Received: 29 January 2025 | Accepted: 4 June 2025

DOI: 10.1111/ics.70000

International Journal of Cosmetic Science
SOCIÉTÉ FRANÇAISE DE COSMÉTOLOGIE

ORIGINAL ARTICLE

The variability of in vivo sunscreen sun protection factor values

Curtis Cole[1] | **Bertrand Colson**[2] | **Steffen Uhlig**[2]

[1]Sun & Skin Consulting LLC, New Holland, New Holland, Pennsylvania, USA

[2]QuoData GmbH, Dresden, Germany

Correspondence
Curtis Cole, Sun & Skin Consulting LLC, New Holland, PA 17557, USA.
Email: curtcolephd@comcast.net

Abstract

Objective: Determination of the sunburn protection provided by a sunscreen product is required globally for sales of these products. Over the past 80 years, many aspects of determining the protection 'factor' have evolved and been modified, with varying levels of impact on the sunburn protection factor (SPF) value. In order to compare new non-invasive and in vitro methods against traditional SPF test protocols, a large, multi-center clinical trial was conducted to establish the level of equivalence of these new methods with the current codified testing standard ISO24444 ([1]: Cosmetics – sun protection methods – in vivo determination of the sun protection factor (SPF), 2019). This report reports the variability found in the in vivo determination of SPF values.

Methods: Thirty-two products of varying levels of sunburn and UVA protection, in a variety of formulation vehicles and ultraviolet (UV) filter combinations and concentrations, were coded and sent to pre-qualified in vivo SPF testing laboratories. The products were divided into eight product groups (four products per product group). For each product group, samples were sent to four laboratories (across product groups, a total of 12 laboratories participated). Precision and trueness estimates were calculated separately for each product group. 'Expected' SPF values were not provided to the test laboratories. However, laboratories were informed as to whether the 'true' SPF was less than or greater than 25.

Results: Interlaboratory variability for samples was proportional to the SPF of the products, with high SPF products having higher variability. Intra-laboratory variability (repeatability) was much lower than the interlaboratory variability.

Conclusions: This study highlights the fact that the interlaboratory variability of SPF results is considerable and is likely greater than expected by the public and regulatory bodies.

Résumé

Objectif: La détermination de la protection contre les coups de soleil offerte par un produit de protection solaire est requise à l'échelle mondiale pour la vente de ces produits. Au cours des 80 dernières années, de nombreux aspects de la

détermination du «facteur» de protection ont évolué et ont été modifiés, avec des niveaux d'impact variables sur la valeur du facteur de protection solaire (SPF). Afin de comparer les nouvelles méthodes non invasives et in vitro aux protocoles de test SPF traditionnels, un essai clinique multicentrique de grande envergure a été mené pour établir le niveau d'équivalence de ces nouvelles méthodes avec la norme de test codifiée actuelle ISO24444. Ce rapport rend compte de la variabilité observée dans la détermination *in vivo* des valeurs SPF.

Méthodes: 32 produits offrant différents niveaux de protection contre les coups de soleil et les rayons UVA, dans diverses formulations et combinaisons de filtres ultraviolets (UV) et concentrations, ont été codés et envoyés à des laboratoires de tests SPF in vivo préqualifiés. Les produits ont été divisés en 8 groupes de produits (4 produits par groupe). Pour chaque groupe de produits, des échantillons ont été envoyés à 4 laboratoires. (Au total, 12 laboratoires ont participé, tous groupes confondus.) Les estimations de la précision et de la justesse ont été calculées séparément pour chaque groupe de produits. Les valeurs SPF « attendues » n'ont pas été communiquées aux laboratoires de test. Cependant, les laboratoires ont été informés que le SPF « réel » était inférieur ou supérieur à 25.

Résultats: la variabilité inter-laboratoires des échantillons était proportionnelle à l'indice SPF des produits, les produits à indice SPF élevé présentant une variabilité plus importante. La variabilité intra-laboratoires (répétabilité) était beaucoup plus faible que la variabilité inter-laboratoires.

Conclusions: Cette étude souligne le fait que la variabilité inter-laboratoires des résultats SPF est considérable et probablement supérieure à celle attendue par le public et les organismes de réglementation.

INTRODUCTION

Topically applied sun protection products come in a wide variety of delivery forms (emulsions, sprays, sticks, and even powders) and typically contain a mixture of multiple ultraviolet radiation blocking filters at varying concentrations. Consumers and dermatologists need to know the level of protection that can expected from a given product to properly choose a product appropriate for their skin phototype (sunburn sensitivity) and the local sun exposure conditions (location, time of day, time of year, etc.) and expected time in sunlight. The sun protection factor is a fundamental measure of the 'potency' of the protection provided by a product when applied under controlled conditions and tested by a defined light source. While the 'factor' implies a time element of protection (SPF 15 implies 15 times longer exposure in the sun compared to no sunscreen protection before becoming sunburned), it is misleading. The best use of the SPF number is for relative comparison of one product versus another, and not for absolute computation of 'safe exposure time' as the sun's sunburning capacity changes by the minute, and the application doses are not controlled in consumer use. Many studies have been conducted and show that the average consumer uses ¼ to ½ of the application dose compared to laboratory SPF testing dosages [1–4] so that absolute calculations of 'safe exposure time' cannot be relied upon. However, it should be possible to assume that a SPF 30 product provides twice the protection compared with an SPF 15 product as sunscreen products are currently tested according to highly prescriptive codified protocols and international testing standards (i.e. FDA Sunscreen monograph, ISO24444:2019). Unfortunately, many challenges of product SPF claims for specific products and manufacturers have arisen in past years, often by consumer safety and advocate agencies that conduct independent testing of products, or by regulatory authorities. Discrepancies between SPF results coming from different laboratories have become more apparent to manufacturers as well, making development of products difficult and time consuming.

Many of these concerns regarding reliability of SPF values were prophetically raised in what appears to be the first comprehensive report on SPF testing written by Harold Blum and 1945 [5]. Blum discusses many of the sources of variability in testing results and concludes:

> The actual evaluation of the protection afforded by a given sunburn preventive, under controlled laboratory conditions, is beset with difficulty and great accuracy is not to be expected. Even with the best laboratory measurements, it is difficult to estimate in more than a general way, the appropriateness of the protection afforded by a given sunburn preventive to the need of a particular condition of exposure to sunlight... All these factors permit claims to be made which, while not actually false, may be quite misleading to the user of a sunburn preventive... Wide differences in the estimation of the protection afforded by sunburn preventives are obtained by different measures and may lead to erroneous conclusions. Individual estimates of the value of preventatives may be widely divergent.

SPF testing protocols have been constantly modified over the past 70 years as regulated by local regulatory authorities and as equipment and procedures have improved. The underlying principles for the determination of the SPF value have remained the same; however, the definitions, procedures, and equipment have changed significantly from the first tests by Blum and others. The first codified protocol was published by the US-FDA in 1978 [6], which outlined testing procedures including the use of either a xenon arc solar simulator or natural sunlight as the test light source. In the 1993 FDA Tentative Final Monograph [7], the possibility to test with natural sunlight was dropped in favour of testing with a solar simulator under laboratory conditions to reduce variability due to the everchanging sun spectrum and intensity. The German DIN (Deutsches Institut für Normung) [8] published in the mid-1980s prescribed the use of a metal halide lamp (Osram Ultravitalux) lamp as the light source, which had a discontinuous spectrum quite different from a xenon solar simulator or sunlight. The European Cosmetics Industry Consortium COLIPA published their SPF testing procedure [9] in 1995, prescribing a xenon arc lamp for the test light source (similar to the US-FDA solar simulator); however, the definitions used for the erythema response and their interpretation were quite different from the US-FDA procedures, and considerable variance between SPF values for European and US products was noted. In an effort to harmonize the many local SPF protocols at the turn of the century, representatives from the European Union, Japan, South Africa and the United States collaborated to produce the 2006 International Harmonized SPF Protocol [10], which was utilized as a primary source for the development of the first ISO24444:2010 [11] SPF Standard for in vivo SPF determination, recognized by nearly every country (notably not the US). Global harmonization using an ISO standard provides the basis for manufacturers to conduct one clinical trial for a product without having to repeat such testing in every country into which it is imported, thus facilitating global trade. Through each successive publication, modifications were made to help improve the repeatability and accuracy of the SPF value.

During the next 5 years, manufacturers and testing laboratories realized that further modifications were still needed to diminish the interlaboratory variability that was still evident. The updated ISO24444:2019 [12] version was drafted with the intention to address areas in the protocol that contributed to this variability. These areas included definitions of the erythema response, prescribing the same UV exposure dose range for the determination of the minimum unprotected ultraviolet (UV) dose causing an erythema response of an individual (Minimal Erythema Dose, denoted MED_u), procedures for application of sunscreens, and providing methods and limits for the uniformity of the UV light source. New Reference Standard sunscreens at higher SPF levels were tested in ring studies to establish a control product within each test range, making the test an equivalency test to validate the product SPF value. Calibration and validation procedures were put in place to eliminate as many variables as possible; nevertheless, sources of variability remain that cannot be as easily addressed. These will be discussed later in this paper. The ALT-SPF Consortium test for the development of alternative test methods of SPF determination was the first controlled ring study conducted evaluating such a wide variety of sunscreen product types and SPF levels and gives a revealing data set to demonstrate the state-of-the art of sunscreen SPF testing across the globe and highlights the level of variability that exists today using traditional in vivo SPF procedures. The ISO24444:2019 [12] in vivo SPF test remains the 'gold standard' and is the measure against which other test methods must be measured for validation and replacement of this 'standard'. Alternative methods suffer from many of the same sources of variability, and some with added sources, such that validation of alternative methods has been a daunting challenge for many years. The use of human subjects for SPF testing has met with increasing challenges due to the known carcinogenic effects of UV light, such that the need to find and validate alternative methods has become imperative. Validating one variable method against another variable method requires careful considerations and statistical analysis to achieve a final method that provides assurance of the reliability of the products being tested. This paper focuses on the magnitude of variability of the 'gold standard' method and the sources thereof.

METHODS

Laboratories

Twelve laboratories were recommended by ISO TC217-WG7 members for in vivo SPF testing of the test products based on their experience using these laboratories. Four independent sunscreen testing experts from WG7 were chosen to audit the protocols, training and testing records, and validations of these laboratories and their equipment. They also conducted online laboratory inspections of procedures against ISO24444:2019 [12] specifications. Deviations were noted and communicated to the laboratory personnel for correction before participation in the study. Previous experience utilizing ISO24444:2019 [12] was required of all the laboratories. Laboratories participating in the study were located in Australia, Brazil, France, Germany, Korea, Romania and Spain.

Test products

Thirty-two products were selected for testing, representing different product types, UV filter types, viscosities and SPF levels. The selection was recommended by ISO WG7 members, consistent with the majority of products available on the global market. These were organized into eight product groups, with four products within each group having similar product characteristics (viscosity, SPF and filter types), see Table 1. Replicate samples of each product were prepared with coded labels so that the test laboratories were unaware of the identity of the product SPF. Each laboratory tested only four products (in blinded replicate), with four labs providing data (in blinded replicate) for each test product.

Design

Blinded replicate samples were assigned to a separate laboratory technician and erythema grader in order to assess intra-laboratory variability. Unlike typical in vivo test procedures, no specific 'target' SPF values were given for each sample. The 'target' SPF is the 'expected' SPF of the test product typically provided by a manufacturer that is used to set the range of exposures used for the dose exposures ('Target SPF' times the unprotected MED) with the 'target' SPF set in the middle (3rd or 4th) exposure subsite within a test site. Use of a 'target SPF' has the potential to bias the expectations of the erythema grader, so the SPF target was omitted for this Reference SPF determination. The only guidance provided was whether a particular test product was greater than or less than 25, in order to minimize unnecessary testing. Laboratories thus could not make assumptions of the SPF result they should expect but were forced to start at a low exposure dose range testing few subjects, and then increase the exposure doses on additional subjects until they started to see erythema responses within the exposure ranges.

Procedures

SPF testing was conducted according to ISO24444:2019 [12]. Because of the volume of products to be tested, only five subjects in each of the test laboratories were tested using one laboratory technician and erythema grader combination. The results from this pair of results were combined with five subjects using the second laboratory and erythema grader combination from that laboratory for $n = 10$ for each test sample. Subjects were not added to achieve the required statistical acceptance criterion described in ISO24444:2019 [12] section 10.3 (Confidence Interval < 17% of the mean).

TABLE 1 Products were grouped according to vehicle type, SPF and UV filter type.

Product group	SPF category	Vehicle description	UV filter type
1	30	Low viscosity emulsion[a]	Organic and Organic/Inorganic
2	6	Medium viscosity emulsion	Organic
3	16	Medium viscosity emulsion	Organic
4	30	Medium viscosity emulsion	Organic and Organic/Inorganic
5	60	Medium viscosity emulsion	Organic and Organic/Inorganic
6	30	High viscosity emulsion	Organic
7	60	Single phase liquid	Organic
8	60	Medium viscosity emulsion	Inorganic (only)

Note: Organic/Inorganic products contained less than 6% Inorganic filters by weight. Inorganic (only) formulae contained only ZnO and TiO_2 as UV filters.

[a]Low viscosity: 3000–4000 cps; Medium viscosity: 4100–29 000 cps; High viscosity: 30 000-40 000 cps.

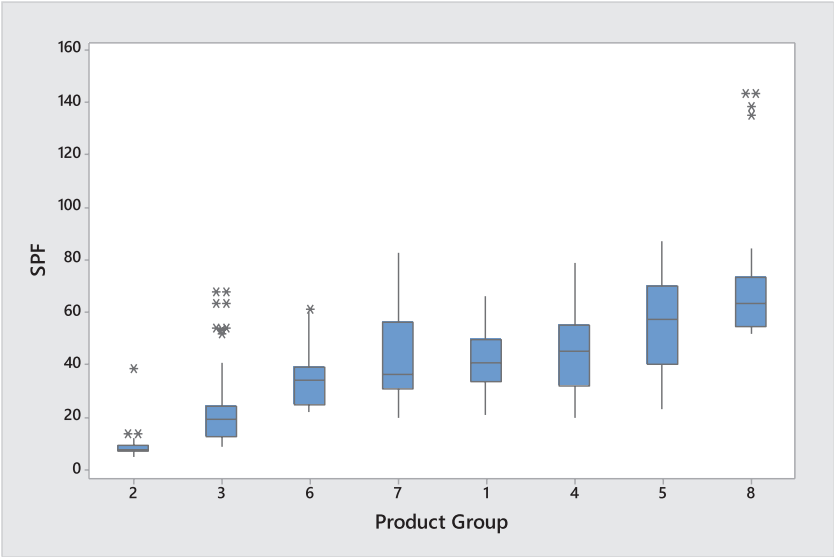

FIGURE 1 Variability of SPF increases with SPF value. '*' represent outlier values.

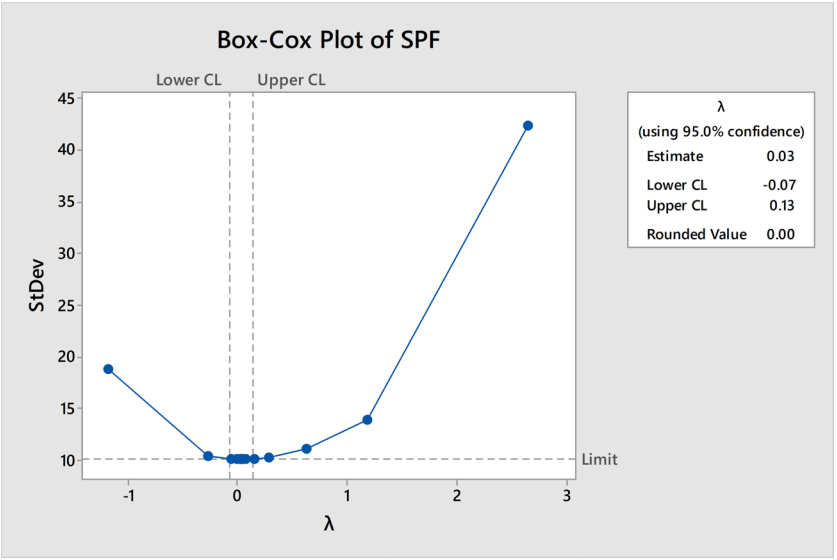

FIGURE 2 Box-Cox plot indicating the data mean SPF value should be log-transformed.

RESULTS

Examination of the data showed the heteroskedastic nature of the SPF results (not unexpected based on previous observations [13–15]; see Figure 1). This simply indicates that the variability of the results increases as the value of the response increases (higher SPF values are more variable than lower SPF values). A Box-Cox plot of the SPF results yielded a λ of −0.12 indicating that a log transformation of the values is appropriate to normalize the distributions as shown in Figure 2.

Figure 3 displays the box plots of the SPF of the various products in each product group. Note that the 'box' area represents 50% of the values reported, while the 'whiskers' above and below the boxed area represent the upper and lower 25% of the reported values. Outliers are shown as the '*'s above or below the whiskers.

The factorial design of the study permits evaluation of the sources and magnitude of the variability across laboratories. As described in Colson et al. [16], for any given product group, four laboratories[1] performed the SPF testing, with different combinations of product appliers and graders (of the erythema responses) involved in the determination of two SPF values for each product[2]. Estimates for the different components of variability were obtained via application of (mixed) linear model to the log-transformed[3] test results. These estimates (expressed as standard deviations) are provided in Table 2 below. The log-domain standard deviations

[1]While a total of 12 different test laboratories were involved for the 8 product groups, for any given product group, only 4 laboratories performed testing.
[2]This design was applied twice, so that a total of 4 SPF test results were obtained per product.
[3]Natural logarithm.

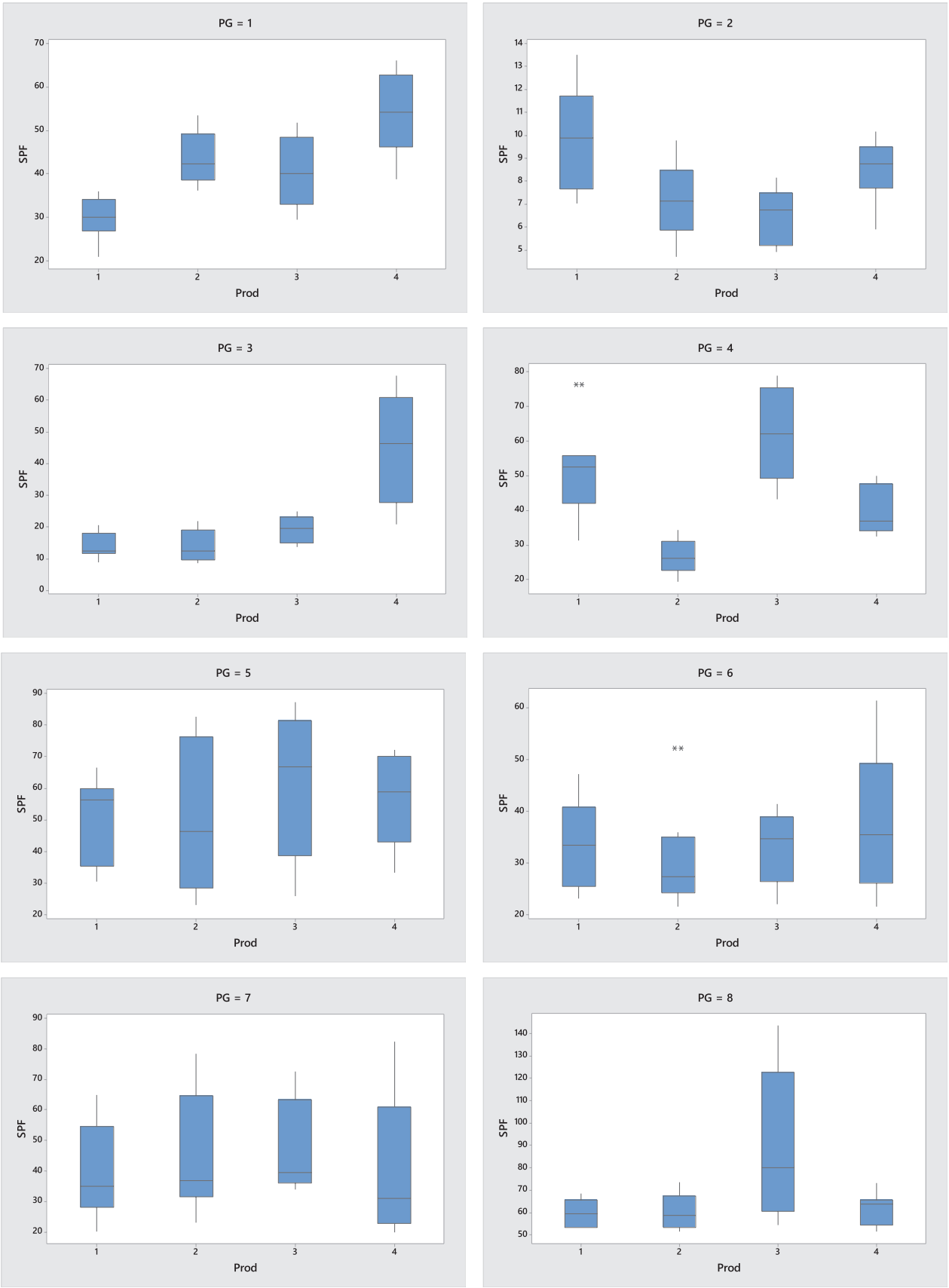

FIGURE 3 Reported SPF values for each of the 4 products within each of the product groups. (See Table 1 for product group descriptions.)

TABLE 2 Precision estimates (expressed as standard deviations) for the in vivo ISO 24444 SPF results.

Product group	Number of		Precision [ln SPF]							
	Results	Labs	s_R	s_L	$s_{L,pers}$	$s_{L•Prod.}$	$s_{L:F1}$	$s_{L:F2}$	$s_{L:F3}$	s_r
1	32	4	0.18	0.07	0.10	0.15	0.06	0.02	0.02	0.04
2	32	4	0.22	0.00	0.15	0.14	0.14	0.03	0.04	0.05
3	32	4	0.35	0.26	0.28	0.21	0.09	0.01	0.05	0.02
4	32	4	0.20	0.10	0.15	0.09	0.09	0.00	0.06	0.09
5	32	4	0.41	0.37	0.38	0.15	0.08	0.00	0.03	0.04
6	32	4	0.29	0.26	0.27	0.05	0.10	0.00	0.00	0.08
7	32	4	0.44	0.42	0.42	0.12	0.00	0.05	0.01	0.04
8	32	4	0.22	0.14	0.14	0.17	0.03	0.00	0.01	0.03

Abbreviations: s_R, reproducibility; s_L, laboratory; s_{Lpers}, persistent laboratory; $s_{Lpers*Prod.}$, laboratory-product interaction; $s_{L:F1}$, factor 1 (panel); $s_{L:F2}$, factor 2 (product applicator); $s_{L:F3}$, factor 3 (grader); s_r, repeatability.

TABLE 3 Repeatability and reproducibility limits for ISO 24444 SPF results.

Method	Product group	Repeatability [ln SPF]		Reproducibility [ln SPF]	
		s_r	r	s_R	R
ISO 24444 [ln SPF]	1	0.04	0.11	0.18	0.5
	2	0.05	0.14	0.22	0.62
	3	0.02	0.06	0.35	0.98
	4	0.09	0.25	0.2	0.56
	5	0.04	0.11	0.41	1.15
	6	0.08	0.22	0.29	0.81
	7	0.04	0.11	0.44	1.23
	8	0.03	0.08	0.22	0.42

Abbreviations: s_r, repeatability standard deviation within a laboratory; r, the repeatability limit (for two test results obtained within one and the same laboratory, under near identical conditions); s_R, reproducibility standard deviation (including between-laboratory effects); R, reproducibility limit (for two test results obtained in different laboratories).

can be interpreted as if they were relative standard deviations in the original SPF domain. That is to say, 0.1 ln SPF is approximately equivalent to 10% SPF, 0.2 ln SPF is approximately 20% SPF, etc. This approximation becomes less accurate as the log value increases. It should be noted that the back-transformed standard deviations are slightly asymmetrical around the geometric mean value.

The reproducibility s_R represents the combination of all the sources of variation: laboratory, lab-product interaction, factorial and repeatability effects. The persistent laboratory variation is the sum of the laboratory effect per se and the factorial effects and thus characterizes laboratory effects that are consistent across products. By contrast, the lab-product interaction effects show laboratory deviations that change from product to product. The results show that both persistent laboratory effects and lab-product interaction effects contribute to overall variability (see Table 3). The contributions to variability from applicator or grader are very small/negligible, and repeatability effects *within a laboratory* are also small. This does not, however, mean that application and grading practices *between* the laboratories are the same. The variability between laboratories is considerable and is easily visualized by inspection of the graphs in Figure 3. Differences in application and grading practices remain the most likely source of laboratory effects per se and lab-product interaction effects.

In the ALT-SPF study, the four laboratories differed from product group to product group (see Figure 4 for an overview of which groups of four laboratories tested which product groups). For this reason, the question arises whether disparities in reproducibility between the product groups are caused by the different groups of four laboratories. To be specific: the question is whether the relatively high reproducibility values in Product Group 5 and Product Group 7 (0.41 ln SPF and 0.44 ln SPF, respectively) are caused by the particular combinations of four laboratories rather than the poor performance of the ISO 24444 method for these two more challenging types of products.

In order to determine to what extent laboratory effects are responsible for the significant differences in reproducibility between product groups, it is advisable to examine the systematic deviations of the laboratories for each individual product. This is best achieved via z-scores, that is, standardized deviations. The following figure provides an overview of z-scores across all laboratories and products. As can be seen, no laboratory displays a consistent conspicuous bias across all the products. Rather, it is evident that the systematic effects are strongly associated with the product groups. For example, L02 and L10, which show a significant bias in PG7, are entirely inconspicuous in other product groups, such as PG8. Similarly, albeit to a lesser

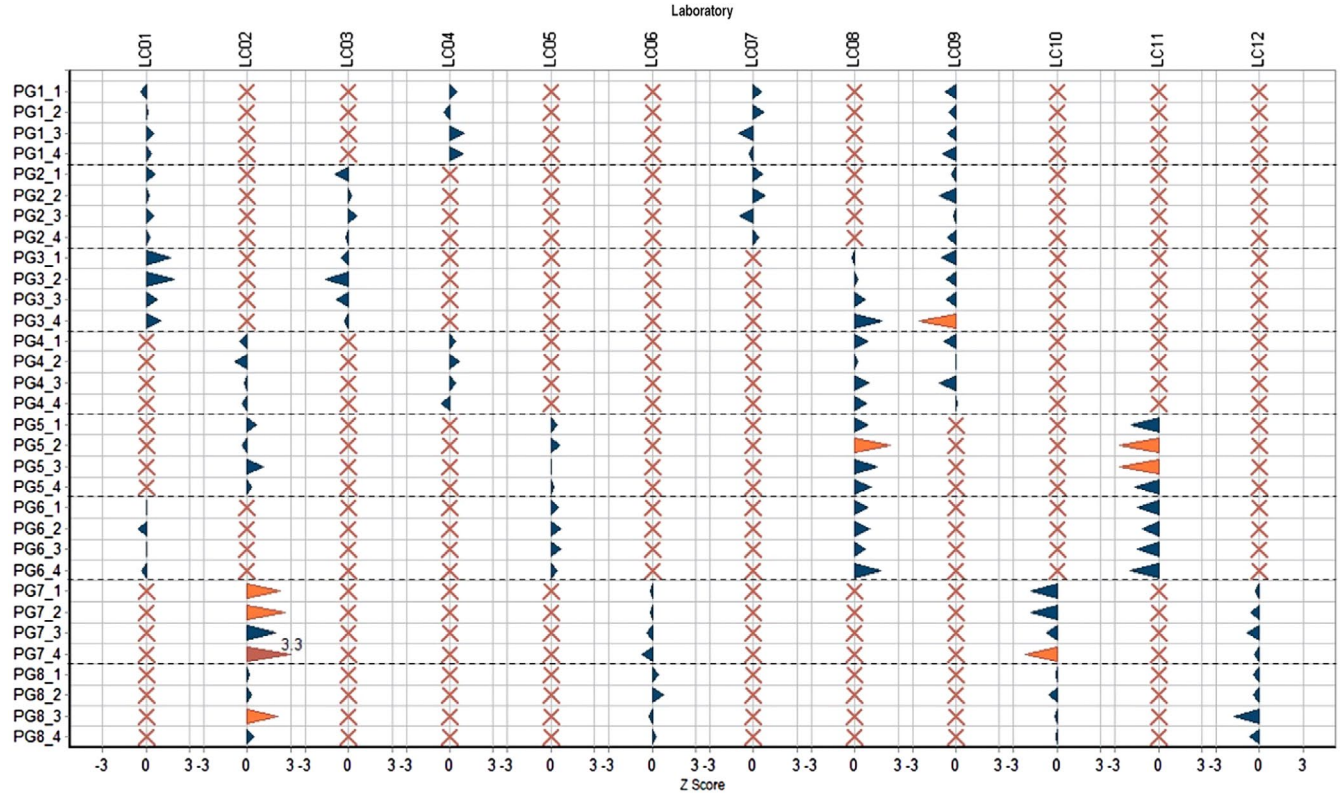

FIGURE 4 *Z*-scores calculated on the basis of consensus mean values with SDPA = robust mean reproducibility estimate of 0.24 ln SPF. Scores whose absolute value lies between 2 and 3 are displayed as orange triangles. Scores whose absolute value is >3 are displayed as red triangles (there is only one such score: $Z = 3.3$ in PG7, product 4).

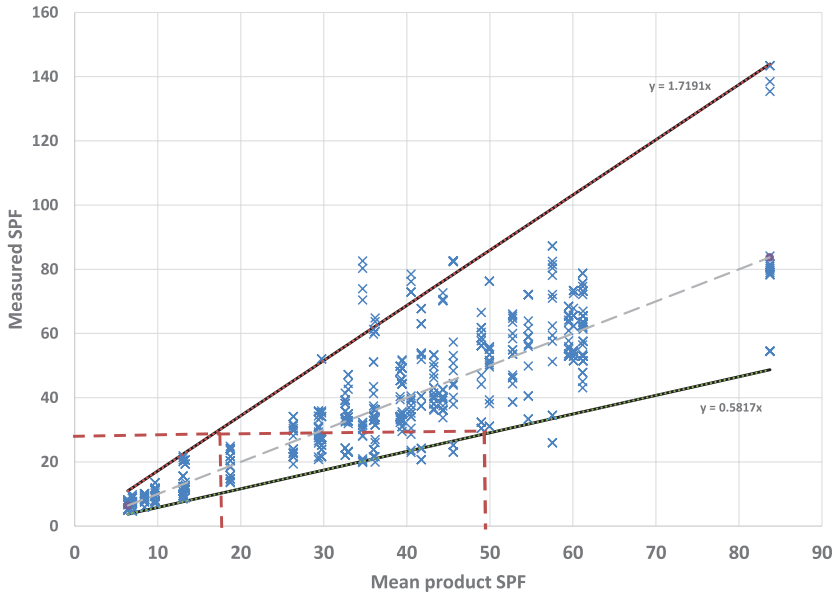

FIGURE 5 Reported in vivo SPF results for the 32 products plotted as a function of the Product Group averages (*x*-axis) along with a 95% Prediction Interval. The dashed red lines represent the expected 'true' SPF range (*x*-axis) starting from a measured SPF 30 test result (*y*-axis).

extent, this can be observed for L08 and L11 in PG5, as well as for L01 and L09 in PG3. This suggests that matrix/formulation effects at least partly explain the differences in observed product group-specific s_R values.

Reported in vivo SPF results (512 values) for the 32 products were plotted as a function of the Product Group averages (*x*-axis) as shown in Figure 5. The upper and lower limits of the prediction range were calculated in such a way that 95% of the data points lie within the prediction range. The calculation is based on an 'average' reproducibility standard deviation value across all product groups (in the log domain). This average value was

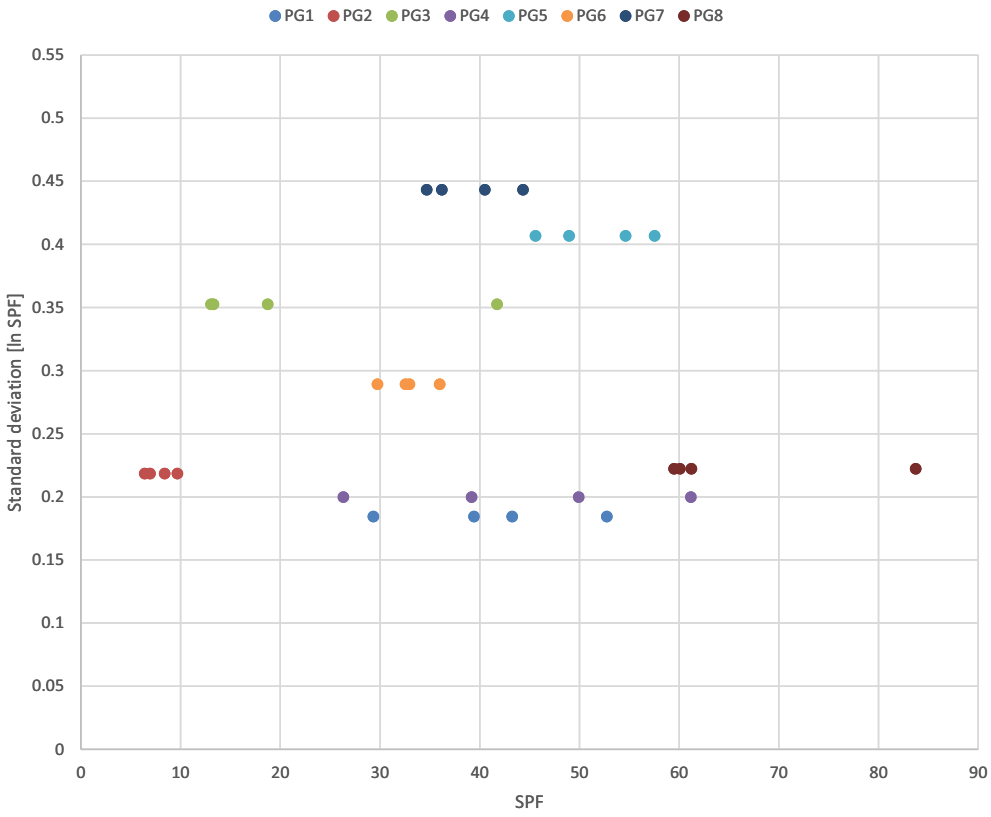

FIGURE 6 Reproducibility estimates (standard deviations) plotted against the product-specific geometric means and colour-coded for product group.

estimated as 0.27 ln SPF. It should be noted that, in the ALT-SPF study, reproducibility varied considerably between product groups (from 0.18 ln SPF for Product Group 1 to 0.44 ln SPF for Product Group 7). Accordingly, the average value of 0.27 ln SPF must be interpreted as a somewhat artificial value whose sole purpose is to *conveniently summarize* ISO 24444 variation *across all matrices and UV filters*. Nonetheless, this approximation can be considered predictive in the sense that for test results obtained from products representing all 8 product groups, 95% can be expected to lie within the prediction range [3].

Of course, in practice, the 'true' SPF value is unknown. In such a scenario, the prediction range can be used as follows. The interval obtained by intersecting the horizontal line starting from the measured SPF value (y-axis) with the upper and lower prediction lines can be considered to represent where the 'true' SPF value can be expected to lie.

For comparison, Figure 6 shows the s_R estimates (standard deviations calculated from the log-transformed SPF test results) plotted against the product-specific mean SPF values.

Finally, we turn to the question of the expected differences between two test results obtained for one and the same sunscreen product. The repeatability and reproducibility limits described in section 4 of ISO 5725-6 [17] provide useful orientation regarding acceptable differences between two such test results. Each limit is obtained by multiplying the standard deviation by the factor 2.8:

$$r = \text{repeatability limit} = 2.8 \cdot s_r,$$

$$R = \text{Reproducibility limit} = 2.8 \cdot s_R.$$

These limits have the following meaning. The difference (absolute value) between two test results obtained in the same laboratory under repeatability (i.e. near identical) conditions can be expected to be $<r$. The difference (absolute value) between two test results obtained in different laboratories can be expected to be $<R$.

For two test results obtained under (near-)identical testing conditions within one laboratory (repeatability conditions), the expected difference can be derived from the repeatability standard deviation.

In order to illustrate how these limits can be used in practice, consider the case that two SPF test results x_1 and x_2 are obtained in different testing institutes for a product corresponding to Product Group 1. The absolute difference of the log-transformed values is then compared to $R = 0.50$ ln SPF. For instance, for $x_1 = 30$ SPF, we have $\ln(x_1) = 3.4$ ln SPF and the acceptable range for $\ln(x_2)$ is thus 2.90 to 3.90 ln SPF, corresponding to the SPF interval [18.1, 49.5] in the raw scale. It should be noted that in the SPF domain, this interval is not symmetric around x_1.

DISCUSSION

SPF values have increased over the past 50 years due to improved formulation techniques, increased acceptance, consumer demand, and medical endorsement. The original US FDA monograph label SPF scheme listed products with SPF 8 to SPF 15 as 'Maximal' protection, with the category of 'Ultra' protection reserved for products with SPF 15 and higher. Product SPF label claims now top SPF 100 in some countries, while many countries have restricted SPF claims for products with tested SPF results above 60 to '60+'. One paper published in 2002 claimed that high SPF (SPF 30–40) sunscreens can be tested with accuracy and reproducibility at different test sites [18]. Another paper reported validation of very high sun protection values (SPF 70–85) by demonstrating their ability to statistically distinguish between an SPF 70 and an SPF 90 product in four different laboratories [19]. While it is possible to show low variability between testing results from different laboratories, these examples can be misleading. Careful selection of the test laboratories based on previous testing results and providing 'target' SPF results can minimize the apparent variability. In both of these cases, only US laboratories testing according to the FDA sunscreen monograph were utilized.

An alternative examination of the variability in testing results was provided by Miksa et al. [20], where the conclusion was given recommending the determination of SPF based on testing of at least 3 (and ideally 4) laboratories to reduce the consumer health risk by ensuring the reliability of the SPF claim. The testing conducted in this testing was also based on a 'target SPF' provided to the test laboratory in advance, potentially improving the test results (yielding less variability). Other authors have noted that the application amount was a critical factor in the SPF value [21], and so this variable is tightly regulated within ±5% using the ISO 24444:2019 protocol. Zago et al. [22] observed that the Coefficient of Variance (σ/μ%) of SPF values from the Bureau Interprofessionnel d'Etudes Analytiques (BIPEA) performance testing ranged from as low as 10% to as much as 50% (36 laboratories testing a variety of sunscreen products and formats over a 17-year time span).

The discrepancies in SPF values between product labels and testing results by third party organizations (consumer advocates, regulatory authorities or competitors) has led to a host of litigations, contested advertising claims and consumer complaints. Manufacturers struggle in developing products based on varying SPF testing results and are faced with testing products in multiple laboratories and then determining which of the results is appropriate for the label claim.

Examination of the SPF box plots in Figure 3 shows the magnitude of variability of in vivo SPF testing, and probably exceeds the level or variability expected by regulators and consumers. It is likely that public perception of the accuracy of SPF is within a few units of SPF, and not 10's of units of SPF as shown here.

What can we learn from these test results regarding the sources of variability that can be used to improve the test method(s)? The results show that contributions to overall variability of the intra-laboratory influences—such as sunscreen applier and the erythema grader—are relatively small. For a given product, replicate test results within a laboratory yield similar SPF values independently of the different technicians who actually perform the test.

The greatest variability is *between* laboratories where training and performance practices appear to be different. There may also be a geographical influence as laboratories used for this testing were located across the globe and on both sides of the equator. While the selection of subjects by their skin darkness (Individual Typology Angle or ITA°) has been partially regulated by ISO24444:2019, it does not account for potential genetic differences in erythema responses from UV radiation [23].

From an optical point of view, the highest source of variability is the uniformity and absolute thickness of the film of UV filters that is spread onto the surface of the skin. The amount used during testing is 1.95 to 2.05 mg/cm^2 as determined by weigh-back calculation of the apparatus and the finger cots used during application. Uniformity of application is determined by examination by Wood's Lamp; however, this technique is highly subjective and non-quantitative and may only provide a very superficial view of the actual uniformity of the sunscreen application. Considering that the film being deposited onto the skin is only 10 micrometres (10^{-6}) thick for a typical emulsion product with 50% volatile carrier (water—even thinner for an alcoholic spray), any minor deviations will have a major impact on the actual transmission values of the UV radiation. To put this in perspective, 10 micrometres is 1/10th the thickness of a human hair, and the sunscreen in this ultra-thin layer is expected to block 90% to 99% of the UV radiation impinging upon it. This variability in application uniformity remains the highest sources of variability and is also the hardest variable to address. In vitro alternative methods have resorted to use of a programmed robot finger to do the application to PMMA plates as the best solution to reduce errors from application non-uniformity.

Another source of variability that is difficult to address is the subjective scoring of erythema responses that is core to the test method conclusions. Part of the confusion in grading differences comes from the original definitions and nomenclature used in describing the qualifying

erythema responses. The original definition in the 1978 FDA [2] sunscreen monograph test described the qualifying skin response for the test as 'a *minimal perceptible erythema*' of a subjects skin—with the lowest dose for this response taken to be the Minimal Erythema Dose (MED). This phrasing suggested that the response is supposed to be a somewhat questionable response—a 'minimally perceptible' erythema. This definition was changed however in the 1999 tentative final monograph [24] to lowest dose causing '*redness* reaching the borders' of the exposure site. This definition was modified again in the 2011 FDA monograph [25] to 'the smallest UV dose that produces *perceptible redness* of the skin with clearly defined borders'. In contrast, the original COLIPA test method [5] and the ISO 24444 SPF test methods [6, 12] (2010 and 2019) have always used the definition of the MED as the energy required to produce the first 'perceptible, *unambiguous* redness reaction'. These inconsistencies in definition have led to significant discrepancy in how laboratories have interpreted the qualifying level of response required for determination of SPF values. Since the unprotected MED is in the denominator of the SPF ratio, it has a dramatic impact on the value of the SPF value in a non-linear way. A review of the unprotected MED values from laboratories across the globe [26] revealed the wide range of UV doses considered by laboratories to be qualifying responses for a 'unambiguous redness reaction'. This led to the mandatory range setting for unprotected MED determination in the ISO 24444:2019 [12] method based on the subject's ITA° as the predictor of the unprotected MED to attempt to normalize the level of erythema that is considered as qualifying for this test method. Even with this mandated range, there is still discrepancy between laboratories regarding 'how much' redness is required for a qualifying response. This mandated dose range for unprotected MED testing has not been required in testing according to the FDA test methods and some laboratories maintain unrealistic 'minimally perceptible erythema' doses for their determination of the SPF of a product resulting in unreproducible high SPF values for products tested by them. The ISO24444:2019 version contains Annex F (A visual guide for erythema grading) with photographic examples and explanations of grading choices several for test sites to help give a visual impression of the extent of erythema and the rationale for the grading scores.

Attempts have been made to utilize an objective measure of the erythema 'redness' with photographic or by reflectance spectroscopy. Initially the COLIPA 1995 test method used a criterion of a Δa value of +2.5 using the *L*a*b* colour space (CIE, 1976) [27] for a qualifying response. This method however has variability in results as well as the measurement is pressure sensitive when the measuring devices are set on the skin. Photographic techniques suffer from lack of standardization of the lighting and camera equipment and have not correlated well with (highly variable) visual assessments.

With increasing pressure to move from invasive and damaging clinical testing on human subjects, it is proper to now turn efforts towards perfecting alternative test methods that can provide the same information with equivalent or better reproducibility. Much of the difficulty of validating alternative measurement methods lies in the inherent variability of the 'gold standard' method and the immense effort required in time and expense to generate the in vivo reference standard results for comparisons. It is unlikely that a multi-laboratory in vivo SPF test of this magnitude will ever be repeated for these reasons.

With the data and statistical observations from this extensive ring study, it is clear that additional improvements in inter-laboratory reproducibility can and should be made. Product matrices (galenic properties) may play a more important role than previously expected (i.e. PG5 and PG7 results) requiring more attention to application practices across laboratories.

New non-invasive tests with ISO endorsement are now available to serve as alternatives to and for improvement upon the 'Gold Standard' SPF test. With this endorsement, laboratories can now have confidence to invest in the new equipment needed to practice these new test procedures, and with training and practice provide the path forward to improving upon the variability difficulties encountered in the traditional 'Gold Standard' in vivo test methods.

ACKNOWLEDGEMENTS

The authors have nothing to report.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Curtis Cole https://orcid.org/0000-0003-2519-3198
Bertrand Colson https://orcid.org/0000-0003-4743-1150

REFERENCES

1. Autier P, Boniol M, Severi G, Dore J. Quantity of sunscreen used by European students. Br J Dermatol. 2001;144:288–91.
2. Azurdia R, Pagliaro J, Diffey B, Rhodes L. Sunscreen application by photosensitive patients is inadequate for protection. Br J Dermatol. 2001;140(2):255–8.

3. Bech-Thomsen N, Wulf H. Sunbathers' application of sunscreen is probably inadequate to obtain the sun protection factor assigned to the preparation. Photodermatol Photoimmunol Photomed. 1992;9:242–4.
4. Lademann J, Schanzer S, Richter H, Pelchrzim R, Zastrow L, Golz K, et al. Sunscreen application at the beach. J Cos Dermatol. 2004;3(2):62–8.
5. Blum H, Eicher M, Terus W. Evaluation of protection against sunburn. Phys Rep. 1945;146:118–25.
6. US-FDA Sunscreen Monograph. Sunscreen drug products for over-the-counter human use. Fed Regist. 1978;43:166.
7. FDA- Sunscreen Monograph. Sunscreen drug products for over-the-counter human use. Fed Regist. 1993;64(98):1978.
8. Deutches Institut für Normung. Experimentelle dermatologische Bewertung des Erythemschutzes von externen Sonnenschutzmitteln für die menschliche Haut. DIN Standard 76:501:1-9 1985.
9. COLIPA. Sun Protection Factor Test Method – COLIPA Publication 94/289. 1994.
10. International Harmonized SPF method. Geneva: International Organization for Standardization; 2006.
11. International Standards Organization. ISO24444:2010 cosmetics – sun protection methods – in vivo determination of the sun protection factor (SPF). Geneva: International Organization for Standardization; 2010.
12. International Standards Organization. Cosmetics – sun protection methods – in vivo determination of the sun protection factor (SPF). ISO24444. Geneva: International Organization for Standardization; 2019.
13. Pissavini M, Tricaud C, Wiener G, Lauer A, Contier M, Kolbe L, et al. Validation of an in vitro sun protection factor (SPF) method in blind ring-testing. Int J Cosmet Sci. 2018;40:263–8. https://doi.org/10.1111/ics.12459
14. Bacardit A. Determining the ability to differentiate results between independent sun protection factor tests using the ISO2444 method. Front Med. 2023;10:1042565. https://doi.org/10.3389/fmed.2023.1042565
15. Pissavini M, Doucet O. Interpretation of SPF in vivo results: analysis and statistical explanation. Cosm & Toil. 2011;126(3):172–84.
16. Colson B, Volhardt J, Uhlig S. ALT-SPF study—validation of alternative methods for the determination of SPF and UVA-PF; design, criteria, and performance. Int J Cos Sci. (this issue). 2025.
17. ISO 5725-6:1994 Accuracy (trueness and precision) of measurement methods and results — Part 6: use in practice of accuracy values. Geneva: International Organization for Standardization; 1994.
18. Agin P, Edmonds S. Testing high SPF sunscreens: a demonstration of the accuracy and reproducibility of the results of testing high SPF formulations by two methods and at different testing sites. Photodermatol Photoimmunol Photomed. 2002;18(4):169–74.
19. Stanfield J, Ou-Yang H, Chen T, Cole C, Appa Y. Multilaboratory validation of very high sun protection factor values. Photodermatol Photoimmunol Photomed. 2002;27:30–4.
20. Miksa S, Lutz D, Guy C, Delamour E. Sunscreen sun protection factor claim based on *in vivo* interlaboratory variability. Int J Cosmet Sci. 2016;38:541–9.
21. Bimczok R, Gers-Barlag H, Mundt C, Klette E, Bielfeldt S, Rudolph T, et al. Influence of applied quantity of sunscreen-products on the sun protection factor-a multicenter study organized by the DFK Task Force Sun Protection. Skin Pharmacol Physiol. 2007;20(1):57–64.
22. Zago I, Ben Bari S, Tirard A, Miksa S, Renoux P, Questel E. Overview of proficiency testing results for the in vivo determination of sun protection factor. Int J Cosmet Sci. 2024;46:1097–104.
23. Rees J. The genetics of sun sensitivity in humans. Am J Hum Genet. 2004;75:739–51.
24. FDA monograph - Sunscreen drug products for over-the-counter human use. Fed Regist. 1999;64:98.
25. FDA monograph- Sunscreen drug products for over-the-counter human use. Fed Regist. 2011;76:117.
26. Cole C. Global data of unprotected skin minimal erythema dose relationship to individual typology angle (ITA°). Photoderm Photoimmunol Photodermatol. 2020;36(6):452–9.
27. CIE. CIE 1976 uniform color spaces. In: Colorimetry. International Commission on Illumination Publication.

How to cite this article: Cole C, Colson B, Uhlig S. The variability of in vivo sunscreen sun protection factor values. Int J Cosmet Sci. 2025;47(Suppl. 1):25–36. https://doi.org/10.1111/ics.70000