https://doi.org/10.2352/CIC.2022.30.1.45

Perception of the appearance of metal-like package printing

Carl Fridolin Weber[1], Labeed Ahmad Solangi[1], Hans Martin Sauer[1], Edgar Dörsam[1], Martin Schmitt-Lewen[2]
[1]Technical University Darmstadt, Darmstadt, Germany [2]Heidelberger Druckmaschinen AG, Walldorf-Wiesloch, Germany

Abstract

A psychophysical study on two series of printed metallized surfaces, which both consisted of ten samples was performed. Two groups participated in the experiment. These were experts, who regularly judge the appearance of printed samples at their daily work, and amateurs who do not regularly visually judge and compare samples. For the experiments, a special light booth for conducting visual experiments with focus on gloss was designed and a ranking experiment was worked out. It was investigated how observers look at these kind of surfaces when asked to judge their glossiness, lightness, roughness, the sharpness of reflected images, and metallicity. All samples had the same size and nearly no hue but differed in gloss and texture. It was examined how the ratings of the targeted characteristics of appearance correlate with each other, and how they correlate with the gloss measured at the specular angles of 20°, 60° and 85° and the distinctness-of-image measured with an IQ-S gloss meter. Additionally, observers were inquired for their individual understanding of gloss.

Introduction and motivation

Packaging of high quality products often comprises metallized and therefore high-reflective graphical elements. One of the methods commonly used for metallization is the application of printing inks that contain aluminum pigments [1]. Depending on the printing process, ink system, pigment type, and substrate used, the appearance can widely differ.

One important indicator that decides about the visual quality of such kind of print product is gloss. In industry, the gloss of print products is often compared and rated by visual assessment. This requires personnel that is experienced in visual assessment of print products. It often takes place under any lighting conditions. Gloss meters are often used as an aid to quantize gloss differences of print products. However, it is known that in many cases the readings of gloss meters do not correspond with human perception [2].

It turns out that gloss is a complex phenomenon that cannot be described using a single statement or number. Hunter [3], who is often cited in this context, divides gloss into six different types: specular gloss, sheen, contrast gloss or luster, absence of bloom gloss, distinctness of image gloss and surface uniformity gloss.

In the context of measuring gloss, there are various gloss metrics and measuring devices. In the print industry, mostly gloss meters that measure the specular gloss at 60° are used [4]. Sometimes, also gloss meters that additionally measure at 20° and 85° come into use. However, there are also gloss meters on the market that allow the measurement of more aspects of gloss as for example the distinctness-of-image (DOI) that correlates to the sharpness of a reflected image or the haze that gives indication about the scattering behavior of light on a specific surface.

Some scientific publications can be found, in which the appearance of metallic surfaces was investigated. These can be divided into studies that use synthetic images of metallic objects and studies that use real world samples. Todd and Norman [5] pursue the question of how the presence or absence of ambient light influences the appearance of computer rendered metal objects of which the metallic appearance and shininess was rated by observers. Harvey and Smithson [6] present the result of a psychophysics study that investigates the influence of low level visual features on the perception and judgement of metallicity using computer graphics. However, the appearance and perception of synthetic images differs from real world samples [7].

Dekker *et al.* [8] used a set of metallic samples, pearlescent samples, and samples containing both metallic and pearlescent materials as representative for car finish coatings to investigate how texture and color combines when assessing the appearance of such kind of special effect coatings. For this study, instrumental and visual data was analyzed and compared. With the aim to improve the consistent control of the appearance of automotive finishes, Mirjalili *et al.* [9] used samples of automotive coatings to find correlations between visually perceived and instrumentally measured appearance. Good correlations between instrumentally measured specular gloss, DOI, waviness and visual judgement were found. Rich *et al.* [10] conducted a study on the agreement between visual assessment and the measurement of the reflectivity of printed metallic inks. They show that for their samples a reliable scale for visual metallic brilliance can be developed from readings of the specular reflectance factor. Gemeinhardt [11] reports on studies on the metallicity of metallic printed samples, which is compared to their lightness. It was concluded that there is a link between measured lightness and perceived metallicity.

In its approach, this paper orients on procedures described in Engeldrum's book [12] and studies conducted by Rich *et al.* [10], and Gigilashvili *et al.* [7, 13]. The aim of the study presented in this paper was to investigate how different groups of observers, who are experts and amateurs in the field of visual judgement, judge printed metallized samples when asked to judge their glossiness, lightness, perceived roughness, the sharpness of reflected images and metallicity. These judgements were compared to the measured gloss values. Further, it was recorded what kind of objects people associate when thinking about gloss and their individual understanding of gloss itself.

Experimental Setup and Methodology

Sample Preparation and Instrumental Assessment

Out of a collection of printed metallic samples produced by flexo and offset printing on a variety of paper and board substrates using UV and water-based inks containing different types of aluminum pigments, two different sets, each consisting of ten samples were extracted. All of them have a negligible small hue component. In the selection for Series A, care was taken to include both similar and diverse samples. It includes a wide range of perceptible differences in gloss, metallic property, image sharpness, texture, surface roughness, and lightness. Thus, this set of samples represents a selection of samples, which asks observers to question their definition of gloss or metallicity. For Series B, samples were selected that were produced using the same printing process, substrate, and ink but differing slightly in print parameters. Therefore, it has a certain consistency and homogeneity. The perceptible differences in gloss, metallicity, surface roughness, and lightness are much smaller compared to Series A.

For the experiments, all samples were inserted into a black cardboard stencil with an outer size of 12 x 15 cm² and an inner window of 6 x 6 cm². This was done to cover the white edges of the specimens and to ensure that the full attention of the observers was focused on the samples. The black stencil also enabled observers to easily pick up the samples without touching the metallized surface and prevented bending of the samples.

For each sample, a number was attributed that can be clearly assigned to the measured values but does not give any information about the sample by itself. The samples were labeled with their specific number on their backside. Figure 1 shows a selection of samples used for the experiments.

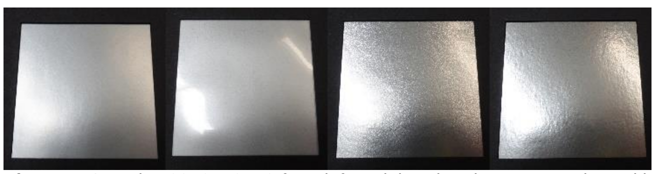

Figure 1. *Samples 20, 25, 2, 10 from left to right. The pictures are taken with the same illumination condition.*

The instrumental measurement results obtained using the IQ-S can be seen in Figure 2 and Figure 3. Samples 9 and 10 were used in both series. The measurement values represent the average of ten measurements. Five in printing direction and five across printing direction.

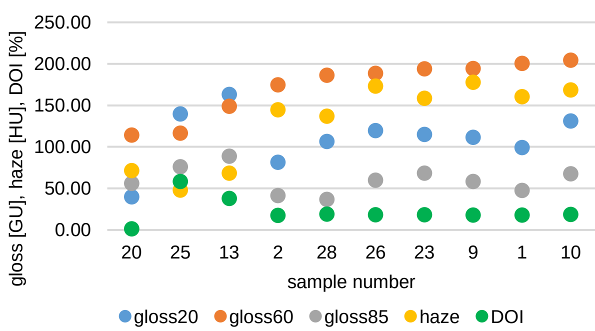

Figure 2. *Instrumental measurement results obtained from Series A. Samples are sorted according to the specular gloss measured at 60° (gloss60).*

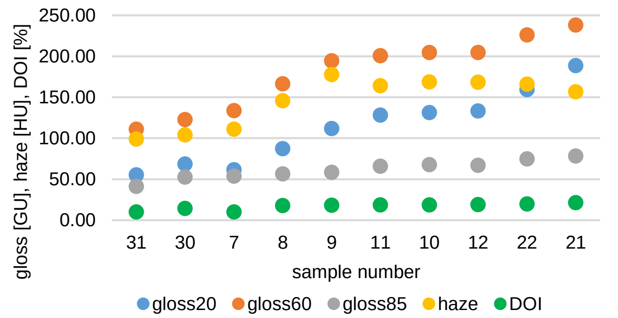

Figure 3. *Instrumental measurement results obtained from Series B. Samples are sorted according to the specular gloss measured at 60° (gloss60).*

Experimental Conditions

For the experiments a light booth was designed. As light source a D65 fluorescent tube of 58 W and a length of 1500 mm from Just Normlicht was used. It was surrounded by a frame out of aluminum profiles with a width of 1905 mm, height of 1090 mm and depth of 600 mm, a black cloth and a white screen out of paper web. Due to the direct light from the fluorescent tube and the diffuse light reflected from the paper web, constant conditions for the sampling of metallic high gloss samples were created. The importance of both direct and diffuse light for gloss perception is emphasized by Sève [14], Kirchner *et al.* [15] and Ged *et al.* [16].

The light booth was designed in a way that the observer, when sitting, only could see the area where the samples were placed but not beyond. Dazzle by the light source was avoided by the white screen. During the experiments, the examiner sat opposite to the observer and had view on the area where the samples were placed. Additionally, with the use of a camera on the observer side, the position of the samples could be recorded. Observers were asked to flip the samples after each ranking task to take a picture of the sample numbers. The setup is shown in Figure 4. The experiments were conducted in dark rooms with closed blinds.

Three objects were given to the observers prior to the experiments as an aid to assess the appearance characteristics. These were two checkerboard patterns – one with a smaller and one with a larger patter and one pen with a barcode attached shown in Figure 5. One object could have been enough but it was chosen to give observers a freedom of choice between the used objects to learn what object is favored most. All experiments were carried out in the German language.

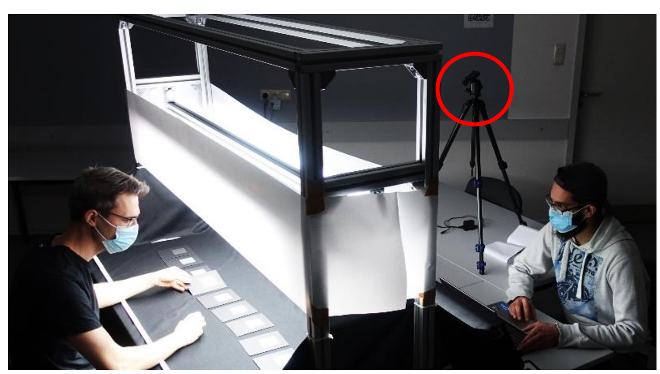

Figure 4. *Light booth and position of observer (left) and examiner (right). The position of the camera is encircled.*

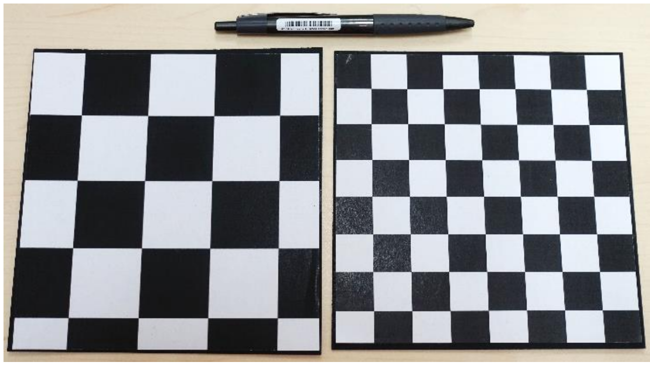

Figure 5. *Checkerboard patterns and ballpoint pen with barcode used by the observers during the experiments.*

Observers

In total, 27 observers participated in the experiment of whom 13 were male and 14 female. All of them speak German as their mother tongue. Ten observers were in the age of 20 to 30 years, four observers in the age of 30 to 40 years, three in the range of 40 to 50 years and ten in the range of 50 to 60 years. All observers had normal or corrected-to-normal vision. 12 of the observers very regularly compare and judge printed samples including metallized samples as part of their everyday business in printing press companies. Statements about how often they judge and compare printed samples ranged from at least five times a week to about twice a month. In the following, this group of observers is referred to as 'experts'. 15 of the observers do not regularly judge and compare printed samples and are referred to as 'amateurs' in the following.

Procedure and Tasks

The experiment was designed to have a maximum duration of 1.5 hours. To avoid fatigue, the observers were allowed to take a five-minute break if necessary. The observers were allowed to arrange the samples to their needs, adjust their viewing angle or stand up and move along the samples to get another perspective. They could pick up the checkerboard patterns and the pencil as help whenever needed. Observers were also encouraged to explain their actions and decisions throughout the experiment. These explanations as well as characteristics of the behavior of the observers were noted down by the examiner.

The samples were given to the observers in a randomized order. Initially, the observers were asked to perform several tasks using the ten samples of Series A. First, they were asked to place the samples in groups of their choice to make themselves familiar with them. Second, they were asked to sort the samples according to their glossiness. Third, the samples were sorted according to their perceived roughness. Fourth, the criterion for the order was the sharpness of reflected images. Fifth, the samples were ordered according to their metallicity. Sixth, samples were ordered according to their lightness. Last, Series A was ordered according to the glossiness for a second time. Hereafter, the samples of Series B had to be ordered according to their glossiness. It was allowed to place several samples on the same rank if they were found to not differ in a queried characteristic, thus preventing arbitrary ordering.

In the end, observers answered additional questions. These were related to the differences noticed between Series A and Series B, what gloss means to the observers and how one would describe gloss. Amateurs were asked, what three objects from their everyday life they perceive as having a high gloss. Experts were asked to state how often they judge printed samples and what kind of additional background knowledge they have about gloss.

Results and Analysis

Answers on questions not related to ranking

The observers described differences between the first and the second sample set congruently. Series B was described as a selection of samples, which have a very similar appearance. Based on first impressions, the perceived gloss was described as nearly identical for all samples in Series B in contrast to Series A. Their surface texture was described as rather coarse or grainy. Observers described individual samples within the set as 'rather matte' or 'rather reflective'. At first glance, compared to Series A, everyone considered it a more difficult task to sort Series B according to gloss. After ranking samples of series B, most of the observers did not find it as difficult as they had initially thought.

When asked for a description of what 'gloss' means, various buzzwords were voiced. These included, for instance: reflection, glare, not mirroring, glowing, lightness, mirror imaging, silvery, twinkling, shiny or pleasant glare. Note that these terms are translated from German. When having a closer look at the terms stated, it can be noticed that some of the descriptions are contrary. For example, some would describe a mirror as having a high or perfect gloss, while others would not describe a mirror as having a high gloss but as having a mirroring capability.

A similar observation can be made when having a look at the mentioned objects from their everyday life perceived as having a high gloss. The stated terms include: watches, automotive paint, silver jewelry, Christmas tree balls, oiled wood, polished metal, piano lacquer, window glass, water tap, mirror, new copper roofing, printed magazines, diamonds, sequins, turned off phone displays, gold jewelry, or tile floors. From the objects mentioned it can be deducted that some of the observers associate gloss with mirroring objects, while others think of objects with a high luster or sparkle.

Observation of Observers in Ranking Experiments

For some of the rankings tasks, many observers took up the samples individually in order to compare them in pairs. For other tasks, all samples were lined up on the table and observers looked at them by slowly walking past the samples and only occasionally swapped the positions of some of them.

During the experiment, most observers used one of the checkerboard patterns at least once. The ballpoint pen, on the other hand, was consulted by only a few. These aids were always used for the assessment of the sharpness of the reflected image. Only few observers, of them mostly experts used the checkerboard pattern as a help to assess gloss. Interestingly, most observers used the checkerboard with the smaller squares.

It could be determined that there was no uniform pattern among the observers with regard to the viewing position. In the course of the experiment, some observers stood up and others sat down. Most of them took up the samples moved them and held them up to the light source at different angles. To assess the criterion of gloss, some observers lined the samples up, then stood up, and bobbed their head and upper body forward and backward to perceive the course of the reflection under different angles.

The criterion of lightness was striking when analyzing the viewing position. There was uncertainty among the observers as to what the lightness actually means. Most observers decided to judge the lightness according to the diffuse reflection of the samples. They chose a very flat or a very steep viewing angle.

The criteria for which most observers made a deliberate attempt to not look into the reflection, were the lightness, the perceived roughness and partly the metallicity. Gloss was mostly assessed under the observation of the direct reflection of the light source.

Evaluation of the Rankings

The first evaluation of the rankings was done using the method *Analysis-Observer-by-Observer* described by Engeldrum [12]. Table 1 shows an excerpt of the evaluation of the first ranking of gloss for seven of the ten samples of Series A and of five observers. The rankings describe the attributes in decreasing form. Thus, the samples on rank 1 are strongest in the considered attribute, The rankings for observer 4 are fractional numbers because sample 9 and sample 10 were ranked equally. Hence, nine positions had to be normalized from one to ten. The ordinal scale was converted into an interval scale by calculating the proportions *p* for every sample and the z scores *z(p)* as described in [17].

Table 1: Excerpt of the evaluation of the first ranking of gloss of samples in Series A.

Observer number	Sample number						
	1	2	9	10	13	20	23
4	7.43	8.71	4.86	4.86	2.29	10	6.14
5	3	5	2	1	7	10	6
6	4	1	2	3	5	10	7
7	3	2	4	5	7	10	6
8	4	5	8	3	1	7	10

In Table 2, the averaged standard deviations of all samples for the respective attributes are shown. These were calculated by taking the standard deviations of the ranks determined by the observers for each sample and calculating the arithmetic mean of them. For the lightness, only the results of eight amateurs were considered, who judged lightness by looking at the diffuse reflection. For the experts, all ratings were considered because all of them judged lightness by looking at the diffuse reflection

Table 2: Averaged standard deviations of rankings of amateurs (15) and experts (12). 'Gloss 1' denotes the first lap observers were asked to rank samples according to gloss. 'Gloss 3' denotes the ranking of Series B.

	Series A						Series B
	Gloss 1	Perceived roughness	Image sharpness	Metallicity	Lightness	Gloss 2	Gloss 3
Averaged standard deviation amateurs	2.17	1.19	1.36	2.71	0.91	2.25	1.63
Averaged standard deviation experts	2.23	1.35	1.44	2.43	0.40	2.16	1.33

It can be seen that the deviation of the ranking of lightness, perceived roughness, and image sharpness are the lowest. One reason for this could be that the observers have a similar idea of what these attributes mean. The observation that the perceived metallicity has the highest standard deviation reveals that this is the most abstract attribute, which leads to the high disagreement among the observers. Furthermore, the standard deviations reveal that the disagreement among the observers was higher for the gloss of samples of Series A compared to the gloss of samples of Series B, although observers initially stated that Series B was more difficult to rank due to the small differences between the samples. The differences between the two rankings of Series A according to gloss are small.

For the non-professional observers the Pearson-r coefficient between the rankings of Gloss 1 and Gloss 2 is r = 0.96, and for the professional observers the correlation is r = 0.95 what shows that both groups have a similar consistency when ranking the same series two times.

However, when comparing the rankings of gloss between the non-professional and professional observers, it turns out that these two groups judged samples of Series A differently, while the rankings of Series B were judged similarly as shown in Figure 6.

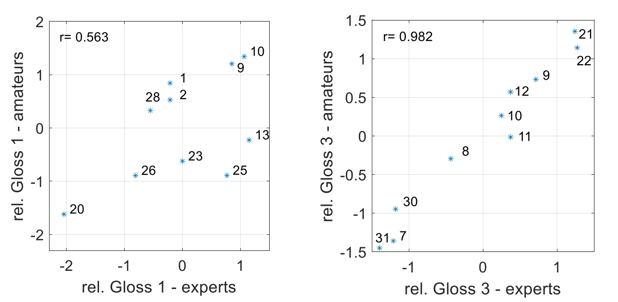

Figure 6. *Comparison of the rankings of the professional and non-professional observer group for Gloss 1 of Series A and Gloss 3 of Series B. The numbers denote the sample numbers.*

Comparison with Instrumental Assessment

The rankings of Series A and Series B according to gloss were compared with the specular gloss measured at 20°, 60°. Results of the rankings of amateurs and experts of Series A according to gloss and comparison with the measured gloss values at the specular angles of 20° and 60° can be seen in Figure 7. To give an impression about the goodness of the accordance between measured and visual data, the Kendall-τ-coefficient was calculated, which evaluates how well the samples were ranked in the same order as they are ranked by the measured gloss values. The reason for this choice was that the least what should be expected of gloss meters is to give measurement values that correlate with the visual rank order. For the amateurs the correlation between specular gloss measured at 20° and the visual ranking is $\tau = 0.02$. The correlation with specular gloss measured at 60° is $\tau = 0.60$. For the experts, the correlation between specular gloss measured at 20° and the visual ranking is $\tau = 0.54$. The correlation with specular gloss measured at 60° is $\tau = 0.22$. It is striking that the visual ranking of both groups correlate with different angles of measurement. It is also interesting to see that none of the two groups ranked Series A in a way that correlates strongly with any of the aspects of gloss measured.

Results of the rankings of amateurs and experts of Series B according to gloss and comparison with the measured gloss values at the specular angles of 20° and 60° can be seen in Figure 8. For the amateurs the correlation between specular gloss measured at 20° and the visual ranking is $\tau = 0.87$. The correlation with specular gloss measured at 60° is $\tau = 0.82$. For the experts, the correlation between specular gloss measured at 20° and the visual ranking is $\tau = 0.76$. The correlation with specular gloss measured at 60° is $\tau = 0.72$.

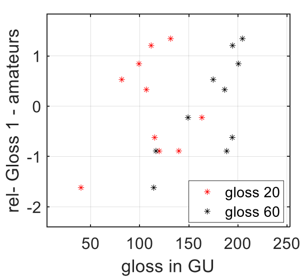

Figure 7. *Comparison between visual rankings of Series A of amateurs (left) and experts (right) according to perceived glossiness (Gloss 1) with specular gloss measured at 20° and 60°.*

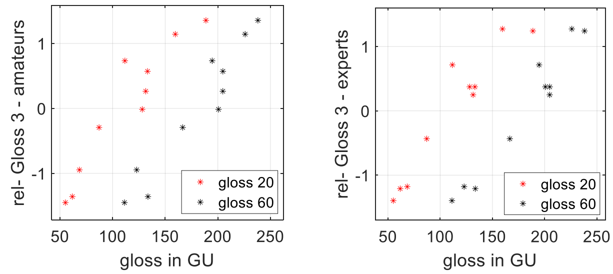

Figure 8. *Comparison between visual rankings of Series B of amateurs (left) and experts (right) according to perceived glossiness (Gloss 3) with specular gloss measured at 20° and 60°.*

For the reason that the receiver aperture of the gloss meter at 20° is smaller than the receiver aperture at 60° as described in ASTM D523-14 [18], gloss measurement values obtained from the 20° geometry naturally correlate stronger with the image sharpness and hence also with the gloss measurement values obtained at 60°. This is reflected for both groups of observers by comparison of the gloss measurement values with the rankings of the perceived sharpness of image reflection. The correlations between both are shown in Table 3. It is striking that the visual ranking of image sharpness correlates slightly better with the specular gloss measured at 20° than with the DOI.

Table 3: Kendall-τ-coefficients calculated for rankings according to perceived image sharpness with measurement values for Series A.

	Gloss20	Gloss60	DOI
Amateurs (15)	0.644	-0.11	0.51
Experts (12)	0.73	-0.11	0.42

Discussion and Conclusion

The main goal of the present research was to examine how different groups of observers judge the appearance of metallized printed samples when asked to rank these according to several attributes such as gloss, perceived roughness, metallicity, perceived sharpness of reflected images, and lightness. These rankings were compared with the measurement values of a gloss meter. For this purpose, a ranking experiment was conducted, two different series of samples were collected, a light booth was designed, and ranking experiments were conducted.

For professional observers, there are strong correlations between perceived gloss and perceived image sharpness of the reflection. This does not apply to the group of non-professional observers. For Series A, the rankings of professional and non-professional observers differ greatly while rankings of Series B strongly correlate. Further, for Series A, specular gloss measurements obtained at 60° correlate more with the rankings of amateurs according to gloss while specular gloss measurements obtained at 20° correlate more with rankings of experts. This means that the 20° angle for high-gloss specimens is not necessarily more suitable than the 60° angle as often described in literature. If it is to hold that there are greater differences in the judgement of gloss between experts and amateurs this would have significant consequences for printing practice were professional observers design metallic embellishments on packaging that could be perceived different by consumers, i.e. amateur observers. However, a greater number of observers and more series of printed metallized samples should validate the experimental results.

Further experiments with different kinds of sample sets will be conducted. It is planned to conduct experiments on printed metallic samples that yield a much higher gloss compared to the samples used in this study. It could also be possible to select samples that only differ in one measured aspect of gloss e.g. DOI, haze, sparkle or specular gloss to further investigate the accordance between measured and perceived characteristics of gloss.

Acknowledgement

The authors express their gratitude for the financial support as well as for the provision of metallized printing samples by the Heidelberger Druckmaschinen AG. We also thank all observers who participated in the experiment.

References

[1] C. F. Weber, D. Spiehl, and E. Dörsam, “Printing methods used in the label and package printing industry for the production metallic embellishments with a focus on metal effect pigments,” *J. Print Media Technol. Res.*, vol. 11, no. 1, pp. 29–45, 2022, doi: 10.14622/JPMTR-2118.

[2] F. B. Leloup, G. Obein, M. R. Pointer, and P. Hanselaer, “Toward the soft metrology of surface gloss: A review,” *Color Res. Appl.*, vol. 39, no. 6, pp. 559–570, 2014, doi: 10.1002/col.21846.

[3] R. S. Hunter and W. H. Harold, *The Measurement of Appearance*, 2nd ed. New York: John Wiley & Sons, Inc., 1987.

[4] C. F. Weber, D. Spiehl, and E. Dörsam, “Comparing gloss meters for gloss measurements on metallic embellishments from the printing industry,” in *IARIGAI Athens 2021 conf. proceedings*, Athens, 2021, pp. 189–199, doi: 10.14622/Advances_47_2021.

[5] J. T. Todd and J. F. Norman, “The visual perception of metal,” *J. of vision*, vol. 18, no. 3, p. 9, 2018, doi: 10.1167/18.3.9.

[6] J. S. Harvey and H. E. Smithson, “Low level visual features support robust material perception in the judgement of metallicity,” *Scientific reports*, vol. 11, no. 1, p. 16396, 2021, doi: 10.1038/s41598-021-95416-6.

[7] D. Gigilashvili, J.-B. Thomas, J. Y. Hardeberg, and M. Pedersen, “Behavioral Investigation of Visual Appearance Assessment,” *color imaging conf*, vol. 2018, no. 1, pp. 294–299, 2018, doi: 10.2352/ISSN.2169-2629.2018.26.294.

[8] N. Dekker, E. J. J. Kirchner, R. Supèr, G. J. van den Kieboom, and R. Gottenbos, “Total appearance differences for metallic and pearlescent materials: Contributions from color and texture,” *Color Res. Appl.*, vol. 36, no. 1, pp. 4–14, 2011, doi: 10.1002/col.20586.

[9] F. Mirjalili, S. Moradian, and F. Ameri, “A new approach to investigate relationships between certain instrumentally measured appearance parameters and their visually perceived equivalents in the automotive industry,” *J. Coat Technol Res*, vol. 11, no. 3, pp. 341–350, 2014, doi: 10.1007/s11998-013-9544-9.

[10] D. C. Rich, R. Marcus, V. Lovell, and T. Kreutz, “Modeling the appearance of metal-like packaging printing,” *Color Res. Appl.*, vol. 42, no. 1, pp. 38–49, 2017, doi: 10.1002/col.22035.

[11] J. Gemeinhardt, „Kommunikation von messtechnischen Kenngrößen zur Beschreibung des metallischen Aussehens von Drucken,” Fogra res. report no. 34.021, 2021.

[12] P. G. Engeldrum, *Psychometric scaling: A toolkit for imaging systems development*, 1st ed. Winchester, Mass: Imcotek Press, 2000.

[13] D. Gigilashvili, J.-B. Thomas, M. Pedersen, and J. Y. Hardeberg, “Perceived Glossiness: Beyond Surface Properties,” *color imaging conf*, vol. 2019, no. 1, pp. 37–42, 2019, doi: 10.2352/issn.2169-2629.2019.27.8.

[14] R. Sève, “Problems connected with the concept of gloss,” *Color Res. Appl.*, vol. 18, no. 4, pp. 241-252, 1993, doi: 10.1002/col.5080180407

[15] E. Kirchner, G. van den Kieboom, Njo, R. Supèr, and R. Gottenbos, “Observation of visual texture of metallic and pearlescent materials,” *Color Res. Appl.*, vol. 32, no. 4, pp. 256–266, 2007, doi: 10.1002/col.20328.

[16] G. Ged, A.-M. Rabal-Almazor, M. E. Himbert, and G. Obein, “Assessing gloss under diffuse and specular lighting,” *Color Res. Appl.*, vol. 45, no. 4, pp. 591–602, 2020, doi: 10.1002/col.22510.

[17] J. A. Ferwerda, Psychometrics 101: How to Design, Conduct, and Analyze Perceptual Experiments in Computer Graphics. Cornell University, nd.

[18] American Society for Testing and Materials (ASTM), *Test Method for Specular Gloss*, ASTM D523 − 14, West Conshohocken, PA, 2018.