

Visual assessment of human skin irritation: a sensitive and reproducible tool

DAVID BASKETTER, FIONA REYNOLDS, MATT ROWSON, CLAIRE TALBOT AND ED WHITTLE

Unilever Environmental Safety Laboratory, Colworth House, Sharnbrook, MK44 1LQ, UK

Human volunteer studies of skin irritation have been carried out for decades, both for research and for safety evaluation purposes. For the majority of this time, and consequently for the majority of the studies, assessment of the skin reactions has been made visually. Typical endpoints include erythema, oedema, dryness and scaling, some or all of which would be rated on a simple scale, eg 0, $\pm$, +, ++, +++. Such approaches can be criticized as subjective, of poor reproducibility, lacking in sensitivity and highly variable between observers and/or institutions. In consequence, instrumental methods of assessment have been strongly promoted and do indeed offer several advantages, not least their objectivity. However, it is possible to use the human eye, which is a very sensitive tool, to make detailed, accurate and reproducible descriptions of skin irritation reactions. To achieve this, it is necessary to give prolonged and thorough training to each observer. In this paper, 3 examples of human volunteer studies, in which different pairs of trained observers independently carried out double blinded scoring of the irritation reactions, are reported. The grading patterns produced were almost identical; statistical analysis showed that properly trained observers are in fact able to reliably measure a grade of erythema to within ± 1 on a 10 point scale; 97.6% of scores were within 2 grade points on this scale. These results provide evidence that visual scoring can be sensitive, reliable and reproducible within a testing institution.

Key words: irritant contact dermatitis; visual assessment; grading of erythema; skin irritation; reproducibility. © Munksgaard, 1997.

Accepted for publication 7 July 1997

For many years, visual assessment of skin reactions has been the primary, often only, means of their evaluation. This approach has taken both diagnostic and experimental dermatology far. However, in recent years, bioengineering techniques have been successfully applied to the measurement of various skin parameters (1). The use of these methods has become rather fashionable in experimental contact dermatitis studies, particularly of skin irritation, as they provide objective and quantitative data. Of particular note is the assessment of transepidermal water loss (TEWL), which provides an indication of the state of the stratum corneum barrier function (2-4). In some workers' hands, TEWL appears to provide an early, sub-clinical indicator of change in barrier function that may either lead to, or predispose to, subsequent frank dermatitis. Techniques for assessing redness, including colorimetry and laser doppler flowmetry are also readily available, but unlike TEWL, probably do not really offer greater sensitivity than the human eye. However, whilst bioengineering techniques offer some significant advantages, they also present certain disadvantages. They can be time consuming for the ex-

perimenter and/or the human volunteer. Also, they can be easily misapplied or misinterpreted by the unwary (5). In some cases, they may simply represent an unnecessary degree of information in relation to the question that is being asked (6). As a consequence, there is still heavy reliance on the visual assessment of skin reactions in human volunteer studies of skin irritation. Thus it is important that this visual observation is carried out in a reliable and reproducible manner. In the work presented in this paper, we have examined the quality, reliability and reproducibility of the visual observation of relatively modest degrees of skin erythema which is carried out in our own laboratory. This work is done as part of the safety evaluation of products or their ingredients, not in relation to product claims (7, 8). All work reported here was carried out in a Good Laboratory Practice compliant facility.

Materials and Methods

Training of skin observers

In order to train skin observers to an adequate degree, each trainee is carefully instructed using a

Table 1. Visual assessment of erythema grading scale

Grade	Erythematous reaction
0	no reaction
1	marginal reaction
2	slight perceptible erythema
3	a greater than slight reaction which is not sufficient to be classed as distinct
4	distinct erythema
5	a greater than distinct reaction which is insufficient to be classed as well developed
6	well developed, possibly spreading erythema
7	a greater reaction which is not sufficient to be classed as strong
8	strong, deep erythema which may extend beyond the treatment site
9	a more intense reaction than above

photographic index of reaction grades, but primarily by tutored observation of reactions in ongoing studies. Subsequently, every trainee is required to carry out independent blinded assessments of patch tests in parallel with a trained observer(s). Each study completed is then analysed for consistency with the trained observer's assessment. Only when the trainee in double blinded independent scoring consistently (usually 3 studies) obtains results comparable to those of the trained observer and which yield the same overall conclusion from the study are they regarded as trained. Typically, this process takes from 1 to 3 years and will have involved evaluation of some 3000 patch test sites in parallel with trained observers. For each study reported here, scoring of skin sites was carried out by pairs of observers, each working independently and blinded to the substances used on each of the treatment sites. Each study was assessed by a different pair of observers; 5 observers were used in total.

Human skin irritation test methods

The variant of the 48 h patch test employed in our laboratory was used (9). It employs 2×23 h Al-test patch (Imeco AB, Stockholm, Sweden) applied to the upper outer arm of panels of approximately 33 volunteers. A known volume of test solutions plus distilled water (negative control) and the positive control, 0.3% SDS, was applied to individual Al-test patches on Leukosilk, according to a randomized sequence. The multiple patch was then applied to the upper outer arm of the panellist and secured with further Leukosilk tape. After 23 h the patch was removed, the sites wiped clean, marked and then assessed 1 h later. The 10 point scoring scheme used is shown in Table 1. At this point, if sufficient erythema (≥ 6) was evident, the site would not be retreated. The remaining sites had a

further patch applied for 23 h after which time the procedure of wiping, marking and assessing was repeated. A further assessment was performed at 72 h to check that any reactions were subsiding. The positive and negative controls are always included to ensure that the panellists are sufficiently sensitive to respond to a known irritant, but largely free of non-specific reactivity. Since individuals, and therefore panels of volunteers, will produce variable irritation reactions, depending on the time of year, initial skin condition, and a variety of endogenous factors, the positive control also acts as a reference point.

Statistical evaluation

Apart from simple inspection, the results were examined in 2 ways. Firstly, the frequency with which there was perfect agreement between the observers, 1 grade difference, 2 grades difference, (etc.) was calculated. Simultaneous 95% confidence intervals were then calculated for the probability of a difference between assessment falling into each category; of particular importance was the interval for when the 2 observers were within 1 grade point of each other. The 2nd approach was to conduct a binomial test on the results, classifying as acceptable results which were within 1 grade point of each other.

Results

The results of the examination of inter-observer agreement/disagreement in the 3 studies are presented in Tables 2–4. In summary, of 1445 assessment points, in just over $\frac{1}{2}$ the observers gave an identical grade; in 1275 cases (82.4%) agreement was within 1 point on the 10 point grading scale. In only 34 cases (2.4%) was the discrepancy >2 points on the grading scale. Taking all 3 studies together, in only 8 cases (0.6%) was the discrepancy >4 points on the grading scale. Such large differences in erythema assessment are more likely

Table 2. Inter-observer (dis)agreement in Study 1

Difference between observers	Frequency over the study	Proportion	Simultaneous 95% confidence for proportion
0	236	0.555	(0.491, 0.617)
1	163	0.384	(0.324, 0.447)
2	23	0.054	(0.0317, 0.0909)
3	1	0.00235	(0.00027, 0.0205)
4	0	0	0
5	0	0	0
6	1	0.00235	(0.00027, 0.0205)
7	1	0.00235	(0.00027, 0.0205)
<2	399	0.939	(0.902, 0.963)

Table 3. Inter-observer (dis)agreement in Study 2

Difference between observers	Frequency over the study	Proportion	Simultaneous 95% confidence for proportion
0	238	0.484	(0.423, 0.545)
1	174	0.354	(0.297, 0.414)
2	54	0.110	(0.0769, 0.154)
3	11	0.0224	(0.0101, 0.0490)
4	9	0.0183	(0.00759, 0.0434)
5	3	0.00610	(0.00143, 0.0256)
6	2	0.00407	(0.00073, 0.0223)
7	1	0.00203	(0.00022, 0.0188)
<2	412	0.837	(0.788, 0.877)

Table 4. Inter-observer (dis)agreement in Study 3

Difference between observers	Frequency over the study	Proportion	Simultaneous 95% confidence for proportion
0	253	0.479	(0.424, 0.535)
1	211	0.400	(0.346, 0.455)
2	59	0.112	(0.0811, 0.152)
3	4	0.00758	(0.00226, 0.0251)
4	1	0.00189	(0.00022, 0.0159)
5	0	0	0
6	0	0	0
7	0	0	0
<2	464	0.879	(0.839, 0.910)

to be due to data recording/transcription errors, but as this cannot be checked retrospectively, all the points have been included in the statistical analysis. It is worth noting that the majority of the reactions evaluated in these studies were in the lower $\frac{1}{2}$ of the 10 point scale. On this scale, a >7 grade erythema would be approximately equivalent to the "+" grading indicated for the scoring of diagnostic patch tests used in the investigation of allergic contact dermatitis (10).

The statistical evaluation of the data confirmed that there was a high probability that observers would score within 1 grade of each other. The 3 studies were examined in isolation as the precise degree of concurrence is dependent on the characteristics of the different pairs of observers. What is evident from the analyses in Tables 2–4 is that the degree of concurrence is consistently high, albeit slightly variable. The binomial test which examined the null hypothesis was comfortably satisfied, since at the 5% level, the p -values were no higher than 0.015. The proportion of occasions when observers agreed to within 1 grade was greater than 80%.

Discussion

The primary purpose of this paper is to demonstrate that, with a suitable level of investment of

time and effort, individuals can be trained to deliver a consistent visual assessment of subtle grades of skin irritation reaction. This is not to decry the great merit in objective measurements of skin reactions. Rather it is to indicate that when a rapid visual assessment is all that is necessary to meet the purpose of the experiment, then such an assessment can in fact be done well. The data presented show clearly that relatively subtle degrees of erythema can be reliably discriminated by suitably trained observers. The type and purpose of the assessments referred to herein are manifestly different from those required for the diagnosis of allergic contact reactions. However, it is also evident in that area, that suitable training and experience can lead to good concordance in the assessment of reactions (11).

References

1. Serup J, Jemec C. *Handbook of non-invasive Methods*. Springer-Verlag, Heidelberg, 1995.
2. Lammintausta K, Maibach H I, Wilson D. Susceptibility to cumulative and acute irritant dermatitis. An experimental approach in human volunteers. *Contact Dermatitis* 1988; 19: 84–90.
3. Tupker R A, Pinnagoda J, Coenraads P J, Nater J P. The influence of repeated exposure to surfactants on human skin as determined by transepidermal water loss and visual scoring. *Contact Dermatitis* 1989; 20: 108–113.
4. Wilhelm K-P, Maibach H I. Susceptibility to acute irritant dermatitis: correlation to individual biophysical properties. *J Invest Dermatol* 1989; 92: 541–545.
5. Pierard G E. Relevance, comparison and validation of techniques. In: Serup J, Jemec C (eds) *Handbook of non-invasive methods*. Heidelberg: Springer-Verlag, 1995: 9–16.
6. Basketter D A, Chamberlain M, Griffiths H A, York M. The classification of skin irritants by human patch test. *Food and Chemical Toxicol* 1997; 35: in press.
7. Basketter D A, Reynolds F. Use of human volunteers for hazard and risk assessment of skin irritation. In: Close B, Coombes R, Hubbard A, Illingworth J (eds): *Volunteers in research and testing*. London: Taylor and Francis, 1997: 117–127.
8. Basketter D A, Reynolds F S, York M. Predictive testing in contact dermatitis – Irritant Dermatitis. In: Goh C L, Koh D (eds): *Clinics in dermatology – Contact dermatitis*. Amsterdam: Elsevier, 1997: in press.
9. Jenkins H L, Adams M G. Progressive evaluation of skin irritancy of cosmetics using human volunteers. *International Journal of Cosmetic Science* 1989; 11: 141–149.
10. Wahlberg J E. Diagnostic tests. In: Rycroft R J G, Menné T, Frosch P J, Benezra C (eds.) *Textbook of contact dermatitis*. Heidelberg: Springer-Verlag, 1992: 241–268.
11. Bruze M, Isaksson M, Edman B, Bjorkner B, Fregert S, Moller H. A study on expert reading of patch test reactions: inter-individual accordance. *Contact Dermatitis* 1995; 32: 331–337.

Address:

D. A. Basketter
Unilever Environmental Safety Laboratory
Colworth House
Sharnbrook, Bedford, MK44 1LQ
UK

This document is a scanned copy of a printed document. No warranty is given about the accuracy of the copy. Users should refer to the original published version of the material.