J. Soc. Cosmet. Chem., 44, 289–297 (September/October 1993)

The methylene blue and image analysis tests for determining skin roughness: A critical assessment based on data in the literature

F.-J. WORTMANN and G. WORTMANN, *Deutsches Wollforschungsinstitut, Veltmanplatz 8, D-51 Aachen, Germany.*

Received November 9, 1992.

Synopsis

The correlation of data published by Schrader *et al.* (2) for skin roughness measurements using the methylene blue and image analysis methods is investigated. The results show that the correlation between the methods is poor and unsatisfactory for prediction and calibration purposes. A classification of the data into four classes shows a pronounced inconsistency between results of the two methods. This in turn leads to serious questions, if not doubts, with respect to the validity, reproducibility, and accuracy of the tests when specifying the performance of skin care products.

INTRODUCTION

On page 980 of his well-known book, Schrader (1) points out (in translation) that

> Scientific procedures for the evaluation of cosmetic products are continuously gaining importance. To fulfill the high quality standards, demanded nowadays by the market, optimization of products just for optical appearance and for processing performance is insufficient. Testing procedures of practical relevance are required to quantify reproducibly the positive and negative effects of cosmetic products.

For skin care products an important effect is related to the reduction of skin roughness. For a recent paper, Schrader and Bielfeldt (2) conducted comparative studies of skin roughness measurements, using image analysis and several *in vivo* skin testing methods, including the methylene blue adsorption test.

According to Schrader (1), the methylene blue test is a simple method for the evaluation of skin roughness. The rougher the skin, the larger its surface and the more intensive the absorption of methylene blue. The methylene blue method has been widely used for industrial quality improvement and control purposes, as well as for reasons of consumer information and protection (3).

In their paper, Schrader *et al.* (2) present, besides other results, a set of data for a range of skin care products, using the methylene blue test as an indirect method and the image analysis method as a direct method for evaluating skin roughness. On investi-

gating the correlation between the results of the methods, they come to the conclusion that "the methylene blue method provides a satisfactory correlation with the image analysis method. It confirms that this simple procedure gives usuable data on the influence of cosmetics on skin roughness."

After a reevaluation of the data provided by Schrader *et al.* (2), the present paper was initially just intended to express some doubts with respect to this conclusion. However, it turned out that a closer look at the data provides, on the one hand, an extended view of certain aspects of the validity and predictive power of both the methylene blue (MB) and image analysis (IA) tests, and, on the other hand, some insight into the general problems of testing skin roughness.

RESULTS

The data on which this paper is based are taken from Table I in reference 2, and are reproduced here in Table I. The data are normalized mean roughness results and relate to 20 individual values (i.e., to 20 test subjects) for each of 22 skin care products. As described in reference 2, "The 20 individual data sets that were obtained at the beginning and the end of each test were normalized. The initial value was set at 100%, so that

Table I
Data Taken from Table I in Reference 2 for Normalized Mean Skin Roughness Values of 20 Test Persons After Applying 22 Different Skin Care Products and the Ratings of These Products Into Four Classes

	Methylene blue test		Image analysis test	
Product no.	Roughness %	Rating	Roughness %	Rating
1	78.7	+	94.86	+
2	81.6	+	92.93	+ +
3	84.8	+	90.94	+ +
4	86.5	+	93.91	+ +
5	88.3	+	95.57	+
6	88.7	+	95.97	+
7	90.0	+	94.70	+
8	90.0	+	97.05	+
9	94.9	o	97.52	+
10	98.0	o	92.83	+ +
11	98.5	o	92.47	+ +
12	98.7	o	97.83	+
13	100.6	o	95.97	+
14	103.6	o	97.77	+
15	104.4	o	102.54	−
16	105.3	o	102.54	−
17	105.5	o	97.76	+
18	109.3	−	98.65	o
19	110.3	−	102.55	−
20	113.5	/	57.09	/
21	114.5	−	94.45	+
22	122.7	−	95.91	+

the value of use indicates the percentage change the product has achieved for each parameter." Values above 100% indicate increased skin roughness, while values below 100% relate to smoothed skin.

The methylene blue roughness (MBR) values are plotted versus those for image analysis roughness (IAR) in Figure 1. This figure shows that there is an outlying point that on the one hand is well contained in the data range for the MB method, but on the other hand is well separated from the range of the IA results. The correlation of the data is negative and not significant. Though the outlying point is contained in the original data of Schrader *et al.* (2), they leave it out without comment for their regression analysis.

Accepting, for the time being, the extreme value as an outlier, a "cleaned" set of data is obtained, summarized in Figure 2, which is equivalent to Figure 1 in reference 2. The parameters for the linear regression line (solid line in Figure 2) are in agreement with reference 2:

slope:	b = 1.70 (0.705)
y-axis intercept:	a = −66.1 (10.27)
correlation coefficient:	r = 0.484
number of data points:	n = 21

The values in brackets are the standard deviations of the parameters. Though the data points relate to 20 individual results and are actually mean roughness values, they will be treated as single values in what follows, since there is no information about, for example, their standard errors. The regression analysis is based on the assumption that the image analysis method provides an independent variable that exhibits an accuracy by

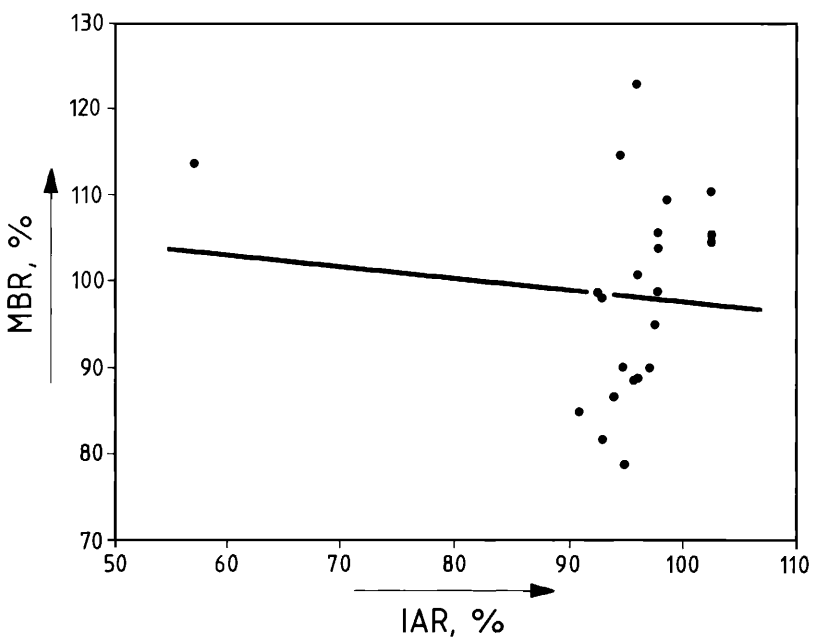

Figure 1. Methylene blue roughness (MBR) vs image analysis roughness test results (IAR) for all data in Table I. Linear regression line (——).

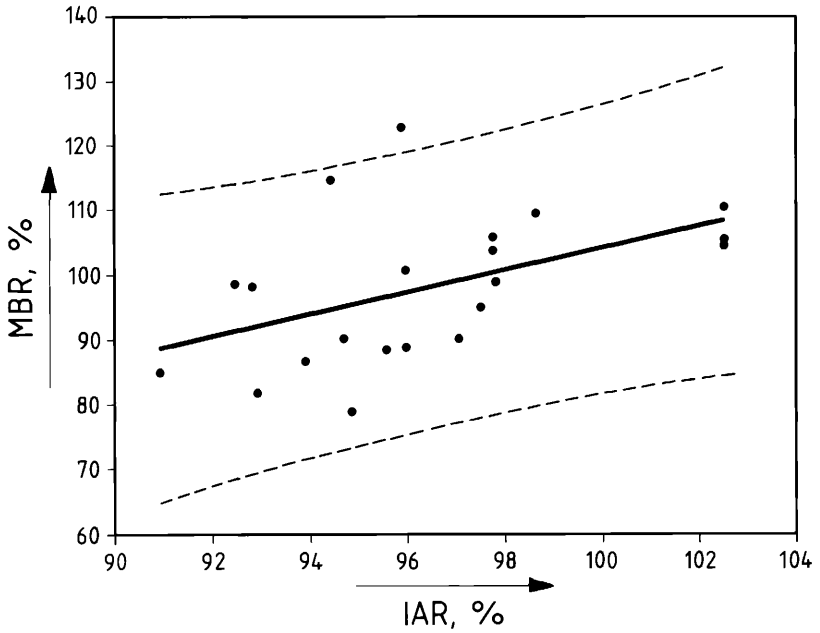

Figure 2. MBR vs IAR results for the "cleaned" data set. Linear regression line (——) and 95% confidence limits for the prediction of single MBR values (– – –).

far superior compared to the MB method. This assumption appears reasonable in view of the added digit in the IAR values as compared to the MBR values in reference 2 (see Table I).

To test the significance of the regression, it is checked whether the true slope of the regression line β is significantly different from zero. This may be done by applying a t-test (4). For the test the statistical significance is set to the usual 95% level.

A parameter $\hat{t}$ is calculated from the data given above according to

$$\hat{t} = b/s_b = 2.41 \quad (1)$$

where s_b is the standard deviation of the slope. $\hat{t}$ is checked against the relevant t-value of the Student distribution for a double-sided test and for DF = n − 2 = 19 degrees of freedom, which is $t_{95\%(2),19} = 2.093$. Since $\hat{t}$ is larger than this value, the hypothesis $\beta = 0$ has to be rejected on the 95% level. However, it is important to note that already an increase to a 98% level ($t = 2.539$) leads to the acceptance of the hypothesis $\beta = 0$ and hence to a rejection of the assumption of correlation between MBR and IAR results. The correlation can hence be considered as being only just significant. This has severe consequences with respect to two types of important conclusions that may be drawn from the correlation.

First, it must be asked, on the basis of the data in Figure 2, within which range an MBR result may be expected when conducting a test with a product for which the IAR result is known.

Predicting an MBR value, $\hat{y}_i$, for a given IAR value, x_i, the range within which the MBR result can be expected with 95% probability is given by Zar (4):

$$\hat{y}_i \pm t_{95\%(2),19}\, S_{\hat{y}_i} \tag{2}$$

where

$$s_{y_i} = \sqrt{s^2_{y\cdot x}\left[1 + 1/n + (x_i - \bar{x})^2 \Big/ \sum_{j=1}^{n} (x_j - \bar{x})^2\right]} \tag{3}$$

$S^2_{y\cdot x}$ is the residual mean square from analysis of variance, where $s_{y\cdot x}$ is known as the standard error of regression. n is the total number of IAR data with the individual values x_j and the mean $\bar{x}$.

On the basis of equations 2 and 3, the so-called "confidence bands for single observations" are calculated for the whole range of IAR values. They are given as broken lines in Figure 2. For further discussion, the values for selected products, representing the data range, are summarized in Table II.

Accepting for the current argument IA as an objective reference method and the MB method as an indirect but fast and practical method, this will lead to the question within which range (95% confidence limits) the true IAR value can be expected if the MBR value for a product is known from experiments. This question involves a method known as "inverse prediction" (4).

Based on the correlation of the IAR and MBR results, the expected IAR value $\hat{x}_i$ can be predicted from an MBR value y_i by

$$\hat{x}_i = (y_i - a)/b \tag{4}$$

The related prediction range is given by:

$$\bar{x} + b(y_i - \bar{y})/K \pm t/K \sqrt{s^2_{y\cdot x}\left[(y_i - \bar{y})^2 \Big/ \sum_{j=1}^{n} (x_j - \bar{x})^2 + K(1 + 1/n)\right]} \tag{5}$$

where

$$K = b^2 - t^2 s^2_b \tag{6}$$

and $t = t_{95\%(2),19} = 2.093$.

Table II
IAR and MBR Values for Selected Cases and the Minimum Values for the 95% Confidence Limits of the Prediction of Single MBR From IAR Values (all values in %)

Product no.	IAR value measured	MBR value measured	MBR value predicted	95% Confidence range for MBR prediction	
				min	max
3	90.94	84.8	88.5	65.1	112.0
21	94.45	114.5	94.5	72.3	116.7
18	98.65	109.3	101.7	79.4	123.9
15	102.54	104.4	108.3	84.5	132.0

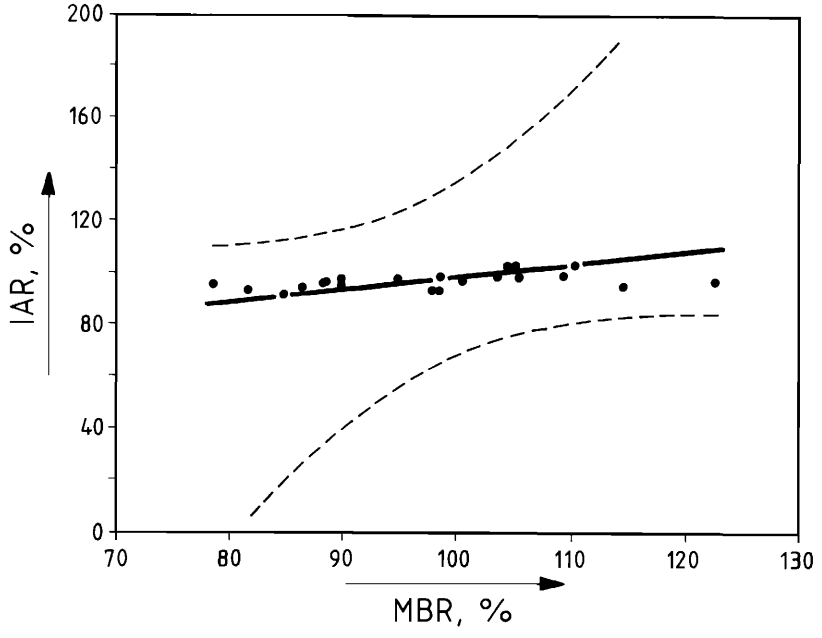

Figure 3. IA vs MB test results (analogous to Figure 2) and 95% confidence limits for the inverse prediction of single IAR values from MBR data (———). Solid line gives the IAR value predicted on the basis of the MBR = f(IAR) correlation given in equation 4.

The 95% confidence bands for the inverse prediction are shown in Figure 3. The solid line gives the predicted $\hat{x}_i$ (=IAR) values. The values for selected products to represent the data range are given in Table III.

To compare the results for the IA and MB tests on a semantic basis, adapted from that applied for consumer-related testing (3), the results for each of the two tests are grouped into four classes, namely *very good* (class 1, + +), *good* (class 2, +), *acceptable* (class 3, o), and *unacceptable* (class 4, −). The class widths were chosen to comprise the range of the cleaned data set in Figure 2, where the position of the classing scheme was set such that the center of class 3 (acceptable, o) was 100%, i.e., the value for unchanged skin

Table III
MBR and IAR Values for Selected Cases and the Minimum and Maximum Values for the 95% Confidence Range of the Prediction of IAR Values From MBR Results (all values in %)

Product no.	MBR value measured	IAR value measured	R value predicted	95% Confidence range for IAR prediction	
				min	max
1	78.7	94.86	85.1	−7.7	109.8
7	90.0	94.70	91.8	39.7	115.9
13	100.6	95.97	98.0	69.8	136.0
19	110.3	102.55	103.7	80.5	171.3
22	122.7	95.91	111.0	83.9	226.7

Table IV
Number of Products in Four Different Classes, Characterized by Their Means and Their Widths (given in brackets) for MB and IA Testing of 22 Skin Care Products

		MB test (16)		IA test (4)	
	Descriptor	Mean	Number of cases	Mean	Number of cases
Class 1	+ +	68	0	92	5
Class 2	+	84	8	96	12
Class 3	o	100	9	100	1
Class 4	−	116	4	104	3

roughness. Table IV summarizes for the test procedures the positions of the classes and the number of products found therein. The ratings for the individual products according to MB and IA testing and based on this scheme are given in Table I.

DISCUSSION

The results of Figure 1 show that there is initially no correlation between MBR and IAR values. Only a reduced data set without the outlier (IAR = 57.09) leads to a correlation that is just significant on the 95% confidence level (see Figure 2). Such a correlation would be of practical value if it could be applied for calibration purposes, namely for the accurate prediction of an unknown MBR value of a product from its IAR value or vice versa.

The 95% confidence range for the predicted MBR values in Figure 2 indicates that already the first task is rather difficult. As Table II shows for four cases, chosen to cover the observed range of IAR values, the predicted MBR values are already between 4 and 10 units off the mark. The situation worsens when the minima and maxima of the possible range for the MBR values are considered. In all cases and regardless of the IAR value, the possible minimum value is within the range of the well-performing and the maximum value in the range of the badly performing products. Even for the product performing best in image analysis (No. 3, IAR = 90.94, + +), its MBR value can be expected to be somewhere between MBR_{min} = 65 (better than + +) or MBR_{max} = 112 (−). The worst performing product in image analysis (No. 15, IAR = 102.54, −) might actually do quite well in MBR testing (MBR_{min} = 84.5, +), but it might also come out as totally unacceptable (MBR_{max} = 132, worse than −).

The comparison between Figures 2 and 3 shows that the situation even worsens if the correlation is applied as a calibration to predict IAR values from values obtained via the MB method, which is more practical to apply. The results summarized in Figure 3 are detailed for five cases in Table III, showing that the predicted IAR values are between 1 and 15 units off the mark.

The 95% range for the IAR-predictions shows that in all cases unrealistic values are obtained, either for the minimum or the maximum, so that no power can be conceded to MB testing to predict IAR values.

Taking the image analysis as the objective test method and providing the independent

variable in the correlation, this would, furthermore, seem to imply that MB testing has no discriminating power with respect to testing skin roughness. This consequence does not fit the general empirical experience with the test (5), and it has, hence, to be concluded that both methods are subject to substantial and possibly similar measurement errors that will make the validity of the discrimination of small effects, and hence of small product differences, questionable.

This point is supported by the existence of the outlier for IA testing. Though we accepted the point as an outlier to conduct the statistical analysis of the data, this decision is in reality rather difficult to justify. The outlier does not relate to a single data point, which might be omitted on the basis of the assumption of a single measurement error, but it actually is the mean roughness value measured for the product applied by 20 test persons. This leads to the conclusion that image analysis is either prone to substantial, random, experimental errors or that in this case a systematic measurement bias passed undetected for a whole series of tests. In any case, the outlier is difficult to understand and its existence certainly gives weight to the assumption that IA testing is subject to experimental errors comparable to those of the methylene blue method.

Comparing the test results for the two methods on a semantic basis, the results in Table IV show that image analysis comes to substantially better ratings in the classification scheme than the methylene blue test. A total of 17 products is rated as *good* or *very good* by IA, while no product is classed as *very good* by MB and only 8 as *good*. In contrast to image analysis, MBR testing indicates that skin roughness is largely unchanged or only slightly improved after cosmetic product use.

While the overall ratings are already substantially different for the two test methods, the individual classifications show a high degree of inconsistency. Only in six cases do both tests come to equal ratings; in 11 cases ratings are obtained that are one class apart. In four cases ratings are even dissimilar to the extent of two classes, so that, for example, products 10 and 11 are rated *very good* (++) by IA and only *acceptable* by MB, and products 21 and 22 are rated *good* by IA and *unacceptable* by the MB test (see Table I).

In view of the similar quality of the correlations found by Schrader *et al.* (2) for the scanning method vs IA ($r = 0.315$) and for skin moisture evaluation vs IA ($r = -0.59$), similar arguments and conclusions as for MB vs IA ($r = 0.484$) can plausibly be expected to apply to these methods. However, a complete evaluation of the data in reference 2 is outside the scope of the present article.

CONCLUSION

Both methods for the assessment of skin roughness, namely the methylene blue and the image analysis tests, obviously exhibit substantial variability in their results, which even with a rather rough classification system lead to inconsistent test results for the 22 skin care products investigated.

This, in turn, leaves some serious questions, if not doubts, with respect to the validity, reproducibility, and accuracy of the tests and, hence, with respect to the requirements set out by Stiftung Warentest (3) for such tests, namely that (in translation), "Modern methods of investigation help us to unequivocally and objectively prove effects and

record even small differences," where the results for specified commercial products are passed on to the consumer for his information and as buying recommendations.

REFERENCES

(1) K. Schrader, *Grundlagen und Rezepturen der Kosmetika,* 2. Auflg. (Hüthig Buch Verlag, Heidelberg, 1989).
(2) K. Schrader and S. Bielfeldt, Comparative studies of skin roughness measurements by image analysis and several *in vivo* skin testing methods, *J. Soc. Cosmet. Chem.,* **42,** 385–391 (1991).
(3) Sonderheft, *Kosmetik* (Stiftung Warentest, Berlin, 1992).
(4) J. H. Zar, *Biostatistical Analysis,* 2nd ed. (Prentice-Hall, Englewood Cliffs, NJ, 1984).
(5) G. Padberg, Modifizierte Methylenblau-Methode zur Prüfung des Rauhigkeitsgrades der Hautschicht, *J. Soc. Cosmet. Chem.,* **20,** 719–728 (1969).