

Article

Skin Sensory Assessors Highly Agree on the Appraisal of Skin Smoothness and Elasticity but Fairly on Softness and Moisturization

Naoki Saito ^{1,*} , Kohei Matsumori ¹ , Taiki Kazama ¹, Naomi Arakawa ¹  and Shogo Okamoto ²

¹ MIRAI Technology Institute, Shiseido Co., Ltd., Yokohama 220-0011, Japan

² Department of Computer Sciences, Tokyo Metropolitan University, Hino 191-0065, Japan

* Correspondence: naoki.saito@shiseido.com

Abstract: We tested the reliability of sensory evaluations of tactile sensation on bare skin and investigated the reliability among evaluation attributes by trained and untrained assessors. Two trained professional panelists and two untrained researchers evaluated skin in terms of several attributes: *smooth-rough*, *elastic-not elastic*, *soft-hard (surface)*, *soft-hard (base)*, *moisturized-dry*. Twenty-two women aged 25–57 years were evaluated, and the sensory evaluation was repeated twice. Correlation coefficients and intraclass correlation coefficients (ICCs) were used to examine intra- and inter-assessor reliability. The sensory evaluation and physical quantities acquired by commercial and non-commercial instruments were moderately correlated. *Smooth-rough* and *elastic-not elastic* showed high or moderate inter-assessor reliabilities with mean correlation coefficients between panelists of 0.81 and 0.58, respectively. Further, the ICC (2,1) values were 0.64 and 0.51, respectively, and the ICC (2,2) values were 0.77 and 0.67, respectively. Conversely, the reliabilities of *soft-hard (surface)*, *soft-hard (base)*, and *moisturized-dry* were low; the mean correlation coefficients between the panelists were 0.36, 0.23, and 0.22; the ICC (2,1) values were 0.27, 0.23, and 0.17; and the ICC (2,2) values were 0.42, 0.29, and 0.26, respectively. Reliability differed between attributes. We found no meaningful differences between the trained and untrained panelists regarding intra- or inter-assessor reliability.

Keywords: sensory evaluation; reliability; tactile sensation



Citation: Saito, N.; Matsumori, K.; Kazama, T.; Arakawa, N.; Okamoto, S. Skin Sensory Assessors Highly Agree on the Appraisal of Skin Smoothness and Elasticity but Fairly on Softness and Moisturization. *Cosmetics* **2022**, *9*, 86. <https://doi.org/10.3390/cosmetics9040086>

Academic Editor: Enzo Berardesca

Received: 6 August 2022

Accepted: 11 August 2022

Published: 18 August 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Sensory evaluation of bare skin is indispensable for developing skin care products with high user satisfaction. It evaluates the appearance and sense of touch on the skin and the changes in skin due to skin care products, and the results can be used for product claims.

Sensory evaluation can be divided into two categories: evaluation by users and evaluation by professional assessors [1]. Evaluation by users is conducted to investigate consumers' preferences under conditions similar to actual usage, whereas evaluations by professional assessors are based on specific criteria. The purpose of the evaluation by professional assessors is to determine the characteristics of each attribute analytically, and reliability is a primary concern. Therefore, this study focused on the reliability of sensory evaluations conducted by professional assessors.

Various guidelines and studies have been published, aiming to improve the reliability of sensory evaluations [1–9]. For example, the International Organization for Standardization (ISO) 11036 is a guideline for texture evaluation, covering texture profile classification, attribute classification and development, reference samples, evaluation methods, scales, panel screening and training, and data analysis [2]. The American Society for Testing and Materials has defined sensory evaluation methods for skin creams, lotions, and shampoos [6,7]. The guidelines on efficacy evaluation issued by the European Cosmetic and Perfumery Association describe the classification of efficacy claims, the disclosure of

information when conducting tests, and the writing of reports [1]. Additionally, the European Group on Efficacy Measurement and Evaluation of Cosmetics and Other Products (EEMCO), a working group of experts, has published guidelines for the sensory evaluation of wrinkles and dryness [7,8].

Currently, cosmetic manufacturers and research institutes that perform efficacy evaluations of cosmetics train professional assessors based on these or similar guidelines [10–16]. For example, Addor et al. [13] evaluated the effect of moisturizing products by instrumental measurements and trained panelists. However, there is limited published information on the reliability of sensory evaluations.

Aust et al. reported on the tactile evaluations of five lotions involving nine trained assessors to show the differences in their textures [17]. They confirmed that the standard deviation of the scores by the expert assessors was small, but they did not indicate this value. Additionally, because the evaluation was conducted only once, intra-assessor reproducibility was not evaluated. Further, Vieira et al. trained 43 assessors to evaluate moisturizing creams [18] and classified the application process into four categories: appearance, pick-up, rub-out, and after-feel. Then, they selected 14 important attributes for each process. The assessors evaluated the three samples for each attribute and reported that there were no significant differences between the first and second evaluations of each attribute and that the responses of the assessors were reliable; however, this study did not provide specific data. Calixo et al. conducted sensory evaluations of four gel creams by 50 Brazilian and 50 French participants and reported high reliabilities of the sensory evaluations of texture between the two cultures [19]. Larnier et al. also developed a photographic scale to evaluate skin photodamage and evaluated the reliability among assessors [20]; however, they did not investigate tactile sensations. Kang et al. assessed dry skin based on the EEMCO guidelines by both dermatologists and pharmacists and tested the inter-assessor reliability using the intraclass correlation coefficient (ICC) [21]. The results showed that the visual scale and crack fissures were in fair agreement, and the overall dry skin score, redness and roughness, and scaling, as defined in the guidelines, were in moderate or substantial agreement. The EEMCO guidelines for assessing dry skin covered several attributes for appearance, and only roughness was related to tactile sensation. Additionally, Kang et al. did not investigate the reproducibility for the assessors [21].

As mentioned, only a few studies have reported on the reliability of sensory evaluations of tactile sensations on skin. Several studies have discussed reliability among assessors but not within assessors. Although the sensory evaluation of bare skin is indispensable for the evaluation of cosmetic products, the fact that the specific degree of reliability is unclear is a major issue. We believe that it is important to understand the current situation to improve the evaluation method. Therefore, this study aimed to determine the reliability of tactile evaluations of bare skin and examine the following: (a) differences in reliability among attributes (i.e., smoothness, elasticity, softness, and moisturization) and (b) the difference in reliability between the attributes with and without training. In addition, to investigate the consistency of the sensory evaluation with the physical quantities obtained by skin measurement instruments, we also report the correlation coefficients between these quantities following earlier studies [22,23].

2. Materials and Methods

The study was approved by the Human Study Ethics Committee of the Shiseido Global Innovation Center (Study No. C02205, C02206). Written informed consent was obtained from the 4 assessors and 22 participants of the study.

2.1. Assessors and Examinee

The assessors were two trained professional assessors (Assessors A and B) and two untrained researchers (Assessors C and D). Assessors A and B had 5 and 16 years of experience, respectively, at the time of the study. They occupationally perform the tactile sensory evaluation of skin and cosmetics and have tested more than several hundred

people in the past. There were no other assessors as experienced as these two in the authors' institute. Only a few professional assessors routinely perform sensory evaluations of bare skin. In contrast, Assessors C and D had never conducted the tactile sensory evaluation of skin at their workplaces. The examinees were 22 women aged 25–57 years.

2.2. Reproducibility Test for Sensory Evaluation

The evaluation site was a 40 mm diameter area on the left cheek next to the nose, and the area was masked with a 0.2 mm thick polypropylene sheet (Figure 1a). The assessors were blindfolded using a face shield to eliminate any visual effects. Normally, expert assessors are not blindfolded when assessing bare skin; however, in this experiment, they were blindfolded to prevent the examinee from being identified (Figure 1b). Assessors practiced in advance in sensory evaluation while being blindfolded. As the assessors could not visually check the evaluation site, the examinee guided the assessor's hand to the evaluation site.

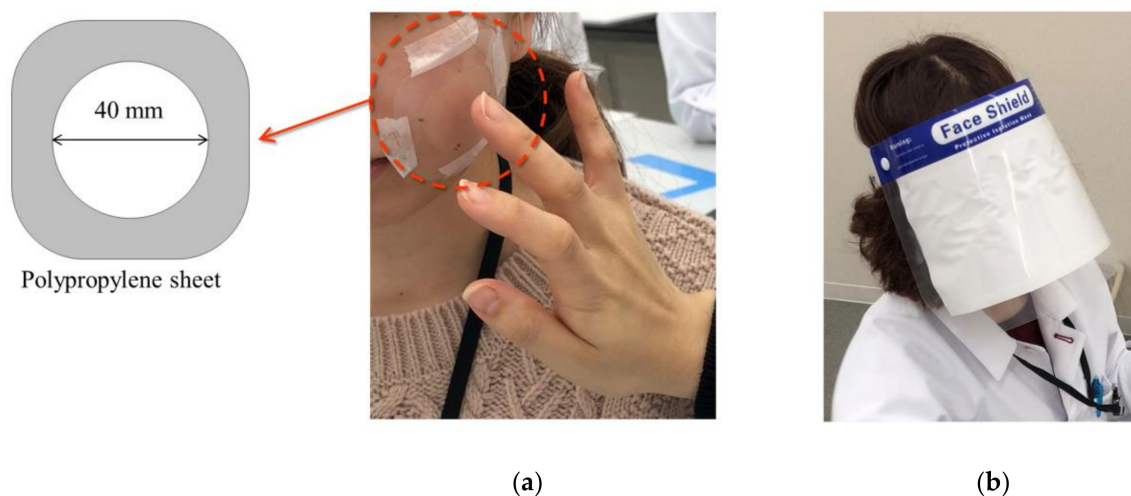


Figure 1. Images of the assessor and examinee during the experiment. (a) Examinee and evaluation area. (b) Assessor.

The experiment was performed over three days with 6–8 examinees per day. Examinees removed their makeup and washed their face before the test, and the sensory evaluation was started 5–10 min after washing their face. The same examinee was tested twice in a blind manner. For example, in one day, eight examinees joined the evaluation; however, the assessors did not know how many examinees there were or how many times each examinee was evaluated. The same examinee was never evaluated twice by the same assessor in a row. For each examinee, the second test was conducted within 30 min after the first test. This design was to prevent temporal changes in the skin condition after cleansing. The test was conducted in a thermo-hydrostatic chamber set at a room temperature of 23 °C and a relative humidity of 45%.

2.3. Sensory Evaluation of Bare Skin

We adopted the sensory evaluation method used by the authors' group. The evaluation attributes were *smooth–rough*, *elastic–not elastic*, *soft–hard (surface)*, *soft–hard (base)*, *moisturized–dry*, and *oily–not oily*. *Elastic–not elastic* referred to stiffness or restoring force of the skin, removing the effect of viscosity. *Soft–hard (surface)* referred to softness up to a depth of approximately 2 mm, and *soft–hard (base)* referred to softness up to a depth of approximately 5 mm. The definitions provided to the panel are listed in Table 1. They were presented in Japanese and English. These attributes are typically used in the sensory evaluation of skin [13,24–26], except for the two types of softness.

Table 1. Definitions of attributes.

Attribute	Definition
Smooth–rough	Overall evaluation of the unevenness and roughness felt when sliding your finger across the surface. Evaluate using reference artificial skin models.
Elastic–not elastic	The force with which the skin pushes back against the finger after pressing. Evaluate using reference artificial skin models.
Soft–hard (surface)	Resistance when pressing (going) to a depth of approximately 2 mm. Evaluate using reference artificial skin models.
Soft–hard (base)	Resistance when pressing (going) to a depth of approximately 5 mm. Evaluate using reference artificial skin models.
Moisturized–dry	A feeling of stickiness to the finger when it is pushed and slid a little. If some areas feel moisturized and others feel dry, evaluate the entire area on average.
Oily–Not oily	Evaluation of the degree to which the oil on the skin surface is felt.

Of the six evaluation attributes, four were evaluated referring to artificial skin models, and the remaining two were evaluated without such models. Two sets, each of which was composed of four models, were used as references. One set was the reference for the *smooth–rough* evaluation, which was made of urethane from a plaster mold of the cheek. Each had a cylindrical shape with a diameter of 45 mm and a height of 6 mm. The other reference set was used to evaluate *soft–hard (surface)*, *soft–hard (base)*, and *elastic–not elastic*. They were cylindrical urethane with diameters of 50 mm and heights of 15 mm. Both reference models were created under the supervision of experienced expert assessors so that the score for each attribute had an even distribution across four levels. Non-disclosure agreements with the manufacturer of these skin models preclude us from providing details on how the artificial skin models were created.

The evaluation was conducted using a scale of 0–5 points, with 11 steps of 0.5 points each. Hand movements for the evaluation were not controlled, and the assessor adopted movements that were easy to evaluate.

2.4. Instrumental Measurements of Skin

Physical conditions of skin were measured after the sensory evaluation tasks.

The state of the furrow of the skin surface was investigated using a video microscope. The amount of furrow and image features representing the non-uniformity of the furrow were obtained from the captured images [27]. The amount of furrow is the number of black dots in the binarized image in which the skin grooves are blackened. The non-uniformity of the furrow is defined as the coefficient of variation in the number of black dots in each square when the acquired image is divided into 12×12 squares, with larger values indicating non-uniformity and smaller values indicating uniformity. The average value was obtained from three measurements taken in the area where the sensory evaluation was conducted.

A Corneometer CM 825 (Courage & Khazaka, Cologne, Germany) and a Skicon-200EX (Yayoi Co., Ltd., Tokyo, Japan) were used to measure the water content of the left cheeks. Measurements were repeated five times, and their arithmetic mean was calculated.

Skin viscoelasticity was evaluated using a Cutometer CT580 (Courage & Khazaka) with a 2 mm suction diameter and a suction pressure of 400 mbar. Twelve parameters defined in [28] (i.e., $R0$ –9 and $F0$ –1) were recorded for later analyses.

2.5. Analysis

We removed some outliers before the main analyses. For each physical parameter of skin, the examinees whose ratings were outside the mean plus or minus twice the standard deviation among all the examinees were excluded as outliers. Each examinee was

tested twice by the same assessor; hence, for individual assessors, examinees for which the difference between the first and second ratings of each attribute was more than twice the standard deviation among all assessors were excluded as outliers [29]. Based on these criteria, out of 88 trials (22 examinees \times four assessors), *smooth-rough* and *moisturized-dry* were removed as outliers in five trials, and *elastic-not elastic*, *soft-hard (surface)*, and *soft-hard (base)* in seven trials.

To seek consistency and reliability between the sensory evaluation and physical parameters of skins, we calculated their Pearson's correlation coefficients. For this calculation, the mean ratings among the four assessors were used. Regarding *smooth-rough*, the correlation coefficients were calculated with two parameters acquired by a video microscope [27]. Regarding *moisturized-dry*, correlations with the Corneometer CM 825 and Skicon-200EX values were calculated. Further, the correlation coefficients between each of the *elastic-not elastic*, *soft-hard (surface)*, and *soft-hard (base)* ratings and each of the twelve parameters provided by the Cutometer CT580 were calculated.

Pearson's correlation coefficients were calculated for each assessor's first and second scores to evaluate the reproducibility within assessors. Additionally, Pearson's correlation coefficient between the two assessors was calculated by using the mean of each examinee's first and second scores to evaluate the reproducibility between the two assessors. The interpretation of the correlation coefficient is described in Table 2 [30].

Table 2. Interpretation of correlation coefficient according to Evans [30].

Coefficient Value	Interpretation
0.00–0.19	Very weak
0.20–0.39	Weak
0.40–0.59	Moderate
0.60–0.79	Strong
0.80–1.00	Very strong

Correlation coefficients can be used to check whether the two data are in a linear relationship, but they cannot evaluate whether the data agree. Therefore, we calculated ICCs. The ICCs were classified into three categories [31]. In this study, ICC (1,1) and ICC (1,2) of Case 1 were calculated as indices of intra-assessor reliability. ICC (1,2) indicates the reliability when the average of two evaluations is used.

The ICC (2,1) and ICC (2,2) for Case 2 and ICC (3,1) and ICC (3,2) for Case 3 were calculated as indices of inter-assessor reliability. Note that Case 2 required an absolute score agreement among assessors, while Case 3 was unaffected by assessment biases. The criteria used for ICCs are listed in Table 3 [32]. The average of two evaluations by each assessor was used for the calculations of Cases 2 and 3. SPSS Statistics (version 23, IBM, Armonk, NY, USA) was used to calculate ICCs. The level of significance was set at $p < 0.05$.

Table 3. Interpretation of intraclass correlation coefficients by Fleiss [32].

Coefficient Value	Interpretation
<0	Less than chance agreement
0.01–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

3. Results

Table 4 shows the results of the evaluation of the furrow of the skin surface, including the mean, standard deviation, maximum, and minimum values for the participants, as well as the correlation coefficient with the *smooth-rough* score. The furrow condition differed among the participants. The correlation coefficients between *smooth* and *rough* and the number and nonuniformity of creases were -0.40 ($p = 0.065$) and 0.46 ($p = 0.030$), respectively.

Table 4. Results of skin furrow measurements.

	Mean	SD	Max	Min	Correlation with <i>Smooth-rough</i>	
					R	p
Number of skin furrows per measured area	7903	622	9277	6784	-0.40	0.065
Non-uniformity of skin furrows	0.35	0.03	0.40	0.28	0.46^*	0.030

* indicates statistical significance at $p < 0.05$. SD, standard deviation.

Table 5 shows the mean, standard deviation, maximum, and minimum values of moisture indices for the participants, as well as the correlation coefficient with the *moisturized-dry* scores. The correlation coefficients between *moisturized-dry* scores and Corneometer CM 825 and Skicon-200EX values were -0.22 ($p = 0.323$) and -0.61 ($p = 0.003$), respectively.

Table 5. Water contents of skin.

	Mean	SD	Max	Min	Correlation with <i>Moisturized-dry</i>	
					R	p
Corneometer CM 825 (a.u.)	19.9	6.5	33.4	8.0	-0.22	0.323
Skincon-200EX (μ S)	185.8	89.4	422.7	79.0	-0.61^{**}	0.003

** indicates statistical significance at $p < 0.01$. SD, standard deviation.

Table 6 shows the results of skin viscoelasticity measurements, including the mean, standard deviation, maximum and minimum values among the examinees, as well as the correlation coefficients between *elastic-not elastic*, *soft-hard (surface)*, and *soft-hard (base)*. *Elastic-not elastic* was significantly correlated with R1 (-0.46 , $p = 0.03$), R2 (0.47 , $p = 0.026$), and R4 (-0.45 , $p = 0.037$) of the Cutometer parameters. *Soft-hard (surface)* was significantly correlated with R0 (-0.49 , $p = 0.021$), R3 (-0.45 , $p = 0.035$), R8 (-0.51 , $p = 0.016$), and F1 (-0.53 , $p = 0.012$). None of the correlations for *soft-hard (base)* were significant whereas the correlations with R8 and F1 were relatively high at -0.40 ($p = 0.066$) and -0.41 ($p = 0.055$), respectively.

The scatter plots of the first and second scores for each attribute and assessor are presented in Figure 2. The results for *oily-not oily* were excluded from the analysis because the corresponding scores were almost zero. This may be attributed to the fact that all examinees were women, and the evaluation was performed immediately after washing the face, which resulted in less sebum overall.

Table 6. Skin viscoelasticity parameters and correlation with sensory evaluation scores.

	Mean	SD	Max	Min	Correlation with <i>Elastic–not elastic</i>		Correlation with <i>Soft–hard (Surface)</i>		Correlation with <i>Soft–hard (Base)</i>	
					<i>R</i>	<i>p</i>	<i>R</i>	<i>p</i>	<i>R</i>	<i>p</i>
R0	0.414	0.049	0.535	0.314	−0.25	0.262	−0.49 *	0.021	−0.33	0.136
R1	0.110	0.038	0.206	0.062	−0.46 *	0.030	−0.11	0.636	−0.01	0.952
R2	0.737	0.074	0.820	0.546	0.47 *	0.026	−0.05	0.823	−0.11	0.612
R3	0.425	0.054	0.552	0.320	−0.24	0.278	−0.45 *	0.035	−0.30	0.172
R4	0.105	0.046	0.232	0.050	−0.45 *	0.037	−0.02	0.918	0.04	0.861
R5	0.637	0.183	0.888	0.248	0.42	0.050	−0.01	0.962	−0.11	0.637
R6	0.230	0.036	0.307	0.160	0.11	0.628	0.18	0.410	0.17	0.445
R7	0.521	0.160	0.750	0.192	0.39	0.073	−0.03	0.897	−0.12	0.602
R8	0.304	0.039	0.393	0.236	0.14	0.549	−0.51 *	0.016	−0.40	0.066
R9	0.011	0.007	0.024	0.000	−0.11	0.617	−0.05	0.829	−0.03	0.912
F0	0.045	0.014	0.068	0.022	−0.11	0.611	−0.10	0.672	−0.04	0.853
F1	0.560	0.073	0.718	0.436	0.16	0.467	−0.53 *	0.012	−0.41	0.055

* indicates statistical significance at $p < 0.05$. SD, standard deviation.

Pearson's correlation coefficients for the first and second scores of each assessor are shown in Table 7 (top). Regarding the mean values of the intra-assessor correlation coefficients among the four assessors, *smooth–rough* exhibited the highest value of 0.77. Conversely, the value of *moisturized–dry* was the lowest with a moderate correlation of 0.52. The other attributes exhibited strong correlations ranging from 0.63–0.68.

When assessors A and B, who were expert assessors, were compared with assessors C and D, who were untrained assessors, no clear differences were found in intra-assessor correlation values. For all assessors, the correlation for *smooth–rough* was very large (i.e., 0.95 for assessor A). For assessor A, it was weak for *elastic–not elastic* and moderate for the other attributes. For assessor B, only the correlation for *moisturized–dry* was moderate, whereas the correlations for the other attributes were strong or very strong. Assessor C exhibited strong correlations for all items and high intra-assessor reproducibility. Assessor D exhibited low values for *smooth–rough* and *moisturized–dry*.

The bottom of Table 7 shows the inter-assessor correlation coefficients. The mean correlation coefficient for *smooth–rough* was highest at 0.81. The second highest inter-assessor correlation was 0.58 for *elastic–not elastic*; those for the other attributes were smaller than 0.36. The correlation coefficient for the expert assessors (A and B) averaged across all attributes was 0.38, which was not largely different from those of other combinations of assessors.

Table 8 shows the ICC (1,1) and ICC (1,2) values. The mean value among the four assessors was highest for *smooth–rough* (0.77), lowest for *moisturized–dry* (0.53), and greater than 0.6 for the other attributes. For assessor A, *smooth–rough* was almost perfect, but the other attributes were below moderate values. For assessor B, only *moisturized–dry* was moderate, and all other attributes scored greater than 0.61 and were substantial. For assessor C, all the attributes were above substantial. Assessor D exhibited low scores for *smooth–rough* and *moisturized–dry*. Overall, the ICC (1,2) value increased compared to the ICC (1,1) value.

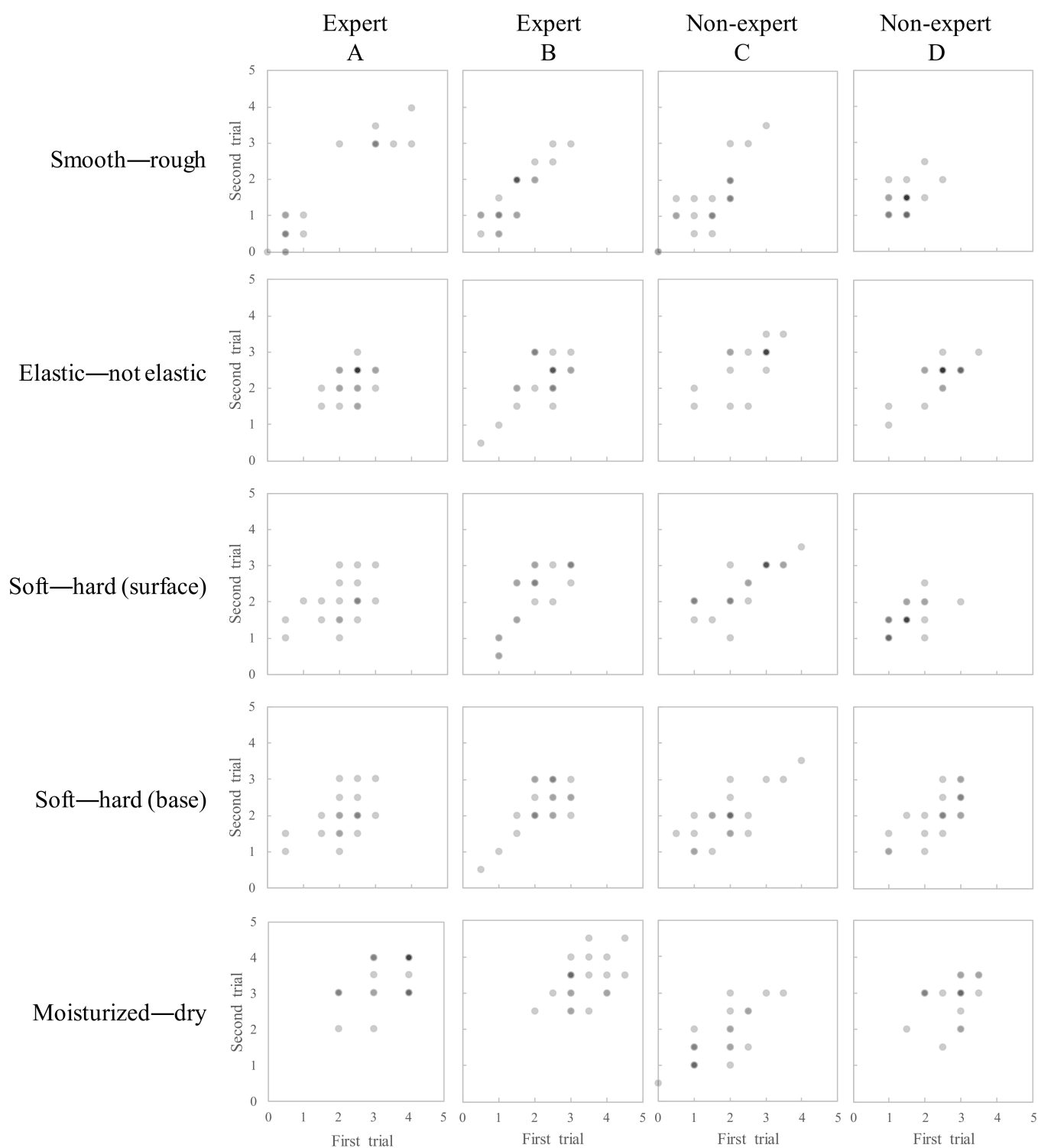


Figure 2. Scatter plots of sensory appraisal scores between the first and second trials for each combination of the panel and attribute. The columns show four assessors: A, B, C, and D. A and B are trained and C and D are non-trained. The rows show the five assessment items. The horizontal axis of each scatter plot represents the first assessment, and the vertical axis represents the second assessment. The shading of the points indicates the density; a darker shading indicates that more samples are gathered at that point.

Table 7. Intra- (top) and inter-assessor (bottom) Pearson’s correlation coefficients.

Assessor	A		B		C	D	
Expert/non-expert	Expert		Expert		Non-expert	Non-expert	Mean
Smooth—rough	0.95 **		0.87 **		0.81 **	0.45 *	0.77
Elastic—not elastic	0.37		0.67 **		0.68 **	0.78 **	0.63
Soft—hard (surface)	0.50 *		0.79 **		0.79 **	0.58 **	0.67
Soft—hard (base)	0.53 *		0.72 **		0.75 **	0.73 **	0.68
Moisturized—dry	0.47 *		0.46 *		0.77 **	0.36	0.52
Mean	0.56		0.70		0.76	0.58	
Comparison	A/B	A/C	A/D	B/C	B/D	C/D	Mean
Smooth—rough	0.83 **	0.84 **	0.81 **	0.81 **	0.72 **	0.84 **	0.81
Elastic—not elastic	0.55 **	0.57 *	0.57 **	0.56 *	0.69 **	0.51 *	0.58
Soft—hard (surface)	0.28	0.59 *	0.20	0.29	0.47 *	0.35	0.36
Soft—hard (base)	−0.13	0.47	0.19	−0.10	0.49 *	0.46 *	0.23
Moisturized—dry	0.36	0.07	0.51 *	−0.13	0.38	0.14	0.22
Mean	0.38	0.51	0.46	0.29	0.55	0.46	

For example, A/B indicates the correlation coefficient of ratings between assessors A and B. *: $p < 0.05$, **: $p < 0.01$.

Table 8. Intraclass correlation coefficient (ICC) (1,1) and ICC (1,2) values.

ICC (1,1)					
Assessor	A	B	C	D	
Expert/non-expert	Expert	Expert	Non-expert	Non-expert	Mean
Smooth—rough	0.95	0.85	0.81	0.47	0.77
Elastic—not elastic	0.34	0.68	0.67	0.76	0.61
Soft—hard (surface)	0.51	0.76	0.76	0.59	0.65
Soft—hard (base)	0.54	0.72	0.74	0.66	0.67
Moisturized—dry	0.48	0.48	0.78	0.38	0.53
Mean	0.56	0.70	0.75	0.57	
ICC (1,2)					
Assessor	A	B	C	D	
Expert/non-expert	Expert	Expert	Non-expert	Non-expert	Mean
Smooth—rough	0.97	0.92	0.90	0.64	0.86
Elastic—not elastic	0.65	0.64	0.87	0.55	0.68
Soft—hard (surface)	0.68	0.86	0.86	0.74	0.79
Soft—hard (base)	0.70	0.84	0.85	0.79	0.80
Moisturized—dry	0.51	0.81	0.80	0.86	0.75
Mean	0.70	0.82	0.86	0.70	

Table 9 shows the results of ICC (2,1) and ICC (2,2). The mean ICC (2,1) of *smooth—rough* was 0.64, which was a substantial agreement. The mean ICC (2,1) of *elastic—not elastic* was 0.51 with a moderate agreement. The values for *soft—hard (surface)* and *soft—hard (base)* were 0.27 and 0.23, respectively, with a fair agreement. *Moisturized—dry* exhibited the lowest mean value of 0.17. The reliability, or ICC (2,1), for the expert assessors (A and B) averaged over all attributes was 0.34, which was not largely different from that of other combinations of assessors. The values of ICC (2,2) were higher than those of ICC (2,1), and the means of *smooth—rough* and *elastic—not elastic* were substantial and that of *soft—hard (surface)* was moderate. The mean ICC (2,2) values for *soft—hard (base)* and *moisturized—dry* were fair.

Table 9. Intraclass correlation coefficient (ICC) (2,1) and ICC (2,2) values.

ICC (2,1)							
Comparison	A/B	A/C	A/D	B/C	B/D	C/D	Mean
Smooth—rough	0.70	0.74	0.40	0.81	0.56	0.61	0.64
Elastic—not elastic	0.48	0.41	0.51	0.48	0.68	0.50	0.51
Soft—hard (surface)	0.27	0.51	0.15	0.28	0.30	0.14	0.27
Soft—hard (base)	−0.13	0.46	0.19	−0.09	0.49	0.44	0.23
Moisturized—dry	0.37	0.02	0.37	−0.03	0.25	0.05	0.17
Mean	0.34	0.43	0.32	0.29	0.46	0.35	
ICC (2,2)							
Comparison	A/B	A/C	A/D	B/C	B/D	C/D	Mean
Smooth—rough	0.82	0.85	0.57	0.89	0.72	0.75	0.77
Elastic—not elastic	0.65	0.59	0.68	0.64	0.81	0.67	0.67
Soft—hard (surface)	0.43	0.67	0.25	0.43	0.46	0.24	0.42
Soft—hard (base)	−0.30	0.63	0.32	−0.20	0.66	0.61	0.29
Moisturized—dry	0.54	0.04	0.54	−0.06	0.40	0.10	0.26
Mean	0.43	0.56	0.47	0.34	0.61	0.48	

4. Discussion

We tested the reliability of sensory evaluations of tactile sensation on bare skin, for which little public information was available. As there was no publicly available information to begin with, the results obtained in this study will be useful for conducting reliable sensory evaluations of tactile sensation on skin in the future.

There was a moderate correlation between the instrumental measurement values and scores for each attribute. As in Table 4, the non-uniform distribution of skin furrows led to the judgment of roughness. *Moisturized—dry* exhibited a low correlation with Corneometer CM825 values, but a high correlation with Skicon-200EX values. This may be because of differences in their measurement principles. The Skicon-200EX is considered more suitable for the surface moisture content of the skin than the Corneometer CM825 [33]. *Elastic—not elastic* was correlated with R1, R2, R4, and R5 calculated from the Cutometer CT580. R1 and R4 indicate the magnitudes of residual deformation at the first and second relaxation, respectively. R2 is the proportion of the elastic recovery of skin deformation after relaxation. R5 is the ratio of the amount of elastic deformation during suction to the amount of immediate recovery during relaxation. R2 and R5 are related to elasticity. *Soft—hard (surface)* and *soft—hard (base)* were correlated with R0, R3, R8, and F1. R0 and R3 represent the first and second maximum suction, respectively. R8 is the amount of recovery during relaxation, and F1 is defined as the area of the waveform showing the time variation in the amount of suction during relaxation, with a smaller area indicating more elastic or less viscous properties. These correlations between the sensory evaluation

ratings and instrumental values indicate that the assessors' sensory evaluations captured the physical aspects of human skin.

The inter-assessor reliability was not very high for softness or moistness in the present experiment. This could be in part attributed to the individual differences in the assessors' finger conditions. Previous studies have reported that differences in finger size and stiffness affect the perception of softness [34,35]. Therefore, the inter-assessor reliability of softness may be reduced by individual differences in finger size and stiffness. It is also known that finger moisture content affects friction, and that friction is associated with sensory evaluation of moistness [36–41]. Adhesion friction, which is thought to be the major force of friction when touching skin with a finger, is expressed as the product of the contact area and the shear strength of the adhesive surface. Different moisture contents of the fingers may result in a different adhesion force and softness of the skin surface [39], which in turn changes the contact area. Therefore, the friction generated by different assessors may differ, resulting in a decrease in inter-assessor reliability of moistness [42–44].

The agreement between the trained assessors was similar to that between the untrained assessors. After the test, expert evaluator A commented that the evaluation was more difficult than usual because the skin on her fingers was rough and different from usual. This suggests that it is difficult for even expert assessors to make reliable evaluations depending on the condition of their fingers. Therefore, it is important to manage the condition of the assessors' fingers by measuring the moisture content and stiffness at each assessment and training.

The values of ICC (1,2) and ICC (2,2) were greater than those of ICC (1,1) and ICC (2,1) suggesting that reliability can be improved by using the average of two repetitions. To improve the reliability of *soft-hard (surface)*, *soft-hard (base)*, and *moisturized-dry*, which exhibited low reliability in this study, the number of evaluation trials by the same panelists will need to be increased.

The following suggestions are made based on the results of this study: (a) *soft-hard (surface)*, *soft-hard (base)*, and *moisturized-dry* are attributes that require intensive training; (b) the condition of assessors' fingers should be thoroughly managed during evaluation and training; and (c) the evaluation should be repeated two or more times and the average value should be used as the score.

The greatest limitation of the study is the small number of panelists in the experiment. For this study, we did not employ and train novice panelists. Instead, we employed two experts with occupational experience of at least five years. It is true that this small sample size limits the generalizability of the study. However, our purpose was to investigate the reliability of such experts' sensory evaluation, and only two such experts were found in the authors' institution. A complementary study needs to be performed in the future in which more panelists are trained to investigate their reliability. However, such trained panelists would not be deemed experts.

5. Conclusions

This study investigated the reliability of sensory evaluations of tactile sensation on bare skin, which is important for the development of skin care products. Earlier studies had rarely reported the quantitative reliabilities for multiple types of tactile attributes, including surface smoothness, hardness, and moisture. We conducted a sensory evaluation study with two expert assessors, two untrained assessors, and 22 female examinees whose cheeks were tested. The scores of the five types of attributes were moderately correlated with the instrumental values using commercial viscoelastic and water-content measurement systems and an in-house image-based surface roughness sensor. We found that the reliability largely depended on the attributes. The reliabilities of *smooth-rough* and *elastic-not elastic* were highly consistent among the assessors. In contrast, the reliability of *soft-hard (surface)*, *soft-hard (base)*, and *moisturized-dry* were in fair agreement. We did not find substantial differences in the reliability between the expert and untrained panelists. Finally, we made a few suggestions to improve the reliability of tactile sensory evaluations of bare skin.

Author Contributions: Conceptualization: N.S., K.M., T.K. and N.A.; methodology, N.S., K.M. and T.K.; formal analysis, N.S.; investigation, N.S., K.M. and T.K.; resources, N.S.; data curation, N.S.; writing—original draft preparation, N.S.; writing—review and editing, N.A. and S.O.; visualization, N.S.; supervision, S.O.; project administration, N.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was approved by the Human Study Ethics Committee of the Shiseido Global Innovation Center (Study No. C02205, C02206). Written informed consent was obtained from the four assessors and 22 participants of the study.

Informed Consent Statement: Informed consent was obtained from all participants involved in the study.

Data Availability Statement: Not applicable.

Acknowledgments: We are grateful to Megumi Mizugaki and Keiko Doi for their help during this study.

Conflicts of Interest: Professional assessor A is one of the authors. However, she was not informed of the details of the test procedure. Further, a blind design was adopted in the experiments.

References

1. Cosmetics Europe: Guideline for the Evaluation of the Efficacy of Cosmetic Products. Available online: https://www.cosmeticseurope.eu/files/4214/6407/6830/Guidelines_for_the_Evaluation_of_the_Efficacy_of_Cosmetic_Products_-_2008.pdf (accessed on 30 July 2022).
2. *Standard 11036; Sensory Analysis—Methodology—Texture Profile*. International Organization for Standardization: Geneva, Switzerland, 2020.
3. Huber, P. Sensory measurement—Evaluation and testing of cosmetic products. In *Cosmetic Science and Technology: Theoretical Principles and Applications*; Sakamoto, K., Lochhead, R.Y., Maibach, H.I., Yamashita, Y., Eds.; Elsevier: Amsterdam, The Netherlands, 2017; pp. 617–633.
4. IFSCC Monograph Principles of Product Evaluation Objective Sensory Methods. Available online: <https://ifsc.org/wp-content/uploads/2018/05/1-Principles-of-Product-Evaluation.pdf> (accessed on 8 February 2022).
5. ASTM International. *Standard Guide for Two Sensory Descriptive Analysis Approaches for Skin Creams and Lotions*; ASTM International: West Conshohocken, PA, USA, 2019; pp. 1490–1519.
6. ASTM International. 2082–12. *Standard Guide for Descriptive Analysis of Shampoo Performance*; ASTM International: West Conshohocken, PA, USA, 2020.
7. Serup, J. EEMCO guidance for the assessment of dry skin (xerosis) and ichthyosis: Clinical scoring systems. *Ski. Res. Technol.* **1995**, *1*, 109–114. [[CrossRef](#)] [[PubMed](#)]
8. Lévêque, J.L. EEMCO guidance for the assessment of skin topography. the European Expert Group on efficacy measurement of cosmetics and other topical products. *J. Eur. Acad. Dermatol. Venereol.* **1999**, *12*, 103–114. [[PubMed](#)]
9. Pensé-Lhéritier, A.M. Recent developments in the sensorial assessment of cosmetic products: A review. *Int. J. Cosmet. Sci.* **2015**, *37*, 465–473. [[CrossRef](#)] [[PubMed](#)]
10. Blaak, J.; Keller, D.; Simon, I.; Schleifinger, M.; Schürer, N.Y.; Staib, P. Consumer panel size in sensory cosmetic product evaluation: A pilot study from a statistical point of view. *J. Cosmet. Dermatol. Sci. Appl.* **2018**, *8*, 97–109. [[CrossRef](#)]
11. Nobile, V. Guidelines on cosmetic efficacy testing on humans. Ethical, technical, and regulatory requirements in the main cosmetics markets. *J. Cosmetol. Trichol.* **2016**, *2*, 107. [[CrossRef](#)]
12. Messaraa, C.; Drevet, J.; Jameson, D.; Zuanazzi, G.; De Ponti, I. Can performance and gentleness be reconciled? A skin care approach for sensitive skin. *Cosmetics* **2022**, *9*, 34. [[CrossRef](#)]
13. Addor, F.A.S.; de Souza, M.C.; Trapp, S.; Peltier, E.; Canosa, J.M. Efficacy and safety of topical dexpanthenol-containing spray and cream in the recovery of the skin integrity compared with petroleum jelly after dermatologic aesthetic procedures. *Cosmetics* **2021**, *8*, 87. [[CrossRef](#)]
14. Stettler, H.; Crowther, J.; Boxshall, A.; Bielfeldt, S.; Lu, B.; de Salvo, R.; Trapp, S.; Blenkiron, P. Biophysical and subject-based assessment of the effects of topical moisturizer usage on xerotic skin—Part II: Visioscan®VC 20plus imaging. *Cosmetics* **2022**, *9*, 5. [[CrossRef](#)]
15. Guest, S.; Mehrabyan, A.; Essick, G.; Phillips, N.; Hopkinson, A.; McGlone, F. Physics and tactile perception of fluid-covered surfaces. *J. Texture Stud.* **2021**, *43*, 77–93. [[CrossRef](#)]
16. Rigano, L.; Montoli, M. Strategy for the development of a new lipstick formula. *Cosmetics* **2021**, *8*, 105. [[CrossRef](#)]
17. Aust, L.B.; Oddo, L.P.; Wild, J.E.; Mills, O.H.; Deupree, J.S. The descriptive analysis of skin care products by a trained panel of judge. *J. Soc. Cosmet. Chem.* **1987**, *38*, 443–449.

18. Vieira, G.S.; Filho, P.A.R. Sensory analysis: Panel training. In Proceedings of the IFSCC Conference, Zurich, Switzerland, 21–23 September 2015; p. 203.
19. Calixto, L.S.; Maia Campos, P.M.; Picard, C.; Savary, G. Brazilian and French sensory perception of complex cosmetic formulations: A cross-cultural study. *Int. J. Cosmet. Sci.* **2020**, *42*, 60–67. [\[CrossRef\]](#)
20. Larnier, C.; Ortonne, J.P.; Venot, A.; Faivre, B.; Béani, J.C.; Thomas, P.; Brown, T.C.; Sendagorta, E. Evaluation of cutaneous photodamage using a photographic scale. *Br. J. Dermatol.* **1994**, *130*, 167–173. [\[CrossRef\]](#)
21. Kang, B.C.; Kim, Y.E.; Kim, Y.J.; Chang, M.J.; Choi, H.D.; Li, K.; Shin, W.G. Optimizing EEMCO guidance for the assessment of dry skin (xerosis) for pharmacies. *Ski. Res. Technol.* **2014**, *20*, 87–91. [\[CrossRef\]](#)
22. Nakatani, M.; Fukuda, T.; Sasamoto, H.; Arakawa, N.; Otaka, H.; Kawasoe, T.; Omata, S. Relationship between perceived softness of bilayered skin models and their mechanical properties measured with a dual-sensor probe. *Int. J. Cosmet. Sci.* **2013**, *35*, 84–88. [\[CrossRef\]](#)
23. Adejokun, D.A.; Dodou, K. Quantitative sensory interpretation of rheological parameters of a cream formulation. *Cosmetics* **2020**, *7*, 2. [\[CrossRef\]](#)
24. Arakawa, N.; Watanabe, T.; Fukushima, K.; Nakatani, M. Sensory words may facilitate certain haptic exploratory procedures in facial cosmetics. *Int. J. Cosmet. Sci.* **2021**, *43*, 78–87. [\[CrossRef\]](#)
25. Iida, I.; Noro, K. An analysis of the reduction of elasticity on the ageing of human skin and the recovering effect of a facial massage. *Ergonomics* **1995**, *38*, 1921–1931. [\[CrossRef\]](#)
26. Shimizu, R.; Nonomura, Y. Preparation of artificial skin that mimics human skin surface and mechanical properties. *J. Oleo Sci.* **2018**, *67*, 47–54. [\[CrossRef\]](#)
27. Takahashi, M. Image analysis of skin surface contour. *Acta Derm. Venereol. Suppl.* **1994**, *185*, 9–14.
28. Cutometer® Dual MPA 580. Available online: <https://www.courage-khazaka.de/en/16-wissenschaftliche-produkte/alle-produkte/266-cutometer-new-e> (accessed on 14 July 2022).
29. Bland, J.M.; Altman, D.G. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1986**, *1*, 307–310. [\[CrossRef\]](#)
30. Evans, J.D. *Straight Forward Statistics for the Behavioral Sciences*; Brooks/Cole Publishing Co.: Thomson, CA, USA, 1996.
31. Shrout, P.E.; Fleiss, J.L. Intraclass correlations: Uses in assessing rater reliability. *Psychol. Bull.* **1979**, *86*, 420–428. [\[CrossRef\]](#)
32. Fleiss, J.L.; Cohen, J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ. Psychol. Meas.* **1973**, *33*, 613–619. [\[CrossRef\]](#)
33. Clarys, P.; Clijsen, R.; Taeymans, J.; Barel, A.O. Hydration measurements of the stratum corneum: Comparison between the capacitance method (digital version of the Corneometer CM 825®) and the impedance method (Skicon-200EX®). *Ski. Res. Technol.* **2012**, *18*, 316–323. [\[CrossRef\]](#)
34. Li, B.; Gerling, G.J. Individual differences impacting skin deformation and tactile discrimination with compliant elastic surfaces. *World Haptics Conf.* **2021**, *2021*, 721–726.
35. Xu, C.; Wang, Y.; Gerling, G.J. Individual performance in compliance discrimination is constrained by skin mechanics but improved under active control. *World Haptics Conf.* **2021**, *2021*, 445–450.
36. Derler, S.; Gerhardt, L.-C. Tribology of skin: Review and analysis of experimental results for the friction coefficient of human skin. *Tribol. Lett.* **2012**, *45*, 1–27. [\[CrossRef\]](#)
37. Sivamani, R.K.; Goodman, J.; Gitis, N.V.; Maibach, H.I. Friction coefficient of skin in real-time. *Ski. Res. Technol.* **2003**, *9*, 235–239. [\[CrossRef\]](#) [\[PubMed\]](#)
38. Koudine, A.A.; Barquins, M.; Anthoine, P.H.; Aubert, L.; Lévêque, J.L. Frictional properties of skin: Proposal of a new approach. *Int. J. Cosmet. Sci.* **2000**, *22*, 11–20. [\[CrossRef\]](#)
39. Dzidek, B.M.; Adams, M.; Zhang, Z.; Johnson, S.; Bocheureau, S.; Hayward, V. Role of occlusion in non-coulombic slip of the finger pad. In *Neuroscience, Devices, Modeling, and Applications*; Auvray, M., Duriez, C., Eds.; Springer: Berlin, Germany, 2014; pp. 109–116.
40. Sakata, Y.; Mayama, H.; Nonomura, Y. Friction dynamics of moisturized human skin under non-linear motion. *Int. J. Cosmet. Sci.* **2021**, *44*, 20–29. [\[CrossRef\]](#)
41. Egawa, M.; Oguri, M.; Hirao, T.; Takahashi, M.; Miyakawa, M. The evaluation of skin friction using a frictional feel analyzer. *Ski. Res. Technol.* **2002**, *8*, 41–51. [\[CrossRef\]](#) [\[PubMed\]](#)
42. Adams, M.J.; Briscoe, B.J.; Johnson, S.A. Friction and lubrication of human skin. *Tribol. Lett.* **2007**, *26*, 239–253. [\[CrossRef\]](#)
43. Nacht, S.; Close, J.A.; Yeung, D.; Gans, E.H. Skin friction coefficient: Changes induced by skin hydration and emollient application and correlation with perceived skin feel. *J. Soc. Cosmet. Chem.* **1981**, *32*, 55–56.
44. Highley, K.R.; Coomey, M.; DenBeste, M.; Wolfram, L.J. Frictional properties of skin. *J. Investig. Dermatol.* **1977**, *69*, 303–305. [\[CrossRef\]](#)