judgments of still images (Sprow, Kuepper, Barańczuk, & Zolliker, 2013). Observers remarked they would like to stop motion at certain points. Formulated as a hypothesis to falsify, interactively moving rendered objects would not help to better judge glossiness of rough surfaces.

## Methods

In this paper, two experiments compare visual perception of physical stimuli with perception of renderings on displays. In the first experiment, different renderings of a surface geometry digitized with a 3-D laser scanner had to be attributed to the most similar of four physical reference samples. In the second experiment, videos taken from a test set of physical samples under different illuminations had to be attributed to the most similar of the same four physical reference samples as in the first experiment.

The physical samples were all cut out of the same plastic flowerpot imitation because this material was of uniform color and had no texture. All samples had a surface relief, which imitated the traces of a mold used in production of clay flowerpots. The traces would be left when taking the form away. The surface relief can verbally be described as longitudinal grooves in a mostly flat surface, interspersed with small round pits. Sample size was about 4 cm $\times$ 4 cm $\times$ 3 mm. The surface geometry was acquired with a 3-D laser scanner from one particular physical sample. That sample was excluded from further use in experiments to make sure that the surface relief would not allow identification.

We applied three different sprayed lacquers to pairs of samples; one pair was left untreated. The pairs were then separated to form a reference set and a test set of four samples each. Gloss was intended to be the same for samples in each pair but differing between pairs (Figure 2). We verified this in natural viewing. Natural viewing means a table was placed in a room with windows facing to the northwest, so daylight entered in

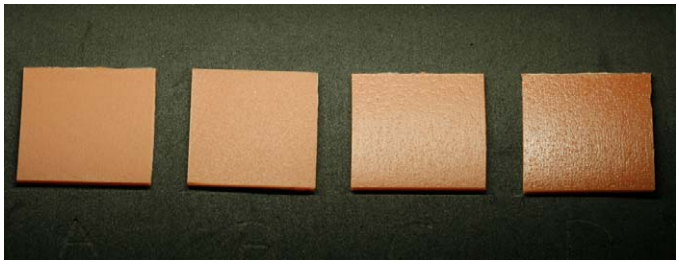

Figure 2. Close-up of physical test samples.

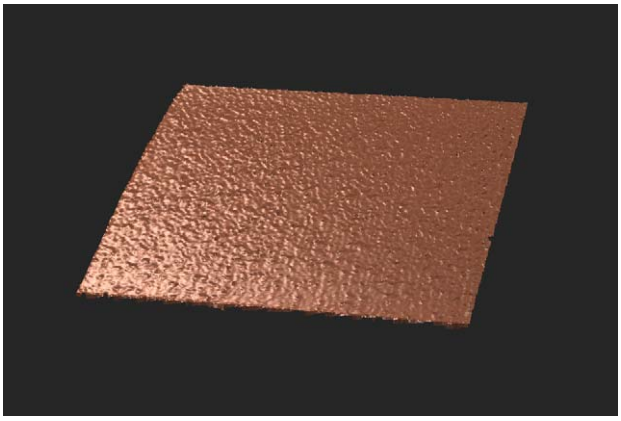

Figure 3. Close-up on the sample geometry rendered with the Phong (1975) model. In the experiment, we used a neutral background of 55.4 cd/m$^2$.

a diffused way. A LED lamp was used as an additional, directed light source (see Figure 7 for a spectrum of the LED).

The experimental setup is displayed in Figure 1. We laid out the four samples of the reference set in a line on the table. Positions within the line were randomized. The lamp was placed above the center of the four reference samples. The height of the standing table was adjusted for each observer so that their eyes were level with the lower end of the lamp casing when standing upright. Observers were free to alter their position relative to the samples, so adjusting the table's height assured that each observer could choose from the same viewing angles relative to the samples and light source.

Observers were first instructed to move freely on their side of the table but not to touch the samples because more than one observer instinctively tried to grasp a sample when approaching the table. This instruction avoided altering the position of samples relative to the light source and also prevented the use of nonvisual information because we planned to exclude those ways of exploration in the following experiments.

Figure 4. Recording phase of Experiment 1.

Figure 5. Replaying phase of Experiment 1.

One sample from the test set was presented in each trial. The observer had to indicate which sample from the reference set was most similar to the test sample. The test sample was taken away after each trial to a place not visible to the observer because test samples were randomly chosen, so it was possible that an observer judged the same test sample more than once, even twice, in a row. No observer passed more than four trials to avoid learning effects.

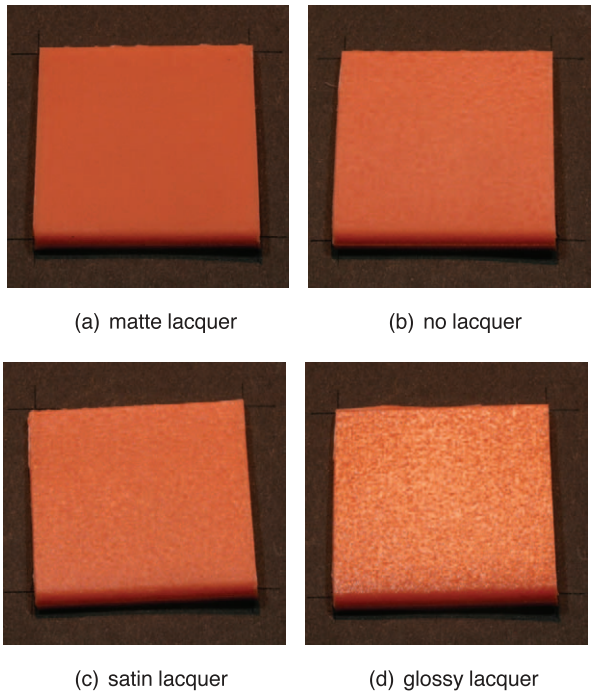


Figure 6. Detail view of the initial frame in Experiment 2 captured under diffuse daylight simulation of XRite Spectralight with additional directed LED illumination on black backing. Compare to Figure 2.

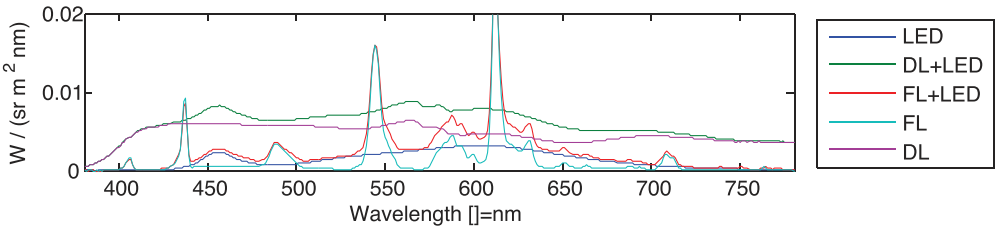


Figure 7. Illumination spectra used for taking pictures in Experiment 2. See Table 3 for luminances and chromaticities.

To present results, we will use the following notation. Let the number of classes (here: reference samples) be fixed as $K$. Let $n_{ij}$ denote the number of times a sample was denoted as belonging to category $i$ by one classifier (here: observer) and to category $j$ by a different classifier (here: known lacquering). The total number of paired judgments or classifications is then $N = \sum_{ij} n_{ij}$.

The results of this verification and the later experiments will be presented as contingency tables comparing two classification procedures. In a $K \times K$ contingency table, each cell referenced by row index $i$ and column index $j$ contains $n_{ij}$ or the relative frequency $p_{ij} = n_{ij}/N$. The row sums $p_{i+} = \sum_i p_{ij}$ and the column sums $p_{+j} = \sum_j p_{ij}$ are the marginal frequencies for the respective classifier.

Contingency for the verification is displayed in Table 1. Some confusion occurs only between the sample treated with matte lacquer and the untreated samples. This is not surprising as Hunter (1975) recommends using grazing illumination and viewing angles of 70° to surface normal for low gloss samples. To reduce this source of confusion, we used a more directed illumination of reference samples with an incident angle of about 45° during the following experiments.

We asked observers in the verification task to sort samples for gloss. This was done after the task to avoid priming for gloss as a criterion of similarity. The results after 10 observers corresponded to an almost unanimous ordering A < B < C < D with only one observer setting B < A. Using this ordering in Table 1 leads to the observation that most of the probability mass is near the diagonal elements this way. This indicates that observers based their similarity judgments mostly on glossiness. An unambiguous ordering for roughness could not be established with the same procedure. Several observers made remarks about being biased by glossiness when judging roughness.

Several indices express reliability or agreement of observers. Most of them use as a null hypothesis that the judgments would be randomly drawn from marginal distributions (Banerjee, Capozzoli, McSweeney, & Sinha, 1999; Cohen et al., 1960; Fleiss & Cohen, 1973). However, prevalence, $p_{i+} \neq (1/K)$; marginal inhomogeneity, $p_{i+} \neq p_{+j}$; and the sample size $N$ can influence some of these indices. Bhapkar (1979) presented chi-square tests for marginal homogeneity. The assumption of marginal homogeneity for the data in Table 1 cannot be rejected ($p = 0.5497$).

Percentage of agreement $A$ is the proportion of the elements on the diagonal $A = (1/N)\sum_{ij} n_{ij} = 0.8625$ for Table 1. Cohen et al.'s (1960) kappa assuming nominal data is $\kappa_n = 0.8145$ for Table 1. Observers producing such results would usually be considered to be very reliable. We can thus assume that confusion is not due to the physical samples, particularly not between satin and glossy lacquer. To reduce confusion between less glossy samples in the following experiment, we used a more directed illumination at a higher angle of incidence relative to the surface normal than in this preliminary verification task.

## Experiment 1

The digitized sample geometry was rendered on a display (EIZO CG220, 1920 × 1200 pixels) using OpenGL and the Psychtoolbox in version 3 (Kleiner, Brainard, & Pelli, 2007). The intensity _o of a particular surface element was determined as a contribution of ambient, diffuse, and specular light flows per light source using a model proposed by Phong (1975):

$$\mathbf{o} = w_a \boldsymbol{i_a} \boldsymbol{r_a} + \cos(\gamma) \boldsymbol{w_d} \boldsymbol{i_d} \boldsymbol{r_d} + \cos^p(\theta) \boldsymbol{w_s} \boldsymbol{i_s} \boldsymbol{r_s} \quad (1)$$

where $\boldsymbol{i}_*$ is the intensity of each inflow, $\boldsymbol{r}_*$ is surface reflectance, and $\boldsymbol{w}_*$ allows weighting contribution of each inflow, $\sum w^* = 1$. Indices stand for ambient,

| | | presented | | | | |
|---|---|---|---|---|---|---|
| attributed to | code | A | B | C | D | $n_{i+}$ |
| matte lacquer | A | 21 | 4 | 0 | 0 | 25 |
| no lacquer | B | 5 | 12 | 1 | 0 | 18 |
| satin lacquer | C | 0 | 0 | 19 | 1 | 20 |
| glossy lacquer | D | 0 | 0 | 0 | 17 | 17 |
| | $n_{+j}$ | 26 | 16 | 20 | 18 | N = 80 |

Table 1. Contingency table of samples in natural viewing.

| Set | $w_a$ | $w_d$ | $w_s$ | p |
|---|---|---|---|---|
| 1 | 0.5100 | 1.0000 | 0.0250 | 56 |
| 2 | 0.3671 | 0.5545 | 0.2367 | 96 |
| 3 | 0.5100 | 1.0000 | 0.0500 | 48 |
| 4 | 0.4756 | 0.4896 | 0.3080 | 90 |
| 5 | 0.5100 | 0.7000 | 0.0500 | 101 |
| 6 | 0.3357 | 0.9345 | 0.4647 | 23 |
| 7 | 0.4495 | 0.9756 | 0.1750 | 62 |
| 8 | 0.5879 | 0.7283 | 0.0983 | 57 |
| 9 | 0.4021 | 0.4832 | 0.1255 | 82 |
| 10 | 0.5100 | 0.7000 | 0.0250 | 97 |

Table 2. Parameter sets for Equation 1. $\boldsymbol{r}_a$ , $\boldsymbol{r}_d = \mathrm{RGB}(0.41, 0.23, 0.19)$, $\boldsymbol{r}_s = \mathrm{RGB}(1, 1, 1)$

diffuse, and specular flow, respectively. The angle $\gamma$ is the angle of incidence to the surface normal, and $\theta$ is the angle between direction of specularly reflected light and the viewing direction. The exponent $p$ is used to vary the angular extent of the specularly reflected light. Vectorial notation just means that these calculations are made for each channel (here: R, G, B) separately; hence, $\boldsymbol{i}_*\boldsymbol{r}_*$ denotes multiplication per component in Equation 1. One should note that diffuse and ambient contributions are independent of the observer position in this model. The $\boldsymbol{i}_*$ was set to (1, 1, 1) for each flow, and $\boldsymbol{r}_*$ was kept constant during Experiment 1. For each observer, we used the same 10 settings of parameters $w_a$, $w_d$, $w_s$, $p$, differing between trials (Table 2).

We randomly drew parameter values from restricted ranges but then arbitrarily fixed four of them (sets 1, 3, 5, and 10) to have very low specular reflection with high ambient and diffuse reflection. We did not intend to make a rendered representation look like a particular test sample, but we fixed $r*$ to approximate the reflectance of the sample measured with XRite i1 spectrometer (45°/0°).

Experiment 1 was conducted in two phases. In the first phase, observers could rotate the rendering around the horizontal and the vertical axes with the help of the computer mouse (Figure 4). The orientations of the rendering in the virtual space were recorded, and we will therefore refer to the first phase as "recording." After an acoustic countdown of 5 s followed by 12 s interaction time, the observers had to attribute the rendering to the most similar of the four physical samples from the reference set presented below the display. Observers were informed that there would be no time limit for the attribution and time to decision would not be recorded.

The second phase of Experiment 1 was identical to the first phase, but users could not interact with the sample; instead, they saw the sample move as in a video. The movement was one of those recorded in the first phase from a different observer; hence, the visual content was identical, and only the viewing mode changed from interactive to passive (Figure 5). We will refer to the second phase as "replaying." The only difference between these two phases was the mode of action, i.e., interactive observers in the first phase and passive observers in the second phase. The renderings could be viewed from frontoparallel up to the oblique angles in which the physical samples were presented, but due to the different position of the illumination in physical space and virtual space (behind the observer), a symmetric match was not possible.

We had 27 observers in three groups. A first group of nine observers was recruited from the Laboratory of Media Technology, includes two of the authors, and can be considered as experts in visual testing. They participated in the preliminary natural viewing task, recording, and replaying. A second group of nine observers participated only in recording, and a third group of nine observers participated only in replaying. The latter two groups can be considered laypersons in visual testing. So there were an equal number of experts and laypersons in both phases. Lay observers volunteered for the task; the members of our lab did it as part of their work. Observers were informed that we would not know of a method to objectively determine a right or wrong answer but that we were interested in homogeneity of judgments with regard to design guidelines for visualization systems.

## Experiment 2

In the second experiment, we tested contributions of diffuse and specular reflections to observer performance. We used a two-axis rotational stage mounted in a viewing booth as used in the graphical industry (X-Rite Spectralight QC) to produce stop-motion videos of the test samples under reproducible conditions. Videos of each of the four test samples were made in three illumination conditions:

- Diffuse and directed illumination
- Diffuse illumination only
- Directed illumination only

As a directed illumination, we used the same LED lamp as in the preliminary verification task; for a spectrum, see Figure 7. The viewing angle relative to the illumination was fixed, and the same sample orientations were used for all illumination conditions and all samples.

The viewing angle and angles of incidence relative to the vertical were below 45°; i.e., the samples were illuminated and photographed from above with the surface normal of the sample describing a motion in the form of the ∞ symbol.

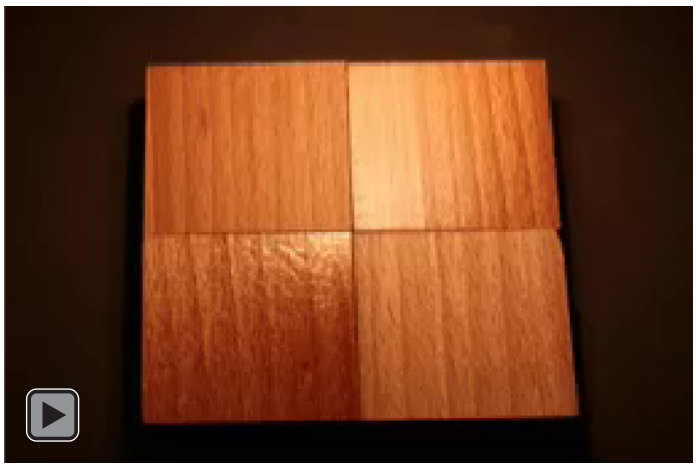

Figure 8. A downsized version of the learning video used in Experiment 2. The size of the learning video and the test videos was 1080 × 720 pixels.

We used a gray and a black cardboard as a background. With the black background, we used the diffuse daylight simulation of the viewing booth, and on the gray background, we used the more yellowish light (FL 3) provided by fluorescent tubes.

We presented the videos on the same EIZO CG220 display we used in Experiment 1. The refresh rate of this LCD is at 60 Hz, so we intended to have 60 frames per second. Psychtoolbox reported less than 1% of the presentation deadlines missed. Maximal emission of this display is 120 cd/m$^2$, and RGB (128, 128, 128) is at 55 cd/m$^2$. These values were measured from a 8 × 8 cm patch centered on the black background with a Konica Minolta CS2000 spectroradiometer that was positioned 45 cm in the direction of the surface normal from the display center. The black matte material on which the physical samples were presented measured from the same position and illuminated as during both Experiments 1 and 2 at 85 cd/m$^2$, and the flat backside of an untreated sample was at 390 cd/m$^2$. The luminance of the GretagMacbeth ColorChecker (McCamy, Marcus, & Davidson, 1976) Neutral 5 patch at the same position was at 484 cd/m$^2$; the Neutral 6.5 patch was at 903 cd/m$^2$.

We used two fluorescent tubes to illuminate the physical samples, 10 cm above the samples. We build a box around the samples with length × depth × height = 21 cm × 6.5 cm × 8 cm, made of grey cardboard with the floor and front walls missing. Illumination entered through a 19 cm × 1.5 cm slit in the top cover on which the lamp casing was solidly seated. The box served as both a baffle and a neutral background. In Figures 4 and 5, this box is removed for visualization purposes, but the pictures in Figure 2 were made with the box above the samples.

We informed observers about the lacquering procedure and that they would see the test set under various illuminations. The observer's task was to attribute the video of the test sample to a sample from the reference set they perceived as identically lacquered. The reference set was presented below the display in the same way as in Experiment 1. We arranged the reference set in Experiment 2 in order of gloss ascending from left to right. We randomized the order of video presentations because learning effects might occur. For this reason, we also did not provide feedback to the observers about whether their choice was correct in the sense that they attributed the test sample in the video to the reference sample treated with the same lacquer.

Observers could interact with the videos (1080 × 720 pixels centered on a RGB 128, 128, 128 background of a 1920 × 1200 pixel display) by slowing them down, reversing play in time, and stopping/restarting at any frame by pressing single keys on a keyboard. Interaction time was not limited, and observers could learn interaction possibilities with a preliminary video of the same motion. This video simultaneously showed four pieces of beech wood each treated with a different lacquer as were the plastic pieces of the reference set used in both Experiment 1 and in one; hence we were priming observers for gloss (Figure 8).

Eight expert observers provided eight observations per illumination condition (four lacquers × two backgrounds). All observers had participated in the previous experiments as well. Presentation was organized in four sessions in such a way that observers could not base their judgments on exclusion principles because they knew that each combination was presented once. We let observers pause several days between sessions for the same reason. To further avoid learning effects, we randomized the order of presentation and alternated between the gray and black backgrounds. Only three out of four movies would be presented in the first three sessions with all movies chosen from the same illumination condition (although, as said, diffuse illumination changed with the backgrounds). In the fourth session, the movies not yet presented to the observer were presented in random order but with backgrounds still alternating. We informed observers prior to the fourth session that they would now be facing both changing illumination conditions and changing reflection properties at the same time.

## Results

### Results of Experiment 1

For each of 10 sets of parameters of Equation 1, we collected 18 judgments in recording and 18 judgments

| Illumination | Luminance | CIE x | CIE y |
|---|---|---|---|
| LED | 165.3 cd/m$^2$ | 0.42 | 0.39 |
| FL | 204.2 cd/m$^2$ | 0.44 | 0.40 |
| DL | 390.0 cd/m$^2$ | 0.31 | 0.33 |
| FL + LED | 341.4 cd/m$^2$ | 0.43 | 0.39 |
| DL + LED | 552.7 cd/m$^2$ | 0.34 | 0.35 |

Table 3. Illumination characteristics in Experiment 2. *Notes*: Spectra were measured from the Neutral 6.5 patch of a GretagMacbeth ColorChecker replacing the samples in Figure 6, and the Konica Minolta CS2000 spectroradiometer (1° angular aperture) replaced the camera. Compare to Figure 7. Values refer to the 2° normal observer model.

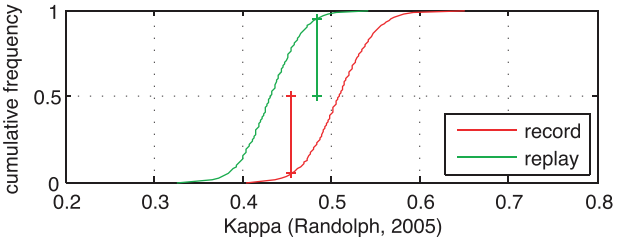


Figure 9. Distributions of Randolph's (2005) $\kappa_f$ in Experiment 1. Per phase, 180 observations were resampled 10,000 times. Note that the data labeled "record" are resampled only from those recordings picked to be replayed.

in replaying, resulting in 180 observations for each phase. Our null hypothesis was that observers in recording and replaying would agree equally often. Testing the hypothesis based on the established models of Fleiss (1971) or Fleiss and Cohen (1973) is not recommended if marginal homogeneity is not given (Feng, 2012; Jakobsson & Westergren, 2005; Sim & Wright, 2005). But the marginal distributions of the data in Table 4 were significantly different ($p = 0.0092$) as verified by Bhapkar's (1979) test.

Randolph's (2005) $\kappa_f$ extended Fleiss' (1971) $\kappa_f$ for the case that categories are nominal and multiple raters judge an equal number of items, but their marginal distributions do not have to be fixed. We therefore used Randolph's $\kappa_f$ to test the hypothesis. We applied resampling methods to generate confidence intervals for Randolph's $\kappa_f$ because asymptotic approaches are questionable (Efron & Tibshirani, 1986; McKenzie et al., 1996). Randolph's $\kappa_f$ is not calculated from a contingency table but from the list of observations the contingency table is based on. To generate such lists, we picked randomly the same number of observations for each Phong (1975) parameter set from the list of observations, resulting again in 180 observations per resample. This resampling procedure was repeated 10,000 times to generate confidence intervals. Randolph's $\kappa_f$ in Figure 9 rejects our null hypothesis that interacting does not alter agreement in judgment of rough and glossy samples. We used the criterion that the 5% quantile in recording would be above the median from replaying.

We analyzed the sample orientations used by interacting observers. The relative frequencies of orientations are plotted as functions of the rotation angles around the horizontal and the vertical axes in Figure 10. Frequencies are on a log scale and were smoothed with a Gaussian. For each set of the 10 sets of rendering parameters, we also plot the frequency of attribution to one of the reference samples. As a general picture, observers look at a small region near the point of specular reflection for renderings that are attributed to glossier reference samples while probing larger regions for less glossy samples. We analyzed angular velocity of rotations in recording as well. The distribution of Euclidean difference between angular positions at time intervals of 0.1 s is highly skewed toward zero. Observers hardly changed the orientation of the rendered object for 20% or more of the interaction time of 12 s—even when ignoring data from the initial 0.5 s. If we accept an ordering of the physical samples for gloss and attribute from one point for the most matte sample to four points for the glossiest sample, the mean is highly negatively correlated with quantiles of the motion statistics (Table 5). This supports our observation that observers tend to make larger movements with matte samples.

| | | recording | | | | |
|---|---|---|---|---|---|---|
| replaying | code | A | B | C | D | $n_{i+}$ |
| matte lacquer | A | 14 | 14 | 3 | 0 | 31 |
| no lacquer | B | 7 | 11 | 13 | 1 | 32 |
| satin lacquer | C | 7 | 16 | 15 | 9 | 47 |
| glossy lacquer | D | 1 | 5 | 22 | 42 | 70 |
| | $n_{+j}$ | 29 | 46 | 53 | 52 | N = 180 |

Table 4. Contingency table of judgments in recording versus replaying.

## Results of Experiment 2

We present results here as contingency Tables 6, 7, and 8, one for each illumination condition. For this, judgments of samples on gray and black backgrounds

| Quantile | *v25* | *v50* | *v90* | *a25* | *a50* | *a90* |
|---|---|---|---|---|---|---|
| Correlation | −0.88 | −0.93 | −0.87 | −0.86 | −0.90 | −0.92 |

Table 5. Correlations between quantiles of object motion statistics for velocity *v*, acceleration *a*, and the mean attribution in Figure 10.

| | A | B | C | D | $n_{i+}$ |
|---|---|---|---|---|---|
| A | 13 | 0 | 0 | 0 | 13 |
| B | 1 | 8 | 3 | 0 | 12 |
| C | 2 | 7 | 10 | 2 | 21 |
| D | 0 | 1 | 3 | 14 | 18 |
| $n_{+j}$ | 16 | 16 | 16 | 16 | N=64 |

Table 6. Contingency under LED illumination. *Notes:* Percentage of agreement = 0.7031. Bhapkar test for marginal homogeneity: $p$ = 0.135.

| | A | B | C | D | $n_{i+}$ |
|---|---|---|---|---|---|
| A | 14 | 2 | 0 | 0 | 16 |
| B | 2 | 10 | 1 | 0 | 13 |
| C | 0 | 4 | 12 | 1 | 17 |
| D | 0 | 0 | 3 | 15 | 18 |
| $n_{+j}$ | 16 | 16 | 16 | 16 | N=64 |

Table 8. Contingency under combined illumination. *Notes:* Percentage of agreement = 0.7969. Bhapkar test for marginal homogeneity: $p$ = 0.424.

were merged. Note that we already compensated for the differences in luminance levels of the illuminations by adapting the exposure time during photographic capturing of the stimuli.

The distributions of Randolph's (2005) $\kappa_f$ resulting from resampling indicates that combining directed illumination from an oblique angle with diffused illumination from above leads to enhanced reliability relative to any of the two illuminations alone (Figure 11). Similar results have been found by Fleming et al. (2003) and Faisman and Langer (2013a) with static stimuli rendered by ray tracing while we use dynamic stimuli rendered by photographic methods.

# Discussion

## Virtual and physical reality

It should be noted that the preliminary verification task is a type 1 task; i.e., a correct answer can be assumed to be reasonably well known. The experiments are type 2 tasks—knowledge of a correct answer may be questionable (Kingdom & Prins, 2010). We did not strive to maximize similarity of appearance between virtual test and physical reference stimuli according to some model so as to use a type 1 argumentation. Therefore, one cannot directly conclude that observers would as well judge glossiness of physical stimuli more reliably when allowed to interact with them.

However, interobserver reliability of judgments in both experiments was significantly different from zero. This indicates as well that there is at least a nonrandom correspondence between physical and virtual stimuli.

| | A | B | C | D | $n_{i+}$ |
|---|---|---|---|---|---|
| A | 12 | 6 | 2 | 0 | 20 |
| B | 4 | 5 | 5 | 2 | 16 |
| C | 0 | 5 | 7 | 2 | 14 |
| D | 0 | 0 | 2 | 12 | 14 |
| $n_{+j}$ | 16 | 16 | 16 | 16 | N=64 |

Table 7. Contingency under diffuse illumination. *Notes:* Percentage of agreement = 0.5625. Bhapkar test for marginal homogeneity: $p$ = 0.467.

## Interaction and surface properties

We could establish two main results with our experiments: Interacting observers agree more often than passive observers when comparing renderings of a surface geometry, and, as a second result, a combination of a nonlocal, diffuse light source with a directed, oblique light source also enhances interobserver reliability with photographic reproduction methods.

We assume that interacting observers had an advantage in the task because they could actively coordinate their visual exploration strategies with the movement of the rendered surface geometry. Furthermore, they had feedback from other modes about whether the rendered object was moving or not, and passive observers had to determine this from visual information only. It would be interesting to investigate this by tracking gaze direction in a replication of Experiment 1.

Gaze tracking would possibly allow investigating whether interactive observers were more involved, i.e., whether attentional effects could play a role.

We should also consider that our experiments and many experiments we discussed in the Introduction contain matching tasks. Matching tasks or paired comparisons with a reference result in saccadic movements between probes and the reference object. It is known that we are blind to some changes during saccadic movements, so it would be interesting to test whether this would also influence perception of changes in highlight positions during saccadic movements.

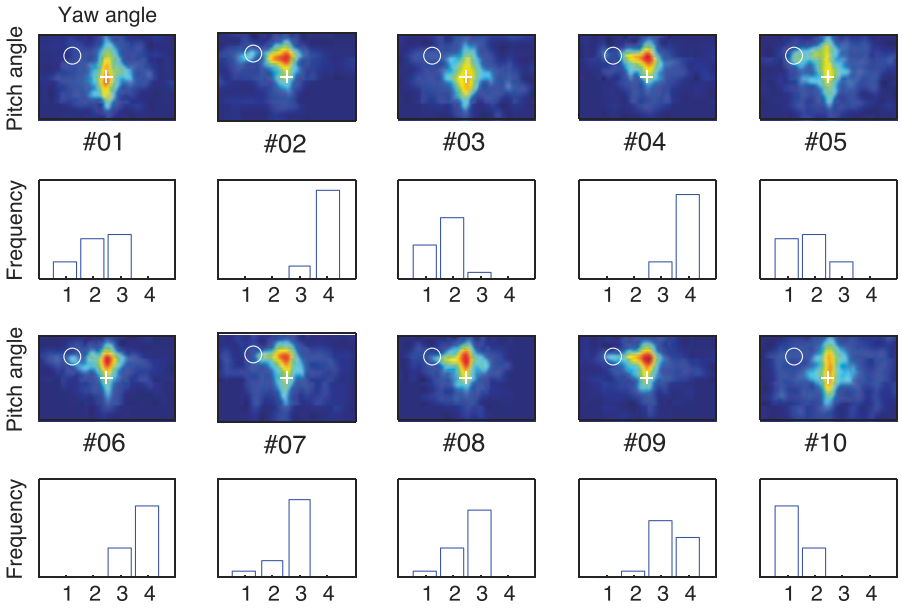


Figure 10. Relative frequencies of orientation and frequency of attributions to reference samples for each parameter set in the recording phase of Experiment 1. Note that sample numbering is ordered by glossiness so that reference sample 4 is glossiest. A white circle marks the region of orientations where a mirror would specularly reflect the point-like light source in the virtual camera. A white cross marks the initial orientation of the surface normal in the same vertical plane as the camera.

Again, this would necessitate coordinating gaze tracking and rendering.

Wendt et al. (2010), for instance, expressed their astonishment that lightness matching in one of their experiments was better when the stimuli did not move. The orientation of the light source in the respective experiment relative to the surface did not change; hence, the diffuse component remained constant, and the specular component changed with time. One should note that motion was perpetual in their experiment; hence, while observers did change their focus from the test stimulus to the reference stimulus, the position between observer and light source had changed. Because their shapes were not symmetric to the rotation axis, position, size, and intensity of highlights were likely to change while diffuse intensity stayed constant.

Interacting observers used a larger range of angles for samples with low gloss. This may be related to an observation of Hunter (1975) that low-gloss samples exhibit the largest differences when samples are illuminated and viewed under a large angle to the sample surface normal. Observers could not achieve this spatial configuration by interacting with the sample. From their verbal comments, we assume they were looking for the best alternative in a larger range of angles, but this assumption would yet have to be verified in an experiment allowing observers to vary their virtual position as well.

We are aware that our results apply to comparing physical with rendered stimuli. Rendering means appearance is limited by the properties of the rendering method and also by the display gamut and the spatial resolution of the display. Furthermore, test stimuli were rendered from a monocular perspective but viewed with both eyes, and binocular disparity was present for viewing the reference samples. As Sakano and Ando (2010) argue, there are two aspects to binocular viewing. The first is that the position of the highlights can be estimated from two perspectives; the second is that highlight intensity changes quickly with the angle. The first property, so they argue, may also be achieved by altering the head position. The fact that combining binocular disparity with head motion led to an even higher impression of gloss in their experiments

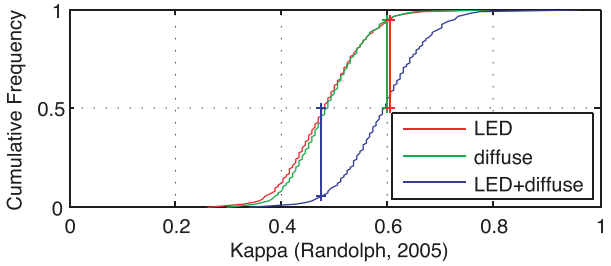


Figure 11. Distribution of Randolph's (2005) $\kappa_f$ in Experiment 2. Per illumination condition, 64 observations were resampled 10,000 times.

is a hint that the human visual system uses both cues simultaneously. However, as Marlow et al. (2012) documented, binocular disparity may not always lead to a higher glossiness with varying depth of surface relief. The influence of highlight intensity changes could be tested by setting a virtual, specularly reflected light source at an infinite distance. It might be tempting as well to analyze the optical flow with data from Experiment 1. However, without knowing where observers did look at each point in time, such data would be of limited use.

In ecological conditions, we may be able to alter the orientation of the object relative to the light source and alter our position relative to an object at the same time. In our Experiment 1, only the first mode of interaction was possible, and Sakano and Ando (2010) studied the second way of interaction. It would be interesting to see whether combination of both interaction possibilities would further enhance reliability of glossiness judgments.

Further limitations of interactive rendering as we used it, relative to many experiments we discussed in the Introduction, should not go unnoticed. The surfaces we rendered were rough on a scale within the resolution limits of the rendering system but mostly smooth on a scale that is usually associated with the term "shape." For many experiments cited, the shape was curved to a point at which occlusions occur, but surfaces were smooth at a mesoscale (Ho et al., 2008). Each surface triangle in Experiment 1 was rendered with a uniform surface normal, but, for instance, Ward's (1992) roughness parameter results in a perceptual correlate of roughness beyond the spatial resolution limits of the visual system. This is implicitly understood in the literature (Hunter, 1975; Ward, 1992). In our stimuli, no cast shadows or occlusions occurred, and they would further help in disambiguating a sequence of retinal images in a mental concept of an object with a surface.

Fleming et al. (2003) showed that observers were able to judge specular reflectance and roughness independently on rendered surfaces that were smooth within the resolution of the visual system. Our results in Experiment 2 extend their findings that natural or real-world illumination enhances judgments of surface reflectance properties like gloss or roughness to physical stimuli. Clearly, interobserver reliability in Experiment 2 is significantly better if both illuminations are mixed.

### Indices of agreement

In Experiment 1, we tested the hypothesis that interacting observers would agree equally often as passive observers in attributing a rendering to a physical sample. The physical samples can be ordered with regard to glossiness. The marginal distributions of Table 4 indicate that passive observers more often attribute the renderings to the glossiest of the physical samples than interacting observers. The marginal distributions are not homogeneous and not uniform. Bhapkar's (1979) test rejected the null hypothesis of marginal homogeneity in Table 4 with $p = 0.012$. This, by itself, indicates that the observers judged samples differently in the two conditions. Although the percentage of agreement and Cohen's $\kappa$ are most used in the literature, using them is not recommended in cases of nonuniform and nonhomogeneous marginal distributions as in Experiment 1 (Banerjee et al., 1999; Sim & Wright, 2005). As Randolph (2005) points out, using a null model of independently drawing from fixed but nonuniform marginal distributions creates confusion rather than providing information. One does not know whether disagreement stems from prevalence or from differences between the marginal distributions of the observers.

Using a uniform prior and no fixed margins assumption, Randolph (2005) assumes that each disagreement has the same importance, and observers have no preexisting assumptions on frequency of each category. The value of Randolph's $\kappa_f$ depends on the number of objects to be judged and the number of categories. Those parameters are identical in both phases of Experiment 1 and for all three illumination conditions of Experiment 2, so we can exclude this source of systematic bias.

The first observer group in Experiment 1 could be considered as experts. Interleaved with replaying, we let them additionally judge their personal recordings again. Percentage of agreement was higher in that task ($A = 0.56$) than between different observers in the recording task ($A = 0.54$). This confirmed our prior decision to exclude their own recordings from being replayed and, at the same time, indicates that the lesser agreement in the replaying phase was not due to fatigue or lack of attention and care. The agreement between experts and between lay observers was not significantly different in replaying. Apart from excluding their own recordings, we randomly chose the content to be replayed, respecting equal proportions for each set of Phong (1975) parameters.

## Conclusions

In our experimental situation, interacting observers agreed significantly more often than passive observers. The agreement was higher in the second experiment when the illumination contained both diffuse and directed components.

We conclude that interaction as well as natural distributions of illumination provide additional cues that allow observers to separate illumination and reflection properties in representations of rough, glossy materials on displays.

We suppose that the ability to coordinate stimulus motion with saccadic movements between test and match stimulus eases the matching task. But conclusion about this point should not be drawn without eye-movement data.

*Keywords: materials perception, gloss, hand-eye coordination*

## Acknowledgments

This research project was supported by a grant of the Swiss National Science Foundation. Pictures of observers were taken by Iris Sprow (Empa).

Commercial relationships: none.
Corresponding author: Matthias Scheller Lichtenauer.
Email: matthias.scheller_lichtenauer@alumni.ethz.ch.
Address: Empa, Laboratory for Media Technology, Swiss Federal Institute for Materials Science and Technology, Dübendorf, Switzerland.

## References

Anderson, B. L., & Kim, J. (2009). Image statistics do not explain the perception of gloss and lightness. *Journal of Vision, 9*(11):10, 1–17, http://journalofvision.org/9/11/10, doi:10.1167/9.11.10. [PubMed] [Article]

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, *27*(1), 3–23.

Bhapkar, V. P. (1979). On tests of marginal symmetry and quasi-symmetry in two and three-dimensional contingency tables. *Biometrics, 35*, 417–426.

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

Brenner, E., Granzier, J. J., & Smeets, J. B. (2011). Color naming reveals our ability to distinguish between a colored background and colored light. *Journal of Vision, 11*(7):8, 1–16, http://journalofvision.org/content/11/7/8/, doi:10.1167/11.7.8. [PubMed] [Article]

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

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.

Doerschner, K., Maloney, L., & Boyaci, H. (2010). Perceived glossiness in high dynamic range scenes. *Journal of Vision, 10*(9):11, 1–11, http://journalofvision.org/content/10/9/11, doi:10.1167/10.9.11. [PubMed] [Article]

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science, 1*(1), 54–75.

Faisman, A., & Langer, M. (2013a). Environment maps and the perception of shape from mirror reflections. In *Perception, ECVP 2013 abstracts* (p. 121). London, UK: Pion Ltd.

Faisman, A., & Langer, M. S. (2013b). Qualitative shape from shading, highlights, and mirror reflections. *Journal of Vision, 13*(5):10, 1–16, http://www.journalofvision.org/content/13/5/10, doi:10.1167/13.5.10. [PubMed] [Article]

Feng, G. C. (2013). Factors affecting intercoder reliability: A Monte Carlo experiment. *Quality & Quantity*, *47*(5), 2959–2982.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378–382.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33,* 613–619.

Fleming, R., Dror, R., & Adelson, E. (2003). Real-world illumination and the perception of surface reflectance properties. *Journal of Vision, 3*(5):3, 347–368, http://journalofvision.org/content/3/5/3, doi:10.1167/3.5.3. [PubMed] [Article]

Ho, Y., Landy, M., & Maloney, L. (2008). Conjoint measurement of gloss and surface texture. *Psychological Science*, *19*(2), 196–204.

Hunter, R. S. (1975). *The measurement of appearance*. New York: Wiley.

Jakobsson, U., & Westergren, A. (2005). Statistical methods for assessing agreement for ordinal data. *Scandinavian Journal of Caring Sciences, 19*(4), 427–431.

Kingdom, F., & Prins, N. (2010). *Psychophysics–A practical introduction*. London: Academic Press.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A.,

Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3? *Perception 36*(14), 1.

Lederman, S., & Klatzky, R. (2009). Haptic perception: A tutorial. *Attention, Perception, & Psychophysics*, *71*(7), 1439–1459.

Maloney, L., & Brainard, D. (2010). Color and material perception: Achievements and challenges. *Journal of Vision, 10*(9):19, 1–6, http://journalofvision.org/content/10/9/19, doi:10.1167/10.9.19. [PubMed] [Article]

Marlow, P., Kim, J., & Anderson, B. L. (2011). The role of brightness and orientation congruence in the perception of surface gloss. *Journal of Vision, 11*(9):16, 1–12, http://journalofvision.org/content/11/9/16, doi:10.1167/11.9.16. [PubMed] [Article]

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

McCamy, C. S., Marcus, H., & Davidson, J. G. (1976). A color-rendition chart. *Journal of Applied Photographic Engineering, 2*(3), 95–99.

McKenzie, D. P., Mackinnon, A. J., Péladeau, N., Onghena, P., Bruce, P. C., Clarke, D. M., ... & McGorry, P. D. (1996). Comparing correlated kappas by resampling: Is one level of agreement significantly different from another? *Journal of Psychiatric Research, 30*(6), 483–492.

Methven, T. S., & Chantler, M. J. (2012). Problems of perceiving gloss on complex surfaces. In *Proceedings of the 3rd international conference on appearance* (pp. 43–47). Edinburgh, UK: Lulu Press.

Muller, C., Brenner, E., & Smeets, J. (2007). Living up to optimal expectations. *Journal of Vision, 7*(3):2, 1–10, http://journalofvision.org/content/7/3/2/, doi:10.1167/7.3.2. [PubMed] [Article]

Nishida, S., & Shinya, M. (1998). Use of image-based information in judgments of surface-reflectance properties. *JOSA A*, *15*(12), 2951–2965.

Obein, G., Knoblauch, K., & Viéot, F. (2004). Difference scaling of gloss: Nonlinearity, binocularity, and constancy. *Journal of Vision, 4*(9):4, 711–720, http://ww.journalofvision.org/content/4/9/4, doi:10.1167/4.9.4. [PubMed] [Article]

Phong, B. (1975). Illumination for computer generated pictures. *Communications of the ACM*, *18*(6), 311–317.

Randolph, J. J. (2005). Free-marginal multirater kappa: An alternative to Fleiss' fixed-marginal multirater kappa. In *Joensuu learning and instruction symposium*. Washington, DC: ERIC.

Sakano, Y., & Ando, H. (2010). Effects of head motion and stereo viewing on perceived glossiness. *Journal of Vision, 10*(9):15, 1–14, http://journalofvision.org/content/10/9/15/, doi:10.1167/10.9.15. [PubMed] [Article]

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy, 85*(3), 257–268.

Sprow, I., Kuepper, D., Barańczuk, Z., & Zolliker, P. (2013). Image quality assessment using a high dynamic range display. In *Proceedings of AIC 2013* (pp. 307–310). Newcastle upon Tyne, UK: AIC.

van Beers, R. J., van Mierlo, C. M., Smeets, J. B., & Brenner, E. (2011). Reweighting visual cues by touch. *Journal of Vision, 11*(10):20, 1–16, http://journalofvision.org/content/11/10/20/, doi:10.1167/11.10.20. [PubMed] [Article]

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

Wendt, G., Faul, F., Ekroll, V., & Mausfeld, R. (2010). Disparity, motion, and color information improve gloss constancy performance. *Journal of Vision, 10*(9):7, 1–17, http://journalofvision.org/content/10/9/7, doi:10.1167/10.9.7. [PubMed] [Article]