Assessment of visibility of facial wrinkle reduction by various types of observers

J.H.D.M. Westerink
Personal Care Institute, Philips Research,
Prof. Holstlaan 4, 5656AA Eindhoven, The Netherlands

Summary

The prime objective of many facial wrinkle-reduction treatments is to achieve visible improvement. Thus the visibility of before/after treatment differences is often part of an efficacy assessment. This paper investigates whether the background knowledge of the people acting as observers in such assessments is of impact on the results: e.g. the subjects themselves are familiar with their faces, skin professionals have much experience in judging skin quality, and thus both might be more sensitive to small changes.

In a clinical study 44 Female subjects were regularly treated during a period of 12 weeks with one of three wrinkle-reduction methods: K, L and M (placebo). Photographs were taken before treatment and at 6 and 12 weeks. The photographs were judged by 3 types of observers:

- 24 Lay observers were given the 0&6-week and the 0&12-week photo pairs of all subjects to indicate the one with the least wrinkles in a two-alternative forced-choice (TAFC) procedure.
- The subjects themselves were given the 0&6-weel and the 0&12-week pair of their own photos (8 replications) to indicate the photo with the least wrinkles (TAFC).
- A trained panel of skin professionals (N=3) each gave 9-point Fitzpatrick wrinkle-severity scores for all individual 0-week and 12-week photos.

We found that the lay observers perceived the same differences as the subjects themselves: significant improvements after 12 weeks for treatment K ($p<0.0005$ and $p=0.005$, respectively), no visible effects for treatments L and M, and, most importantly, a significant difference between treatments K and M/placebo ($p=0.02$ and $p=0.04$, respectively). Also the trained panel found this difference between K and M ($p=0.013$), but here it was due to a significant deterioration over time of the 'placebo-treated' wrinkles (M, $p=0.03$).

Thus in conclusion we have found no indications that extra knowledge - in the form of familiarity with the own face or in the form of professional training - results in the identification of more treatments that show significantly visible wrinkle-reduction.

Keywords: skin smoothness, wrinkle reduction, visual assessment, observer types

Introduction

In many studies assessing the efficacy of anti-aging treatments, e.g. wrinkle reduction of the face, photographs play an important role, often in addition to instrumental skin measurements [1]. Even in everyday life, photos are meant to bridge time, and this is exactly their use in efficacy assessments. Whereas it might be difficult to remember how one looked some time ago, it is much easier to compare photos taken before and after treatment. This will allow small improvements to become noticed.
We wondered whether some observers would be more likely to detect small changes than others. A group of potentially critical judges of wrinkle reduction are skin professionals, like dermatologists and beauticians. They are used to inspect and interpret minute differences in skin quality which may be meaningless to the layman's eye. In addition, they have the possibility to adhere in their judgments to internationally agreed classifications of wrinkle severity [2][3][4].

A second group of potentially critical judges are the persons who applied the treatment to their own face. We can expect that, since they are highly familiar with their own face, they would be more likely to see small differences in wrinkle severity. On the one hand, this would enlarge our chances of finding significant visibility of wrinkle reduction. On the other hand, differences seen by the treated person themselves are probably the most relevant in terms of user satisfaction. Thus we were interested to assess whether type of observer and background knowledge is of influence on the (significant) detection of visible wrinkle reduction.

Experiment set-up

In the autumn of 2000, a clinical study was performed that allowed us to explore the influence of observer background knowledge in wrinkle-reduction visibility assessments. The study investigated several possibilities for skin conditioning and skin rejuvenation. The study involved 44 female subjects (aged 37-58), each of whom applied a wrinkle-reduction treatment to both sides of her face twice a week for 10 minutes. Two treatment principles were investigated (indicated with K or L). Subjects were treated with one of these methods or a placebo treatment (indicated with M) during a period of 12 weeks. Each subject was assigned to one of these treatments for the full duration of the study, and they were not told to which group they were assigned. Neither could the subjects feel which type of treatment principle was applied.

During the 12-week treatment period various aspects of the skin were monitored: mechanical properties [5], skin profile & texture and thickness & structure. In addition to that, we took photos at week 0 (baseline, before the first treatment), at week 6 and week 12 (final result). In order to take the photos under identical circumstances, even if they were a number of weeks apart, we employed a chin-rest for the subject to rest her chin in. At a fixed distance from the chin-rest, and at three fixed angles, the photo camera could be positioned. In this way, each time three photos were make of each subject: front view, left-side view and right-side view, see Figure 1 for an example. In addition, a professional photographer ensured that the lighting conditions were maximally similar over sessions. Prints were made of all negatives ensuring a magnification factor of 1 (yielding photos at real-life size). Also, care was taken that despite slight differences in exposure conditions, the photos were printed with a constant color reproduction by matching the color and density (darkness) of the chin-rest.

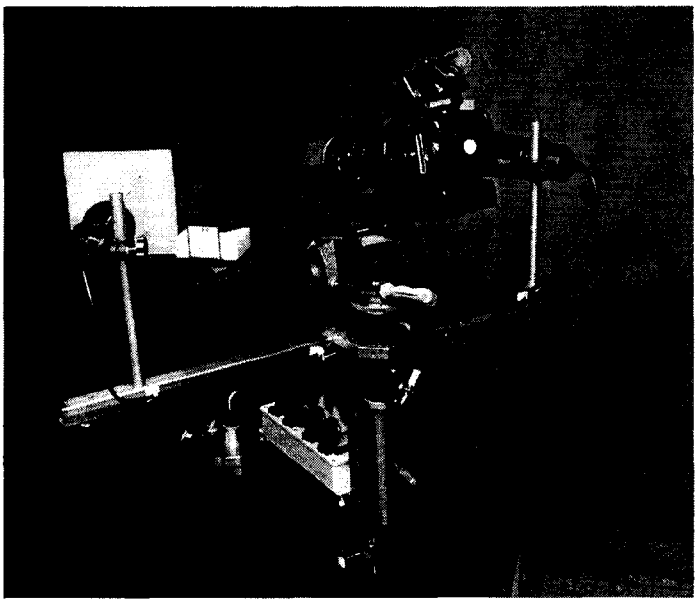

Figure 1. Set-up for taking photographs. The subject rested her chin on the white chin-rest. A little plate on the chin-rest identified the subject with her subject number. Illumination came from two sources on either side of the face. The camera & illumination sources can be rotated to 3 fixed positions for left, front and right views of the face at a fixed distance, and presently it is in the position to take the left-side photo.

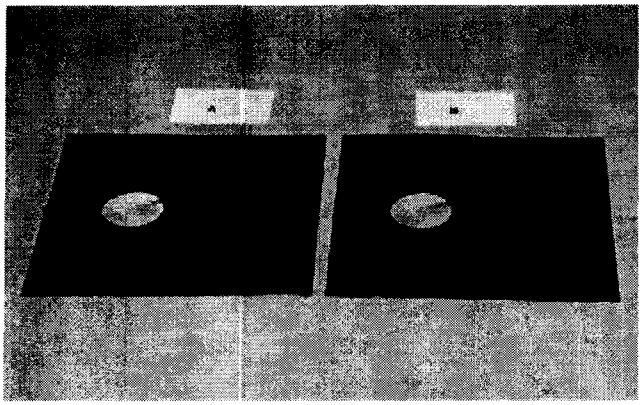

Figure 2. A pair of photos with their black overlays.

These photo prints then were the basis for the assessment of the visibility of wrinkle reduction of the treatments. They were judged by various groups of people (the subjects themselves, a panel of trained dermatologists and independent layman observers) on the basis of before-treatment & after-treatment photographs of the subjects' faces.

Photo comparisons by independent observers

Visibility of wrinkle reduction was evaluated by 24 independent, lay observers. They had no professional skin-related knowledge, nor did they know the subjects on the photographs.

We asked for their judgments using a two-alternative forced choice (TAFC) method. This methodology is specifically suited to measure small differences. The procedure is as follows: two photographs are presented to the observer, and the observer is asked to indicate on which of these two he/she thinks *the wrinkles are less severe.*

In order to prevent that the observer is distracted by the clothing or make-up of the subjects, most of the photograph is covered by a black overlay (see Figure 2), leaving the temple part with the crow's feet visible. The hole in the overlay showed that part of the face where the other (objective) measurements had been taken as well, and it served to help the observers to concentrate on this area.

We only presented photographs taken of the sides of the faces and restricted ourselves to only one side of each of the 44 subjects (the so-called 'priority side'). We carefully randomized which facial sides were designated 'priority side', having approximately equal numbers of left and right priority sides in each treatment group. Of each priority side, we presented the 0&6 weeks and 0&12 weeks combinations, amounting to 88 photo pairs.

The photo pairs of all subjects were put into a pseudo-random presentation order. Sometimes, the 0-week photo is presented on the observer's left-hand side (position A), sometimes on the observer's right-hand side (B), and we took care that both occurred equally often for each pair. The pseudo-random presentation order furthermore ensured that each pair was presented once to each observer, that subjects, and treatment type & duration were evenly distributed over the presentation order, and that they were counterbalanced over observers in order to minimize order effects.

The experiment leader presented the photo pairs sequentially on a table, with letters A and B well indicated directly above the photos. The desk was illuminated with a strong desk lamp in addition to normal ceiling illumination, in order to ensure controlled, adequate and reflection-free illumination of the photos. The task of the observer is for each presented pair to indicate in which photo (A or B) he/she thinks *the wrinkles are least severe.* This will not always have been obvious to the observer, and there might have been (many) cases in which he/she has had to make an intuitive guess. This was acknowledged to the observers in advance. In addition, the observer was asked to indicate the judgments of which he/she was absolutely sure.

Subjects judging their own photos

A few weeks after their last treatments, the subjects were presented with their own photos. They were not presented with the photos of the other subjects. The same overlays were used as in the

experiment with the independent observers (see above) leaving exactly the same area visible through the hole.
The evaluation was done as follows: The experiment leader presented pairs of photos using the same TAFC methodology as used for the judgments of the independent observers. Most important were the 16 presentations of the subject's priority side (see section above): 8 repeated presentations for the 0&6 week combination, and 8 for the 0&12 week combination. Half of the 0&6 pairs were presented with the baseline photograph in the A position, for the other half it was in the B position, and similarly for the 0&12 combinations. These 16 pairs were interleaved with 4 pairs of the non-priority side (the side not evaluated by the independent observers) and 4 pairs featuring the forehead (two 0&6 combinations and two 0&12 combinations of each, one pair of each with the baseline photo in position A and one pair with the baseline photo in position B). The presentation order was different for each subject, guaranteeing that all types of combinations and baseline photo positions were evenly spread over each presentation order, and that aggregated over all observing subjects all combinations were presented equally often in the beginning, middle and final parts of the presentation order.

Photo judgments by a trained panel
All of the side-view photos were judged for their wrinkle severity by a panel of 3 skin professionals: a professor in dermatology, a dermatologist-in-training and an ex-beautician now working as a nurse in a dermatology clinic. These professionals can be considered a trained panel, because they had compared and calibrated their judgments for a series of test-photos prior to the photo evaluation. This is why we adhered to their usual test procedure, even if that meant that their way of judging was not similar to that of the independent observers and the subjects themselves.
The photographs were viewed by the panel members without lay-over, in order to show the full extent of wrinkles in their complexion. Each of the photographs was judged in absolute sense for its wrinkle severity according to a pre-defined 9-point scale defined by a set of example photographs: a 9 indicates very severe wrinkles, a 0 indicates absence of wrinkles (see Fitzpatrick e.a., 1996, [6] for more details on the scale and training procedures used). The photographs were judged by all three panel members individually in no particular order.

Analysis of the data
The necessary transformations of the raw data were performed with Excel and SPSS:
- For each photo pair in the photo comparisons it was counted how often the judges chose the photo taken after the 6 (or 12 week) treatment as the one with the least wrinkles.
- The Fitzpatrick judgments of the trained panel were directly scored on an integer scale ranging between 1 and 9.

Also the consecutive statistical analysis was performed in Excel or SPSS:
- In analyzing statistical differences between treatment groups or treatments, we used t-tests for parameters on an integer or percentage scale (Fitzpatrick scale data, percentages).
- Sometimes we checked whether such a parameter was significantly different from what we would theoretically expect if no change in wrinkle visibility had occurred, and also in this case we used t-tests.

Results

Photo comparisons by independent lay observers
When a group of untrained observers judges wrinkle severity, we stand a chance that they have different criteria of what is important in their decision: one person might value the amount of wrinkles, while another might be more susceptible to their depth. We investigated our data in this respect with the aid of Figure 3, where the percentage of choices for the photo taken after 6 or 12 weeks of treatment, is plotted as a function of the percentage of observers who indicated to be certain of their choice. It is clear from this graph that in the instances where the observers are generally certain of their choice, they do not disagree on the photo they chose as having the least wrinkles. This means that although criterion differences still might exist in this area, they do not influence the results. On the left-hand side of the graph (more uncertainty), we find less unanimity in the observers responses, which might be due to differences in criterion, and –more likely- to random forced choices in no-difference situations.

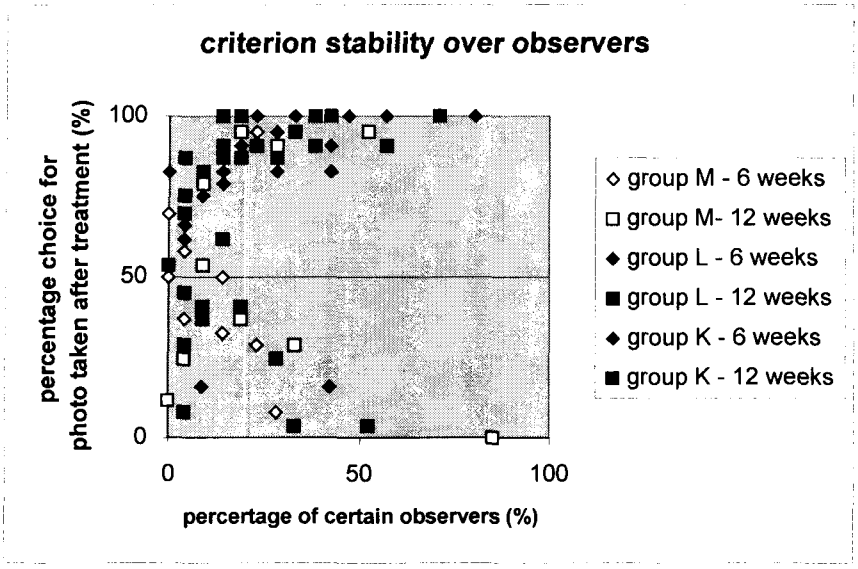

Figure 3. Criterion stability over lay observers.

In order to differentiate between the three treatment groups, the 12&0-week comparison data of each subject is arranged according to group in Figure 4. For almost all subjects in group K the independent observers generally agree that the photo taken after treatment shows the least severe wrinkles, and thus for this group we find that the mean percentage is significantly higher than chance (50%), $p<0.0005$. This is not the case for the other two groups ($p>0.49$ for both). Comparing the groups, we find that group K shows significantly higher percentages than the other two groups ($p=0.011$ for the difference of means with group L, and $p=0.015$ for the difference of means with group M).

For the comparison between weeks 6&0 (not shown), we find that both group K and group L have percentage means that are significantly higher than chance ($p<0.001$ and $p=0.002$, respectively). As a consequence, the only significant difference between groups is that between groups K and M ($p=0.003$) in favor of group K.

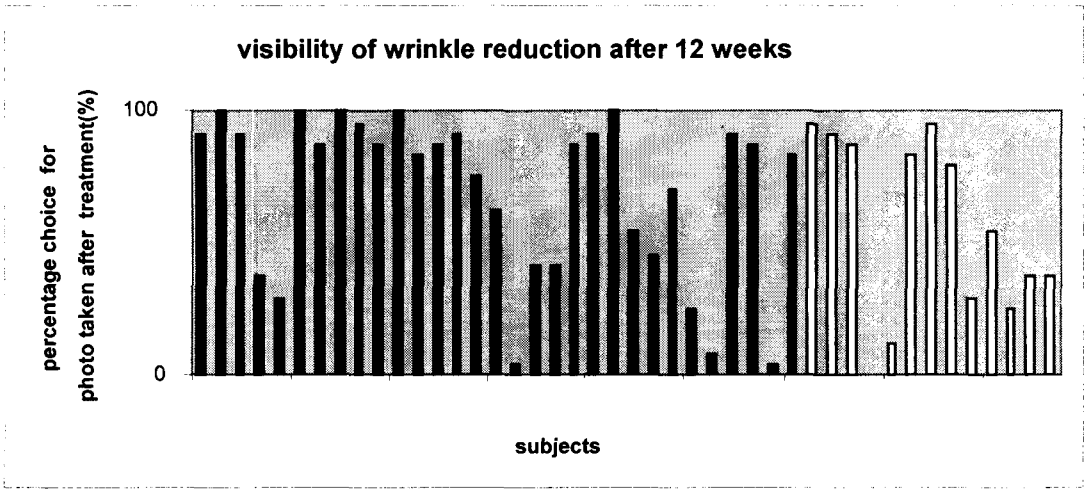

Figure 4. Visibility of wrinkle reduction as judged by 24 independent observers. The 15 subjects of group K are indicated by black bars on the left, the 16 subjects of group L by hatched bars in the middle and the 13 subjects of group M by white bars on the right.

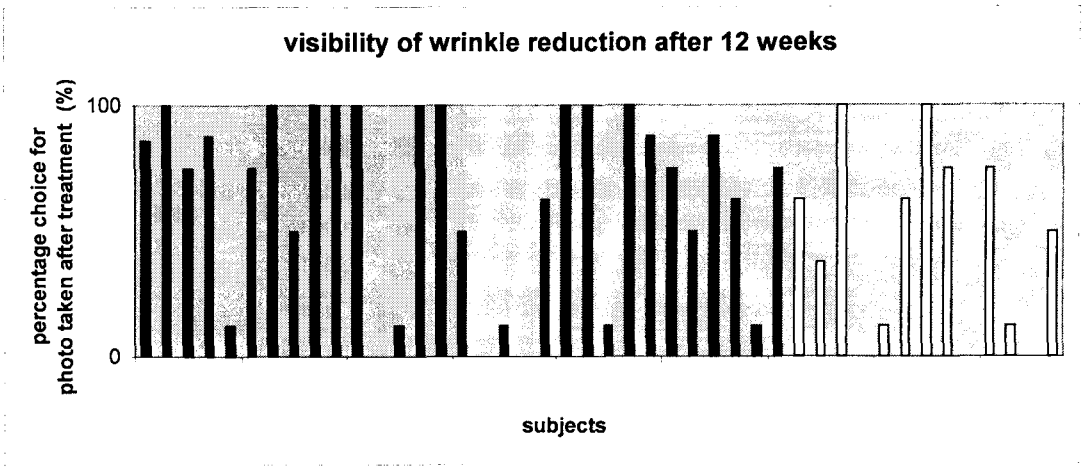

Figure 5. Photo comparisons by the subjects themselves. The 15 subjects of group K are indicated by black bars on the left, the 16 subjects of group L by hatched bars in the middle and the 13 subjects of group M by white bars on the right.

Photo comparisons by the subjects themselves

The photo comparisons done by the subjects themselves for their priority-side photos show more or less the same results as those done by the independent observers (see Figure 5). Group K is the only group where we find the average percentage of choices for the photo taken after treatment to be significantly higher than no change (chance, 50%) for both the 6-week and the 12-week photos ($p<0.001$ and $p=0.005$, respectively). The only other average percentage that is significantly higher than 50% is found for group L, but only at t=6 weeks ($p=0.001$). The only between-group percentage differences that are significant are found between groups K and M, both for 6 weeks and for 12 weeks ($p=0.03$ and $p=0.01$, respectively).

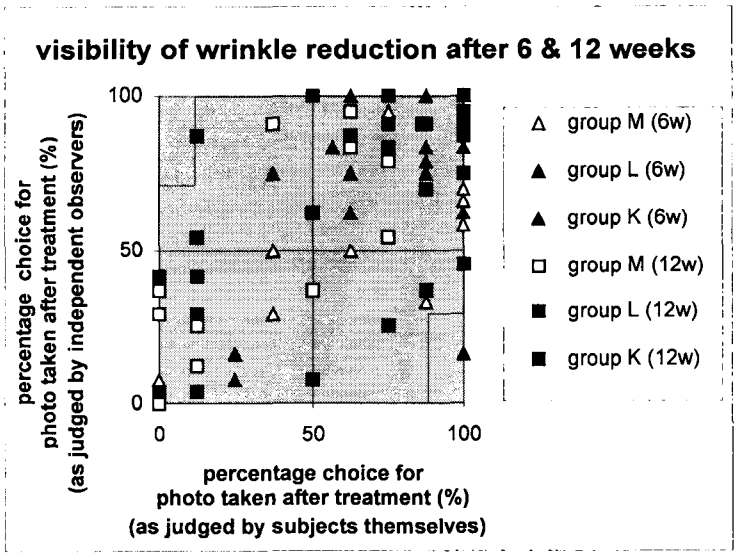

Figure 6. Photo comparisons by the subjects themselves and by independent observers. The lines at 50% indicate no change (chance visibility).The boxes in the upper-left and lower-right hand corners indicate areas of contradiction: an individual subject in the upper-left area judges herself consistently (significantly) to show wrinkle-reduction, while the independent observers (significantly) agree that on the contrary, the wrinkle are enhanced. These significances are calculated on the basis of a binomial distribution be sure ($\alpha=0.05$) of a visible wrinkle reduction, and were found to occur for percentages of 71% and 88% for the independent observers and the subjects themselves, respectively.

Despite the fact that the averaged observations of the subjects themselves are very much in line with those of the independent observers, a scatter plot shows quite some scatter (Figure 6). In general the data show a distinct correlation (corr.=0.66), but nevertheless a considerable variation in the individual data. This emphasizes the effects of statistics and of variations in personal smoothness criteria: judges (observers & individual subjects) appreciate different aspects of smoothness, e.g. one judge might value 'few wrinkles', the other 'shallow wrinkles'. The graph also shows that these effects do not lead to many outright contradictions (in only 2 out of 88 cases one subject sees a significant improvement where the independent observers see a significant deterioration, or vice versa).

Photo evaluations by a trained panel

The Fitzpatrick scores given by the members of the expert panel are absolute scores, and we subtracted those of the 0-week photo from those of the 12-week photo. These differences were averaged over both sides of the face and over panel members. The results are shown in Figure 7. In general, the differences are small, almost always smaller than a full grade. Strangely enough, the graph shows a significant wrinkle increase over time for the placebo group (group M, p=0.03). As a consequence, the difference between group K and group M is significant (p=0.013), even though group K itself does not show a significant improvement over time.

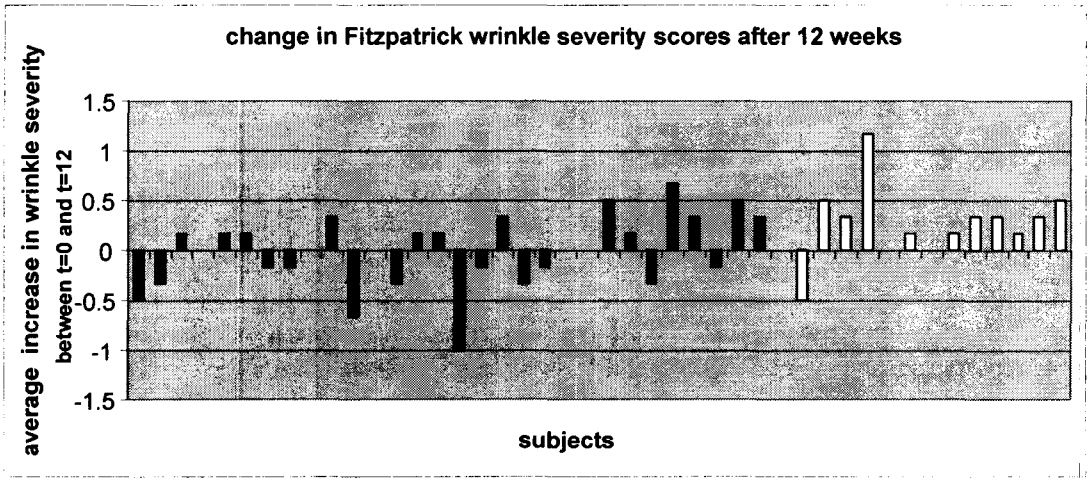

Figure 7. Results of Fitzpatrick evaluation by a trained panel. Depicted The 15 subjects of group K are indicated by black bars on the left, the 16 subjects of group L by hatched bars in the middle and the 13 subjects of group M by white bars on the right.

Discussion

In general, we find that all three observer types find more or less the same significant differences between treatment groups (see Table 1): significant differences between treatment groups K and M for the average percentages/scores of both the 6-week and the 12-week photos, in favor of group K. Since the subjects in treatment group M received the placebo treatment, this allows us to conclude that treatment K indeed does achieve a visible wrinkle reduction. The three observer types also agree that none of them finds a significant difference between treatment groups L and M. Apparently treatment L does not yield visible effects. One less important difference between the observer groups is that only the independent observers find an additional significant difference between the 12-week photos of group K and group L.

Table 1. Overview of significances found.

Difference tested	Observer type		
	Independent observers	Subjects themselves	Trained panel
K6 different from K0	p<0.001	p<0.001	-
K12 different from K0	p<0.0005	p=0.005	n.s.
L6 different from L0	p=0.002	p=0.001	-
L12 different from L0	n.s.	n.s.	n.s.
M6 different from M0	n.s.	n.s.	-
M12 different from M0	n.s.	n.s.	p=0.03
K6 different from L6	n.s.	n.s.	-
K12 different from L12	p=0.011	n.s.	n.s.
K6 different from M6	p=0.003	p=0.03	-
K12 different from M12	p=0.015	p=0.01	p=0.013
L6 different from M6	n.s.	n.s.	-
L12 different from M12	n.s.	n.s.	n.s.

We had expected that maybe the subjects themselves would be better judges of wrinkle differences than independent observers, because they would know their own skin best. This does not come forward from the data: both groups yield (approximately) the same number of significances. Instead of having a stricter criterion, it appears that the subjects show more variability in their judgment criteria, as can be seen from Figure 6, where there are 2 data points found to lie well outside the expected region. We explained this by assuming that subjects involved might have distinctly different criteria than the majority of the independent observers. Similar differences in criterion are likely to occur amongst the independent observers themselves, especially when they indicate not to be certain of their choice (left side of Figure 3), but at least for them, these differences occur equally in the evaluation of each subject.

The advantage of a trained panel is that judgments are obtained relatively quickly and easily, once the panel is trained, and moreover, are positioned on an absolute wrinkle severity scale known from literature. Also, we expect that the criteria used are relatively stable. Nevertheless, there is one indication that their criterion might be different from that of the independent observers, and that is they find a wrinkle increase after 12 weeks for group M, whereas at that point in time both the subjects themselves and the independent observers find a wrinkle decrease for group K. We do not know the reason behind this difference. Possibly the fact that the trained panel has judged the full photos, that is: without overlay, has been of influence. Usually observers state that on tanned skin wrinkles are less conspicuous, and this might have had more of an influence on the full photos as judged by the trained panel. Especially the 0-week (summer) photos comprised many cases of tanned skin color, and this would exactly fit the trained-panel results that in the placebo-treatment group (M) wrinkles enhance over time.
Whether this explanation is correct or not, the fact remains that the trained observers find the same significant difference between the groups as the other observers, and their extra background knowledge is not apparent in terms of an additional significant difference between other treatment groups.

For each of the observer types in this study, we chose panel size, number of presentations and judgment procedure on the basis of practicality and feasibility in actual test situations. Thus our interpretations and conclusions must concern the combined effect of observer type and related test set-up. If we had chosen the test set-up to be equal for all three observer groups, intrinsic differences between observer types might have come forward, that have remained unnoticed so far.

Conclusions

The preceding discussions can be summarized into the following conclusions regarding the methods used:

- Visual evaluations on the basis of photos are a robust method to evaluate visibility of wrinkle reduction. Independent observers, a trained panel of professionals and the subjects themselves yield similar results as far as the comparison of treatments is concerned.
- The method involving (forced choice) photo comparisons by independent observers is slightly preferred because it yields slightly more significant differences between groups, and because the averaged results are less prone to criterion variability.
- Expected advantages of more background knowledge from the subjects themselves and the members of the trained panel were not apparent in our data.

Acknowledgements

Very many persons contributed to obtaining the results, and I would like to thank them all for their efforts: All observers and all subjects, who showed considerable endurance in additionally participating in photo evaluations, Jan Engel, who helped with discussions on the statistical analysis, Sam Bazelmans, for his tenacity in getting the right color reproduction on the photo prints and Tom Nuys, Femke Wagemakers, Nele Vervaet, Arjen Cense and Patricia van der Bruggen for their help in planning and performing the experiments.

References

[1] S.S. Hawkins, D.I. Perrett, B. Tiddeman, D.M. Burt, C. Desantis, L. Meyers, K. Hoyberg, S.L. Wrigth and R.L. Weinkauf, 2002, Novel approaches in texture measurement for cosmetic anti-aging evaluation, Proceedings 22nd IFSCC Congress, Edinburgh, UK.

[2] C. Larnier, J.-P. Ortone, A. Venot, B. Faivre, J.-C. Béani, P. Thomas, T.C. Brown and E. Sendagorta, 1994, Evaluation of cutaneous photodamage using a photographic scale, *British Journal of Dermatology* **130**, 167-173.

[3] C.E.M. Griffiths, T.S. Wang, T.A. Hamilton, J.J. Voorhees and C.N. Ellis, 1992, Aphotonumeric scale for the assessment of cutaneous photodamage, *Arch. Dermatol* **128**, 347-351.

[4] K. Tsukahara, Y. Takema, H. Kazama, Y. Yorimoto, T. Fujimara, S. Moriwaki, T. Kitahara, M. Kawai and G. Imokawa, 2000, *J. Cosmet. Sci.* **51**, 127-139.

[5] L. Schlangen, D. Brokken and P. van Kemenade, 2003, Correlations between small aperture skin suction parameters: statistical analysis and mechanical model, *Skin Research and Technology* **9**, 122-130.

[6] R.E. Fitzpatrick, M.P. Goldman, N.M. Satur and W.D. Tope, 1996, Pulsed carbon dioxide resurfacing of photo-aged facial skin, *Arch. Dermatol* **132**, 395-402.