Journal of Cosmetics, Dermatological Sciences and Applications, 2018, 8, 97-109
http://www.scirp.org/journal/jcdsa
ISSN Online: 2161-4512
ISSN Print: 2161-4105

Consumer Panel Size in Sensory Cosmetic Product Evaluation: A Pilot Study from a Statistical Point of View

Jürgen Blaak[1], Daniela Keller[2], Isabel Simon[1], Marina Schleißinger[1], Nanna Y. Schürer[3], Peter Staib[1*]

[1]Research & Development and Regulatory Affairs, Kneipp GmbH, Würzburg, Germany
[2]Statistik & Beratung Daniela Keller, Kürnach, Germany
[3]Department of Dermatology, Environmental Medicine and Health Theory, University of Osnabrück, Germany
Email: *peter.staib@kneipp.de

How to cite this paper: Blaak, J., Keller, D., Simon, I., Schleißinger, M., Schürer, N.Y. and Staib, P. (2018) Consumer Panel Size in Sensory Cosmetic Product Evaluation: A Pilot Study from a Statistical Point of View. *Journal of Cosmetics, Dermatological Sciences and Applications*, **8**, 97-109.
https://doi.org/10.4236/jcdsa.2018.83012

Received: June 19, 2018
Accepted: August 28, 2018
Published: August 31, 2018

Abstract

Subjective evaluations are nowadays applied more commonly in cosmetic product assessment. They are used in quality control, product development steps and efficacy studies for claim support. Several studies have been published to determine the adequate number of panelists, but recommendations and guidelines dealing with this topic are rare in the cosmetic sector. The aim of the present pilot study was to recommend a suitable study plan and define the adequate consumer panel size for cosmetic consumer assessment. A questionnaire-based product evaluation study, with three different cosmetic products, was organized as a consumer test using a seven-point scale. As a last step, a specific statistical calculation was performed to define the minimum sample size. It showed that the minimum sample size, besides the obvious statistical parameters of standard deviation and confidence interval, also depends on age and gender of the panelists and product assessment item. Utilizing a CI of 95% a minimum of 60 panelists seems to be sufficient for home-use-test (HUT) with a given seven-point scale. A minimum of 101 panelists are shown to be sufficient utilizing a CI of 99%.

Keywords

Claim Substantiation, Consumer Panel Size, Product Evaluation, Sample Size Calculation

1. Introduction

Subjective investigations are nowadays applied more commonly because of the

essential need for direct consumer feedback among manufacturers of fast moving consumer goods, like cosmetics [1]. They are fundamental in cosmetic science [2] and have been proven to be a highly effective tool, where subjects are the measuring instrument [1]. The results of product evaluation through panelists provide information that could not be obtained by methodologies in which instrumental measurements are used [3]. Optimally, the results obtain detailed, robust, consistent and reproducible results, which are stable over time [4]. The principle uses of sensory techniques are in quality control, product development processes or efficacy studies for claim substantiation. Through analysis, information from consumer personal test responses can be assessed and manufacturers can estimate the potential viability of new product designs or developments in the marketplace [1] [5] [6]. Moreover, evaluation of product sensory characteristics is an essential component in the maintenance, optimization and quality improvement of consumer goods [7]. A distinction is made between three types of assessors in a panel: naive, selected and trained [5] [8]. The expert panel, with trained assessors, has its roots in traditions such as perfumer, mast brewer and winemaker [5]. The assessors judge the intensity of product attributes or the reaction of a certain body area [9]. Extensive training led to a calibrated sensory panel where the assessor can perform like an "instrument" [5]. Therefore, they can be considered as a replacement for unavailable instruments [9]. This procedure implies high quality and stable results [4]. However, the assessors' training can have the effect of uniform criteria of acceptability and, consequently, a low dispersion of the criteria [10]. The creation and continuity of a well-trained expert panel involve high costs and are time-consuming [11]. Experienced market researchers know that consumers differ dramatically from each other [5]. Thus, the trend of panels with semi-experienced, or even naive, panelists has emerged recently as the assumption that consumers are able to describe products successfully is more and more accepted within the sensory science community [4]. Naive assessors have had no training in sensory methodology and can be defined as consumers [8]; therefore, home-use-tests (HUT) represent the ultimate approach in consumer research [1]. With HUT, products are tested under natural conditions of use [1]. Manufacturers in the food industry make notable use of the trend as it has become a powerful tool in sensory research [4]. That means that naive panelists are accepted in sensory communities comparable to trained panelists [11]. Moreover, consumer panels bring the benefit of being faster, partly less costly compared to experimental investigations and, above all, provide direct feedback from consumers [11] [12]. In the present work "consumer test" is defined as cosmetic product evaluation with inexperienced panelists, performed as HUT. The rise in the use of consumer tests has increased the need for developing guidelines to ensure standardization and best practices [11]. Regarding this methodology, base size is an important key issue. Because it is substantial for outcome and statistics, it should be considered at the beginning of study planning. Base size in sensory evaluation is discussed in different corresponding textbooks, but the recommended number of panelists varies considerably [5]

[10] [13] [14] [15]. Costs and statistical quality depend upon the number of participants [16]. In general, one may say a reduction in base size leads to an increase in risk and decrease in costs [16]. In product evaluations, specifically in claim substantiation studies, a larger base size is recommended [5] [17]. Base size influences the stability of the averages and affects individual differences [5]. Therefore, a large base size is suitable to cancel out individual differences [17]. However, researchers miss a clear guidance from the literature as to how many consumers are adequate [10]. Hence, variations in base size occur in practice and many performed studies did not give a reason for choosing the number of consumers they used for measuring sensory acceptability [10] [18]. Several studies have been published evaluating the number of consumers needed for acceptability tests. Thereby, the involved and recommended number of panelists ranges extremely. Moskowitz [5] recommends a base size of 40 to 60 subjects, and describes that more than 100 assessors are not necessary. Stone and Sidel [7] suggest involving 25 to 50 panelists in laboratory tests and about twice as many in HUT. Meillgaard *et al.* [1] stated 50 to 300 participants per location for central location tests and 75 to 300 participants per city for HUT. Mammasse and Schlich [19] determined that an adequate base size depends mainly on the level of complexity of the product space, which has also been stated by Basker [20]. For that reason, Mammasse and Schlich declared that defining a global base size valid for every study is not possible [19]. The US Army also did research on study design and evaluations of food acceptances in the first half of the 20th century [21]. Due to the results, a large number of scientists and technologists developed systematic techniques later in the century. Sensory testing as a formalized mythology has only been recently developed by scientists [1]. Regarding the base size question in consumer tests, one can find individual papers published in journals like *Food Quality and Preference* or *Journal of Food Science* or *Food Technology* regarding food research and development [1]. To our knowledge, studies discussing the adequate panel size in the cosmetic sector are rare. Cosmetic companies utilize sensory panels and consumer use trials for the same research, development and economic reasons as food manufactures [22]. Moreover, any efficacy claim on cosmetic products should be scientifically verified, e.g. prescribed in the current EU cosmetic legislation (Regulation (EC) No. 1223/2009; Regulation (EU) No. 655/2013). Nevertheless, guidelines with information regarding recommendations on panel size for HUT are few published. Kramer and Thiemann [17], Montgomery [23] and Gacula and Sigh [24] established formulas to calculate panel size on the basis of four, estimated parameters: Type I-error, Type II-error, SE and differences in means. Gacula and Singh [24] stated that the larger the base size, the smaller the standard error and margin of error in an estimate. In 2005, Hough *et al.* [10] were the first to present the basic concepts necessary to estimate the number of panelists for consumer studies. In addition, they set out variability data on previous consumer studies from different countries, which are necessary for estimating base size, but also to estimate base size for different values of error types (Type I and Type II). Nowadays, free

software packages, like G*Power, are available, by which calculation of the individual base size for a certain experiment is possible [25]. In summary, panel size plays a crucial role in planning consumer tests, not the least of which is to fulfill legal requirements. However, publications, recommendations and guidelines dealing with this topic are rare in the cosmetic sector. Against this background, the aim of the present study was to give an indication for future study planning concerning consumer panel size in cosmetic HUT. Therefore, we posed the question: How many consumer panelists are necessary to get meaningful and informative data and results? To address this issue, we performed a questionnaire-based evaluation with three different cosmetic emulsions as HUT, followed by a specific statistical calculation.

2. Experimental

2.1. Participants, Test Products and Instructions

For the present HUT, 110 German users of skin care products were recruited in the winter months by University of Osnabrück where a large study panel is established. In total, 101 questionnaires came back correctly filled in and the data were analyzed. Sample consisted of 89 females and 12 males, aged between 18 and 85 years (mean 41.9 years, SD ± 16.1; median 42.0). The inclusion criteria reflecting ethical aspects were: informed consent and willingness to actively participate in the study, age ≥ 18 years, regular use of skin care products and, due to the tested product range, "slightly or moderately dry skin". No further age- or gender-specific restrictions were made. Subjects with chronic skin disorders or acute skin lesions were excluded from the study. Three different leave-on products (face cream, hand cream, body lotion) were presented in white coded samples to avoid bias. The test formulations were emulsions (oil in water, O/W), developed to be suitable for slightly or moderately dry skin conditions and already on the market as part of a commonly available skin care range (Kneipp GmbH, Würzburg, Germany) (**Table 1**). Safety assessment according to EU Cosmetic Regulation No. 1223/2009 including dermatological compatibility examination such as closed epicutaneous test has been conducted for all test formulations. The HUT was performed as a within-subject design, and all participants got a product set and questionnaire for each product. Every product was used for seven days; hence, the application period amounted to 21 days in total. Participants were allowed to choose which product sequence to use, introducing randomization and thereby prevent order effects. Furthermore, the participants were instructed not to apply any types of products with skin care properties in the test area throughout the entire course of the study. While the use of cleaning products like soap was allowed, the participants were permitted to apply any other cosmetic products, e.g. sunscreen or decorative cosmetics.

2.2. Subjective Product Evaluation

The questionnaire—originally applied in German—consisted of two main parts:

1) the application recommendation and 2) the product evaluation. By means of the application recommendations, the panelists were introduced to the test design and received brief instructions on what to do. In the product evaluation, seven characteristic parameters were queried. Those were skin tolerability, skin care effect, skin feeling, product absorption, smell, overall impression and recommendation. For each parameter, the participants were asked to evaluate to what extend they agreed or disagreed with specified statements on an endorsement scale, shown in Table 2. The gradations of the applied scale were equidistant and numerically labelled. The equidistant points served to define the degree and orientation of the continuum effectively [26]. To ease the usage, the ends/extremes were accompanied by extreme terms/annotations. A horizontal, seven-point scale (Likert item) was used because it included the range between "1 = strongly agree" and "7 = strongly disagree", which automatically includes a center result. With this, the participants were offered greater freedom to express their sensory perception through selecting a middle opinion. Rating scales with seven, nine or 10 consumer response categories are generally most preferred. Results from Preston and Colman [27] revealed that scores from scales with between seven and 10 response categories are the most reliable. The most valid and discriminating were those more than six response categories.

Table 1. Composition of the test products according to INCI.

Face cream (no. VIHC160)
Aqua, *Prunus amygdalus dulcis* Oil, Glycerin, Cetearyl Alcohol, Cetearyl Olivate, Tocopheryl Acetate, Sorbitan Olivate, *Persea gratissima* Oil, Panthenol, Butyrospermum Parkii Butter, Candelilla Cera, *Olea europaea* Fruit Oil, Phytosterols, *Prunus amygdalus dulcis* Flower Extract, Retinyl Palmitate, *Citrus Aurantium dulcis* Peel Oil, *Pogostemon cablin* Oil, Limonene, Linalool, Coumarin, Citronellol, Geraniol, p-Anisic-Acid, Parfum, *Helianthus annuus* Hybrid Oil, Caprylyl Glycol, Acrylates/C10-30 Alkyl Acrylate Crosspolymer, Arginine, Tocopherol
Hand cream (no. MBH126)
Aqua, Glycerin, Caprylic/Capric Triglyceride, *Prunus Amygdalus Dulcis* Oil, Tocopheryl Acetate, Panthenol, Distarch Phosphate, Cetearyl Olivate, Sorbitan Olivate, Candelilla Cera, Cetearyl Alcohol, Butyrospermum Parkii Butter, Argania Spinosa Kernel Oil, *Prunus amygdalus dulcis* Flower Extract, Bisabolol, Retinyl Palmitate, *Citrus Aurantium dulcis* Peel Oil, *Pogostemon cablin* Oil, Limonene, Linalool, Coumarin, Citronellol, Geraniol, Citral, p-Anisic-Acid, Parfum, *Helianthus annuus* Hybrid Oil, Caprylyl Glycol, Acrylates/C10-30 Alkyl Acrylate Crosspolymer, Sodium Stearoyl Glutamate, Tocopherol
Body lotion (no. VIHL131)
Aqua, *Prunus Amygdalus Dulcis* Oil, Glycerin, Tocopheryl Acetate, Cetearyl Olivate, Cetearyl Alcohol, Butyrospermum Parkii Butter, Panthenol, Sorbitan Olivate, *Prunus amygdalus dulcis* Flower Extract, Retinyl Palmitate, *Citrus Aurantium dulcis* Peel Oil, *Pogostemon cablin* Oil, Limonene, Linalool, Coumarin, Citronellol, Geraniol, Citral, p-Anisic-Acid, Parfum, Caprylyl Glycol, Xanthan Gum, Acrylates/C10-30 Alkyl Acrylate Crosspolymer, Arginine, *Helianthus annuus* Hybrid Oil, Tocopherol

Table 2. Questionnaire (seven-point scale) for product evaluation after a seven-day usage period (English translation; originally in German).

How do you assess the following statement?	Evaluation scale						
	Strongly agree						Strongly disagree
	1	2	3	4	5	6	7
The product is very well-tolerated.	O	O	O	O	O	O	O
The product has an intensive care effect.	O	O	O	O	O	O	O
The product provides a very pleasant skin sensation.	O	O	O	O	O	O	O
The product absorbs perfectly.	O	O	O	O	O	O	O
The product possesses a very pleasant smell.	O	O	O	O	O	O	O
My overall impression of the product is very positive.	O	O	O	O	O	O	O
I would recommend the product to others.	O	O	O	O	O	O	O

2.3. Statistics and Sample Size Calculation

The statistical analysis was divided into two consecutive steps: first, a descriptive analysis of each item for each product (face cream, hand cream, body lotion) was performed, by calculation of the means, the corresponding standard deviations (SD), the min/max values and the 95% confidence interval (CI). Furthermore, data analysis by Q-Q plot showed nearly normal distribution. Afterwards, the data were divided by age (≤42 and 42+ years, median: 42 years) and gender, followed by descriptive calculations as aforementioned. Due to the imbalance between the female- and male-groups (n = 89 vs. n = 12), we only analyzed age-specific relations (Spearman rank correlation coefficient). The level of significance was set at $P = 0.05$. In a second stage, a sample size calculation was conducted. It was based on the formula for confidence intervals for means, as this was the desired method for presenting the results of later studies. Presentation with confidence intervals enables interpretation of the results with respect to location (mean) and estimation of confidence. The confidence depends on the width of the interval and the probability (level of confidence). For the present study, confidence levels of 95% and 99% were chosen, and a width of one and two units. Therefore, sample size for different situations concerning confidence can be compared. Additionally, the descriptive data of separate groups (age and gender) and for different product assessment items were used. All statistical calculations were conducted using the R statistical software (Version 3.2.2, The R Foundation for Statistical Computing, Vienna, Austria).

3. Results

Guidelines and publications concerning consumer panel size in cosmetic product evaluations via HUT are rare. Therefore, the present study aimed to give a recommendation to plan studies and evaluate the consumer panel size for cosmetic HUT. In total, 110 German users of skin care products were recruited. Fi-

nally, 101 questionnaires came back correctly filled in and the data were analyzed. The sample consisted of 89 females and 12 males, aged between 18 and 85 years (mean 41.9 years, ±16.1 SD; median 42.0). Table 3 demonstrates the combined evaluation of the three different product categories (mean value per panelist for face cream, hand cream and body lotion). In total, the evaluation rate ranged between 1.92 ± 1.23 SD (… well tolerated) and 3.07 ± 1.75 SD (… very pleasant smell). Moreover, the ratings (% of panelists) are displayed in Table 3; e.g. 54% of the volunteers "strongly agree" with the statement "The product is very well-tolerated". Additionally, Figure 1 represents consumer responses from "strongly agree" over "intermediate" to "strongly disagree" for each product assessment item.

To visualize aged-related effects on product evaluation, two age groups were generated based on the median age (≤42 years, 42+ years). Table 4 illustrates the correlation between age (in years) and product assessment. The statistical analysis revealed no significant correlation to subject ages, except for two items "... absorbs perfectly" and "... recommendation". Concerning these two product assessment items, a significant negative correlation (−0.305 rho, −0.223 rho) was calculated. Negative correlation means that agreement increases (decreasing values) with increasing age.

Based on the sample size calculation, the minimum sample size ranges between n = 5 to n = 58 (95% CI) and n = 8 to n = 101 (99% CI), presented in Table 5. The ranges of the SD were freely determined by the authors and are defined by the highest and lowest SD values. The minimum sample size varies depending on the applied CI. However, the minimum sample size depends not only on SD and CI, but also depends on the product assessment item as well as age and gender of volunteers (Table 6). The 95% CI was used for further analytical considerations [28].

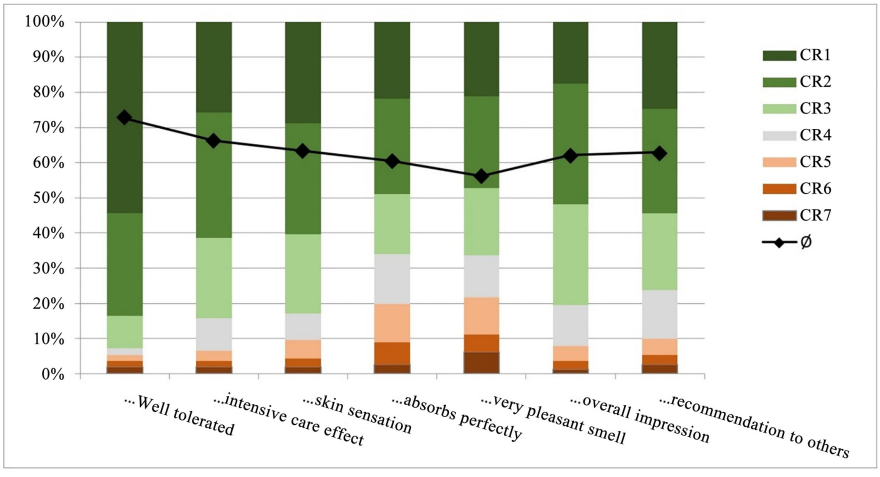

Figure 1. Descriptive data and rating (%) of panelist product assessments by Likert scale. Combined data of face cream, hand cream and body lotion. Color code: green spectrum = range of agreement, grey = intermediate, orange = range of disagreement, CR = Consumer Response (CR1 = "strongly agree" and CR7 = "strongly disagree"); ϕ = Mean %.

Table 3. Descriptive data and rating (%) of panelist product assessments by Likert scale (1 = "strongly agree" and 7 = "strongly disagree"). Combined data of face cream, hand cream and body lotion. SD = Standard deviation; CI = Confidence interval; Color code: orange = ≥ 10% of panelist, blue ≤ 10% and green ≤ 5%.

	Descriptive data			Rating (% of panelists)						
Product assessment item	Mean	SD	95% CI	1	2	3	4	5	6	7
… well-tolerated	1.92	±1.23	1.68 - 2.16	54	29	9	2	2	2	2
… intensive care effect	2.36	±1.25	2.11 - 2.60	26	36	23	9	3	2	2
… skin sensation	2.56	±1.41	2.29 - 2.84	29	32	22	8	5	2	2
… absorbs perfectly	2.77	±1.43	2.49 - 3.05	22	27	17	14	11	6	3
… very pleasant smell	3.07	±1.75	2.73 - 3.41	21	26	19	12	11	5	6
… overall impression	2.65	±1.20	2.42 - 2.89	17	34	29	12	4	2	1
… recommendation	2.59	±1.37	2.33 - 2.86	25	30	22	14	3	3	3

Table 4. Impact of consumer age on product evaluation. Age groups were defined on age median (42 years). CI = Confidence interval; P = P-value; rho = Spearman rank correlation coefficient; ns = No significance (correlation is significant at the 0.05 level).

	Age group (years)					
	≤42 (n = 51)		42+ (n = 50)		Correlation	
Product assessment item	Mean	95% CI	Mean	95% CI	rho	P
… well-tolerated	2.02	1.70 - 2.34	1.82	1.46 - 2.18	−0.111	ns
… intensive care effect	2.43	2.11 - 2.76	2.28	1.91 - 2.65	−0.127	ns
… skin sensation	2.73	2.38 - 3.07	2.40	1.97 - 2.83	−0.143	ns
… absorbs perfectly	3.24	2.84 - 3.63	2.30	1.95 - 2.65	−0.305	0.002
… very pleasant smell	2.94	2.46 - 3.42	3.20	2.72 - 3.68	0.050	ns
… overall impression	2.82	2.53 - 3.11	2.48	2.11 - 2.85	−0.195	ns
… recommendation	2.84	2.45 - 3.24	2.34	1.99 - 2.69	−0.223	0.025

Table 5. Minimum sample size (n) for different standard deviations and a desired CI of 95% (width one unit) or 99% (width one unit). CI = Confidence interval; a = Lowest observed value (SD 0.5149); b = Highest observed value (SD 1.9388).

	95% CI		99% CI	
Standard deviation	n	z-score	n	z-score
0.51a	5	1.96	8	2.58
0.52 - 1.00	16	1.96	27	2.58
1.00 - 1.50	35	1.96	60	2.58
1.50 - 1.94b	58	1.96	101	2.58

Table 6. The minimum sample size under a desired confidence interval of 95% ranges from n = 5 to n = 58 and is influenced by assessment item, gender, age group and test product. Color code: green = n = 58, blue = n = 35, orange = n = 16 and red = n = 5.

	Minimum sample size (n)				
	Age group		Gender		
Product assessment item	≤42 (n = 51)	42+ (n = 50)	Female (n = 89)	Male (n = 12)	Total (n = 101)
Face cream (no. VIHC160)					
… well-tolerated	16	16	35	5	35
… intensive care effect	16	16	35	16	35
… skin sensation	16	35	35	35	35
… absorbs perfectly	16	16	35	58	35
… very pleasant smell	35	35	58	35	58
… overall impression	16	16	35	16	35
… recommendation	16	16	35	35	35
Hand cream (no. MBH126)					
… well-tolerated	16	16	16	16	16
… intensive care effect	16	35	35	16	35
… skin sensation	16	35	35	16	35
… absorbs perfectly	58	35	35	35	35
… very pleasant smell	58	58	58	58	58
… overall impression	16	35	35	16	35
… recommendation	35	35	35	35	35
Body lotion (no. VIHL131)					
… well-tolerated	58	58	58	5	58
… intensive care effect	35	58	58	16	58
… skin sensation	35	58	58	16	58
… absorbs perfectly	58	58	58	35	58
… very pleasant smell	58	58	58	58	58
… overall impression	35	58	58	16	35
… recommendation	58	58	58	35	58

Evaluation of the body lotion required the highest sample size (n = 58) compared to face and hand cream, which was independent from the specific assessment item. For the face cream, the calculated sample size was lower (n = 35) with exception of the item "… very pleasant smell" (n = 58). Concerning the hand cream, n = 35 was also the calculated sample size, but with exception of "… well-tolerated" (n = 16) and "… very pleasant smell" (n = 58). Taking the results together, they demonstrate that assessment of product scent ("… very pleasant smell") needs the highest sample size of the three product categories.

In addition, the age- and gender-related impact on product assessment and calculated sample size is shown in Table 6. Comparing the two age groups, a tendency towards a higher sample size in the group 42+ years is shown, especially in the evaluation of the body lotion. Comparing gender results, there is a tendency to a higher sample size in the female group, especially for the evaluation of skin tolerance ("… well-tolerated").

4. Discussion

Subjective investigations, like HUT, are more and more relevant [1] and essential in cosmetic product evaluation and in cosmetic claim substantiation [2]. Product assessment via panelists generates results that can not be provided by biophysical devices [3]. They have been demonstrated to be very effective [1] and are essential in the maintenance, optimization and quality improvement of consumer goods [7]. When planning a HUT, the number of involved panelists is important because it is a key factor for costs, statistics and outcome: a decline in sample size leads to a decrease in costs and validity [16]. In other words, sample size affects the stability of the average and influences individual differences [5]. However, there is a wide range concerning the number of panelists and relevant recommendations vary greatly [5] [10] [13] [14] [15]. Therefore, scientists lack clear orientation from the literature as to how many consumers are adequate [10]. Concerning the field of cosmetic science, relevant publications, recommendations and guidelines dealing with this topic are rare and/or inexact. Cosmetics Europe [28], for example, recommends "the number of subjects/size of a study should always be large enough to provide a reliable answer to the questions addressed (*i.e.* have sufficient power). The number is usually determined by the primary objective of the study through a formal sample size calculation or by a justification based on statistical and/or methodological expertise (background data, former study, etc.)". Taking the situation of insufficient data into account, we performed a questionnaire-based product evaluation with three different cosmetic emulsions to generate clearer guidelines concerning sample size in HUT.

In total, 101 questionnaires were analyzed (89 f, 12 m; median age 42.0). The calculated minimum sample size ranged between n = 5 to n = 58 (95% CI) and n = 8 to n = 101 (99% CI). The present results are in the line with Moskowitz [5], who recommended a base size of 40 to 60 subjects. In addition, it has been recommended to involve approximately 50 to 100 panelists in laboratory tests and 50 to 100 in HUT [7]. In contrast, 75 to 300 panelists for HUT were suggested by Meillgaard *et al.* [1], which reflects a very broad range.Comparing the face, hand and body lotions showed that the body lotion needed the highest sample size (n = 58, 95% CI), which may be caused by different application areas, area size and application mode of a body lotion. One might conclude and postulate that products used for large skin areas, thereby, are unspecific in their application and require more panelists in a HUT. Moreover, it was demonstrat-

ed that assessment of the product scent ("… very pleasant smell") required the highest sample size in the three tested product types. Evaluating the scent of a product is mainly driven by hedonistic effects and, therefore, seems to be beneficial to include more rather than less panelists, *i.e.* minimum of 58 (Table 6). Regarding gender-related effects, a higher minimum sample size was required for the female group, especially for assessment of skin tolerance ("… well-tolerated"). Evaluation of the product absorbance depended on age, which could be interpreted as impact of skin dryness. Although "slightly or moderately dry skin" was one inclusion criteria, skin conditions of the subjects were not queried in detail—which could be helpful to analyze such data.

5. Conclusions

Efficacy claims on cosmetic products should be substantiated in accordance with regulation (EC) No. 1223/2009 and regulation (EU) No. 655/2013, whereby HUT could be used to verify such claims. Nevertheless, guidelines with detailed information or recommendations concerning the number of panelists are rare. We performed a questionnaire-based product evaluation with three different cosmetic emulsions as HUT, followed by a specific statistical analysis to answer the question: How many consumer panelists are necessary to get meaningful and informative data and results?

Based on the present study, and using a 95% CI (width one unit), the following recommendations are provided:

- The **impact of age** on cosmetic product evaluation is important. Statistical analyses of results regarding age groups could be helpful and provide further information.
- When addressing "skin tolerance", **gender-related effects** should not be ignored. It seems to be necessary to include more panelists (minimum 60), when such a study is planned with females (compared to a sheer male panel).
- Regarding **fragrance evaluation** items, more panelists (minimum 60) are required compared to other product assessment items.
- Evaluation of test products with **unspecific application mode/area** needs more panelists (minimum 60).
- It is important to specify the level of **skin dryness** as narrowly as possible, especially by assessing product absorbance behavior.

In summary, the minimum sample size depends not only on SD and CI, but also depends on the product assessment item as well as age and gender of the panelists. It is shown that a minimum of 60 (95% CI, width one unit) or a minimum of 101 (99% CI, width one unit) panelists seems to be sufficient for a similar questionnaire-based HUT with a given seven-point scale in cosmetic product evaluation. However, it is very important to know that the present conclusions and recommendations are based on the present constellation, e.g. test products, subjects and product assessment items, and, therefore, should be interpreted and applied carefully. Further research with different constellations is

required in order to define generally valid statements concerning the number of panelists in cosmetic product evaluation.

Acknowledgements

The study was initiated and sponsored by Kneipp GmbH. In addition, the test products were provided by Kneipp and are available in the market (Kneipp® Mandelblüten Hautzart). We thank Ms. Ann-Kathrin Haller for technical assistance. No further conflicts of interest are declared.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Meilgaard, M.C., Civille, G.V. and Carv, B.T. (1999) Sensory Evaluation Techniques. 4th Edition, CRC Press, Boca Raton. https://doi.org/10.1201/9781439832271

[2] Nobile, V. (2016) Guidelines on Cosmetic Efficacy Testing on Humans. Ethical, Technical, and Regulatory Requirements in the Main Cosmetics Markets. *Journal of Cosmetology & Trichology*, **2**, 1-10. https://doi.org/10.4172/2471-9323.1000107

[3] Mead, R. and Gay, C. (1995) Sequential Design of Sensory Trials. *Food Quality and Preference*, **6**, 271-280. https://doi.org/10.1016/0950-3293(95)00029-1

[4] Moussaoui, K.A. and Varela, P. (2010) Exploring Consumer Product Profiling Techniques and Their Linkage to a Quantitative Descriptive Analysis. *Food Quality and Preference*, **21**, 1088-1099. https://doi.org/10.1016/j.foodqual.2010.09.005

[5] Moskowitz, H.R. (1996) Base Size in Product Testing: A Psychophysical Viewpoint and Analysis. *Food Quality and Preference*, **8**, 247-255. https://doi.org/10.1016/S0950-3293(97)00003-7

[6] Lawless, H.T. and Heymann, H. (2010) Sensory Evaluation of Food. Principles and Practices. 2nd Edition, Springer Verlag, New York. https://doi.org/10.1007/978-1-4419-6488-5

[7] Stone, H., Sidel, J.L. and Taylor, S. (1993) Sensory Evaluation Practice. 2nd Edition, Academic Press, New York.

[8] Scriven, F. (2005) Issues and Viewpoints—Two Types of Sensory Panels or Are There More? *Journal of Sensory Studies*, **20**, 526-538. https://doi.org/10.1111/j.1745-459X.2005.00044.x

[9] Lodén, M. (2000) Efficacy Testing of Cosmetics and Other Topical Products. *IFSCC Magazine*, **3**, 47-64.

[10] Hough, G., Wakeling, I., Mucci, A., Chambers IV, E., Gallardo, I.M. and Alves, L.R. (2006) Number of Consumers Necessary for Sensory Acceptability Tests. *Food Quality and Preference*, **17**, 522-526. https://doi.org/10.1016/j.foodqual.2005.07.002

[11] Vidal, L., Cadena, R.S., Antùnez, L., Giménez, A., Varela, P. and Ares, G. (2014) Stability of Sample Configurations from Projective Mapping: How Many Consumers Are Necessary? *Food Quality and Preference*, **34**, 79-87. https://doi.org/10.1016/j.foodqual.2013.12.006

[12] Valentin, D., Chollet, S., Lelièvre, M. and Abdi, H. (2012) Quick and Dirty but Still Pretty Good: A Review of New Descriptive Methods in Food Science. *International*

Journal of Food Science & Technology, **47**, 1563-1578.
https://doi.org/10.1111/j.1365-2621.2012.03022.x

[13] Gacula, M.C. and Singh, J. (1984) Statistical Methods in Food and Consumer Research. Academic Press, Orlando.

[14] Lawless, H.T. and Heymann, H. (1998) Sensory Evaluation of Food. Principles and Practices. Chapman and Hall, New York.

[15] Gacula, M.C. (1993) Design and Analysis of Sensory Optimization. Wiley-Blackwell, Trumbull.

[16] Moskowitz, H.R. (2008) Chapter 14: Sample Size N, or Number of Respondents in Viewpoints and Controversies in Sensory Science and Consumer Product Testing. In: Moskowitz, H.R., Muñoz, A.M. and Gacula, M.C., Eds., *Viewpoints and Controversies in Sensory Science and Consumer Product Testing*, John Wiley & Sons, Trumbull, 241-254.

[17] Kraemer, H.C. and Thiemann, S. (1987) How Many Subjects? Statistical Power Analysis in Research. Sage Publications, Newbury Park.

[18] Gacula, M. and Rutenbeck, S. (2006) Sample Size in Consumer Test and Descriptive Analysis. *Journal of Sensory Studies*, **21**,129-145.
https://doi.org/10.1111/j.1745-459X.2006.00055.x

[19] Mammasse, N. and Schlich, P. (2014) Adequate Number of Consumers in a Liking Test. Insights from Resampling in Seven Studies. *Food Quality and Preference*, **31**, 124-128. https://doi.org/10.1016/j.foodqual.2012.01.009

[20] Basker, D. (1997) The Number of Assessors Required for Taste Panels. *Chemical Senses*, **2**, 493-496. https://doi.org/10.1093/chemse/2.4.493

[21] Meiselman, H.L. and Schutz, H.G. (2002) History of Food Acceptance Research in the US Army. *Appetite*, **40**, 199-216.
https://doi.org/10.1016/S0195-6663(03)00007-2

[22] O'Donnel, J. (2011) Sensory and Stability Testing. *IFSCC Magazine*, **14**, 217-221.

[23] Montgomery, D.C. (1991) Design and Analysis of Experiments. 3rd Edition, John Wiley and Sons, New York.

[24] Gacula, M. and Singh, J. (2002) A Review of Strategy for the Estimation of Sample Size in Research Investigation. *Journal of Sensory Studies*, **17**, 583-598.
https://doi.org/10.1111/j.1745-459X.2002.tb00367.x

[25] Faul, F., Erdfelder, E., Lang, A.-G. and Buchner, A. (2007) G*Power: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomatical Science. *Behavior Research Methods*, **29**, 175-191. https://doi.org/10.3758/BF03193146

[26] Silva, A.N., Silva, R.D.C.S.N.D., Ferreira, M.A.M., Minim, V.P.R., Costa, T.D.M.T.D. and Perez, R. (2013) Performance of Hedonic Scales in Sensory Acceptability of Strawberry Yogurt. *Food Quality and Preference*, **30**, 9-21.
https://doi.org/10.1016/j.foodqual.2013.04.001

[27] Preston, C.C. and Colman, A.M. (2000) Optimal Number of Response Categories in Rating Scales: Reliability, Validity, Discriminating Power, and Respondent Preferences. *Acta Psychologica*, **104**, 1-15.
https://doi.org/10.1016/S0001-6918(99)00050-5

[28] The European Cosmetics Association (2008) Cosmetics Europe: Guidelines for the Evaluation of the Efficacy of Cosmetic Products, Brussels. Rev. Efficacy Evaluation Guidelines.