

Historical perspective on the use of visual grading scales in evaluating skin irritation and sensitization

Miranda A. Farage¹, Howard I. Maibach², Klaus E. Andersen³, Jean-Marie Lachapelle⁴, Petra Kern⁵, Cindy Ryan¹, Jeanne Ely¹ and Amita Kanti¹

¹The Procter & Gamble Company, Cincinnati, 45224 OH, USA, ²Department of Dermatology, University of California, San Francisco, CA 94102, USA,

³Department of Dermatology and Allergy Centre, Odense University Hospital, DK-5000 Odense, Denmark, ⁴Department of Dermatology, Catholic University of Louvain, Louvain, Belgium, and ⁵The Procter & Gamble Company, Brussels, Belgium

doi:10.1111/j.1600-0536.2011.01912.x

Summary

Visual assessment of skin reactions has long been used to evaluate the safety of chemicals and preparations that contact the skin, and to meet regulatory requirements. This article reviews the history of visual grading scales, and the results of investigations into the reliability of the method. Some examples are provided to illustrate the diverse array of protocols that use visual scoring to evaluate skin irritation. Furthermore, as bioengineering methods are developed that can quantitate certain aspects of skin irritant and sensitization reactions, it is important to consider whether such measures should supplement or replace visual assessment. Examples of investigations comparing the outcomes of studies that use visual scoring and those that use bioengineering methods are discussed. These examples provide little evidence that bioengineering measures provide an improvement in overall quality in comparison with current testing methods that rely on visual assessment. In addition, such measuring techniques can add considerably to the complexity of testing protocols. When benefits and cost are weighed in the balance, the visual assessment scales popularized by Draize and others remain an effective, practical method of evaluation.

Key words: contact dermatitis; grading erythema; patch testing; skin irritation; visual assessment.

A History

The four cardinal signs of an inflammatory response are redness, heat, swelling, and pain, and are said to have been described as early as 178 A.D. by Celsus (1). Skin irritation and sensitization are types of inflammatory processes that exhibit these signs. The degrees of redness, heat, swelling and pain are proportional to the extent of the reaction. For decades, patch testing with visual assessment of these signs has been the cornerstone of evaluating skin irritant and sensitization reactions. The

history of patch testing has been reviewed by Devos and van der Valk (2) and, more recently, by Lachapelle (3). There are reports of patch tests being used for dermatological evaluations as far back as 1847, when Städel (Germany) described a method called the blotting paper strip, which he had devised to examine the effects triggered by *Anacardium occidentale*, or the cashew plant. Jadassohn (1896) (Austria) applied a mercurial preparation topically and reproduced dermatitis in sensitized subjects. Bloch (1911) (Basel) described a patch testing technique for allergens using visual assessment. In the United States, Sulzberger (1931) published on the utility of patch testing in the diagnosis of hypersensitivity (4). In 1932, Graves described the use of patch testing in determining the aetiology of some types of dermatosis reactions, specifically in identifying contact sensitization reactions caused by plants (5).

Correspondence: Miranda A. Farage, Feminine Care Innovation Center, Clinical Sciences, The Procter & Gamble Company, 6110 Center Hill Ave., Box 136, Cincinnati, Ohio 45224, USA. Tel: +1 513 634 5594; Fax: +1 866 622 0465. E-mail: farage.m@pg.com

Accepted for publication 19 February 2011

Unfortunately, these early publications did not provide details on the visual assessment of the reactions, but generally described positive reactions as definite surface changes on the skin, or reactions that included redness, itching, and vesiculation or pustules. Some early authors acknowledged that there were varying degrees of response, and specified that a distinction should be made between a minimal reaction (mere redness), and a maximal reaction (infiltration, and non-purulent and/or purulent bulla). In the 1930s, German dermatologists standardized diagnostic patch testing to a certain degree, describing both patch test technique and scoring of test reactions using grading scales (6, 7). A standard series of patch tests for contact sensitization, with a standardized grading scale, was published by Bonnevie (1939) in Copenhagen (Table 1a) (8).

Early patch test appliances used gauze squares or blotting paper to hold the test material. These were applied to a strip of adhesive tape. A layer of oiled silk between the gauze and the tape provided some occlusion. Improvements in patching appliances were described by Guild in 1939 (11) and Rokstad in 1940 (12). These authors described the use of a glass square and a small plastic

chamber, respectively, to hold the test substances in place. These early appliances developed into a number of standardized patch testing systems, such as the Finn Chambers®, the Webril™ patch, the Hill Top™ patch, and the Van der Bend® chamber. Finn Chambers® have been widely used by many national contact dermatitis research groups in routine diagnostic testing of patients with suspected contact allergies (13, 14).

In 1944, Schwartz and Peck described in greater detail a method of diagnostic patch testing for human subjects in order to identify the potential sources of dermatitis reactions (9). They described the use of gauze squares saturated with the test substance, and applied to the skin for 24 hr. The gauze was covered with a layer of cellophane and adhesive tape. A four-point numerical scale was specified for scoring, with higher numbers indicating more extensive reactions (Table 1b). Interestingly, they also proposed the use of two diagnostic patch tests applied 10–14 days apart as a means of predicting whether a material was a contact sensitizer, using the same four-point scale. This was an early reference to the use of patch testing in prophetic tests for contact sensitization. Protocols for such investigations were subsequently developed.

In using patch testing for contact allergy, authors often stress the importance of avoiding irritant reactions, which can be difficult to distinguish from allergic reactions. However, patch testing and other protocols involving dermal exposure have been developed precisely for the purpose of evaluating irritation. In the same year as the Schwartz and Peck publication, in response to the introduction of many new compounds and topical preparations to the marketplace, scientists at the Food and Drug Administration (FDA) recognized a need to bring some standardization to tests conducted to study local toxicity to the skin and mucous membranes (10). These FDA scientists described several protocols for evaluation of the toxicity (acute and repeated dose) and primary irritation of materials applied to the skin. They included a standard visual assessment scale for primary irritant reactions on the skin of animals and humans, whereby numerical values were assigned to correspond with descriptions of the physiological reactions of redness and oedema (Table 1c). Such a numerical scale was proposed to facilitate comparisons of different compounds, and enable arithmetical interpretations.

The use of diagnostic patch testing in humans became very widespread in the 1950s and 1960s. In 1970, members of the International Contact Dermatitis Research Group (ICDRG) published a common grading scale to facilitate accurate communication of patch test results (Table 2a) (15). This scale became widely used for allergic contact dermatitis.

Table 1. Early standardized visual assessment scales

Description	Numerical equivalent
a. Human diagnostic patch test, 1939 (8)	
Spotty erythema	(+)
Few or more infiltrated papules	+, +(+) , ++
With minimal vesicles	++(+)
Typical oozing eczema in the test area	+++
Pronounced erythema, strong infiltration and closely spaced vesicles	++++(+)
With bullae	++++
b. Human diagnostic patch test, 1944 (9)	
Erythema on the area of application	1+
Erythema and oedema	2+
Erythema, oedema, papules, and a few vesicles	3+
Erythema, oedema, many vesicles, and, in some cases, ulceration	4+
c. Primary irritation, animals and humans, 1944 (10)	
Erythema and eschar formation	
Very slight erythema (barely perceptible)	1
Well-defined erythema	2
Moderate to severe erythema	3
Severe erythema (beet redness) to slight eschar formation (injuries in depth)	4
Oedema formation	
Very slight oedema (barely perceptible)	1
Slight oedema (edges of area well defined by definite raising)	2
Moderate oedema (area raised ~ 1 mm)	3
Severe oedema (raised > 1 mm and extending beyond area of exposure)	4

Table 2. Revised standardized visual assessment scales recommended by regulatory agencies and professional societies (1970s)

Description	Numerical equivalent
a. Human diagnostic patch test recommended by the ICDRG (1970) (15)	
Not tested	NT
Doubtful reaction	? +
Weak (non-vesicular) reaction	+
Strong (oedematous or vesicular) reaction	++
Extreme reaction	+++
b. Primary irritation in animals recommended by the NAS (1977) (16)	
Erythema and eschar formation	
No erythema	0
Very slight erythema (barely perceptible)	1
Well-defined erythema	2
Moderate to severe erythema	3
Severe erythema (beet redness) to slight eschar formation (injuries in depth)	4
Oedema formation	
No oedema	0
Very slight oedema (barely perceptible)	1
Slight oedema (edges of area well defined by definite raising)	2
Moderate oedema (area raised ~ 1 mm)	3
Severe oedema (raised > 1 mm and extending beyond area of exposure)	4
c. Primary irritation in humans recommended by the NAS (1977) (16)	
No response	0
Questionable or faint, indistinct erythema	1/2 or ±
Well-defined erythema	1
Erythema with slight to moderate oedema	2
Vesicles (small blisters) or papules (small, circumscribed elevations)	3
Bullous (large blister), spreading, or other severe reaction	4

ICDRG, International Contact Dermatitis Research Group; NAS, National Academy of Sciences.

With the passage of the Consumer Product Safety Act, and the creation of the Consumer Product Safety Commission (CPSC) in 1972, the regulation of consumer product safety was placed at the federal level in the United States, with a common set of uniform standards (17). The CPSC requested that the Committee on Toxicology of the National Academy of Sciences (NAS) develop methods for evaluating dermal irritation in animals and humans. These were published in a procedures manual in 1977 (16), and used for many years by regulatory agencies, private companies, and research organizations. The test for skin irritation in animals used the visual assessment scale shown in Table 2b. This scale was very similar to that described by Draize et al. (10) (Table 1c),

except for the addition of scores of '0' for no erythema and no oedema.

The protocol for evaluating skin irritation in animals originally published in the 1977 NAS procedures manual, and the accompanying visual assessment scale, remained virtually unchanged for decades. Such protocols were specified by worldwide regulatory agencies as part of the risk assessment process (18–20). In addition, the American Society for Testing and Materials published a version of the protocol (21). Results of skin irritation testing in animals provided the essential data for determining whether materials or products posed a hazard with regard to skin irritation, and therefore required cautionary or warning labelling (22, 23). This protocol is still used by many, although, since the 1990s, there has been a shift towards alternative methods that use fewer or no animals (24).

A protocol for evaluating irritant reactions in humans was also specified in the NAS procedures manual in 1977 (16). This protocol differed from the recommendation of Draize et al., in that it specified a 4-hr exposure time rather than 24-hr, and suggested an alternative visual assessment scale (Table 2c) based on one developed by Marzulli and Maibach in 1975 (25). This integrated scoring scale for erythema and oedema accommodated the wider variety of visible responses seen in humans, that is, papules, vesicles, and spreading reactions.

The NAS standards published in the 1977 procedures manual did not specify particular protocols for evaluating contact sensitization in either animals or humans; however, several protocols were mentioned as being acceptable. For evaluations in animals, these included methods utilizing intradermal injection (10), topical application (26, 27), or a combination of injection and topical application (28). Evaluations in humans included protocols utilizing topical application, with (29) or without (10, 30) prior irritation of the patch test site. All of these methods used visual assessment to evaluate the results, and many specified a scale to score the extent of erythema, and a means for describing the degree of oedema, induration, vesiculation, and necrosis.

As in the case of test protocols for skin irritation, standard protocols for evaluating contact sensitization using visual assessment scales were published worldwide by regulatory agencies (31, 32) and voluntary testing organizations (33). Over the years, protocol variations using visual assessment scales have been developed to evaluate contact sensitization in animals (34, 35) and humans (36, 37). As with evaluations of irritation, there has been a shift in recent years towards alternative methods that use fewer animals. Methods such as the local

lymph node assay are being incorporated into the overall approach for evaluating sensitization (38, 39).

For decades, visual assessment of skin reactions has been used to evaluate the safety of chemicals and preparations that contact the skin, and to meet regulatory requirements for product labelling. However, product manufacturers also have a need to evaluate other aspects of dermal safety that may not be adequately addressed with the protocols discussed above. As a result, numerous testing methods have been developed to provide information on skin effects and product mildness. These protocols vary widely in exposure methods, and are often tailored for specific types of products, and designed to simulate or exaggerate actual exposure conditions that consumers may experience in the real world. Some examples are provided in Table 3. This list is not intended to be a comprehensive list of all test methods currently available or in use, and nor is it an endorsement of these methods. Rather, it is meant to convey the broad and diverse nature of test protocols that have evolved to evaluate unique circumstances of exposure. Although the means of exposure in these tests vary, they share a common reliance on a visual scoring system to determine skin effects.

Reliability

Visual assessment of skin reactions has been the cornerstone of skin testing for decades. Reliable testing facilities make every effort to train the skin graders thoroughly, and to control the grading conditions to reduce the potential for variation. However, it remains a subjective evaluation, and the question is sometimes raised about the reproducibility of the evaluation. Investigators have conducted studies demonstrating the reliability of visual scoring, as reviewed below.

Falk et al. (50) investigated agreement in visual scoring, using both an irritant [sodium lauryl sulfate (SLS)] and induced photoreactions. Volunteer test subjects were

patch tested with a concentration series of SLS (0.5–6%) and increasing exposure to ultraviolet (UV)-B irradiation (42–126 mJ/cm²). Test sites were evaluated by both the subjects themselves (on a two-point scale, i.e. positive or negative) and the investigators (using five-point scales, with any score of >0 being considered a positive). In the irritant patch series, self-readings correlated well with investigator determinations, with absolute agreement being seen in 76.9% of cases. Disagreements occurred when the subject reacted to a broad range of concentrations of SLS. In these instances, the test subjects tended to under-report the reactions. In the UVB series, absolute agreement between investigator and self-reading was found in 85% of cases. The weighted kappa (κ) for agreement between observations was 0.76 for SLS reactions and 0.83 for UVB reactions.

Ivens et al. (51) evaluated the agreement in visual grading among five dermatologists, using photographs of 55 patch tests performed with different allergens. The observers were asked to score the patch tests using the both the ICDRG standard five-point scale and a simplified three-point scale (positive, negative, and irritant). With the ICDRG scale, the results showed fair agreement when pairwise comparisons were conducted among the five dermatologists, with κ -values mostly in the range of 0.42–0.68 (one outlier produced a κ -value of 0.37). Agreement was much higher with the three-point scale with κ -value for pairwise comparisons in the range of 0.61–0.80.

Uter et al. (52) also used digital photographic images to evaluate the agreement on visual scoring. These investigators used 20 different digital images of patch test reactions to contact sensitizers or irritants. These were examined by 122 participants from several countries. Participants included consultant dermatologists, researchers, practising dermatologists, and professors in dermatology. Each participant rated each image on a five-point scale for contact sensitization, or as an irritant reaction. On average, 63.5% of all ratings were in exact agreement. This percentage increased to 87.8% when the ratings were considered on the basis of non-positive (no evident reaction) and positive (any skin reaction, including faint erythema).

A systematic study of the agreement on scoring clinical erythema was conducted by Lock-Andersen and Wulf (53). Twenty-one healthy, Caucasian volunteer subjects received a series of graded UV exposures (five increments and no exposure) symmetrically on both buttocks, resulting in a total of 42 phototests for clinical scoring. Twenty hours to 24 hr after UV irradiation, eight physicians working in a dermatology department where phototesting is a routine procedure visually scored

Table 3. Examples of test protocols that use visual scoring

Method	References
Forearm wash test	(40, 41)
Repeat application tests	(42)
Repeat open application tests	(43)
Cumulative irritation tests	(43)
Chamber scarification test	(44)
Behind-the-knee test	(45)
Fabric wear test	(46, 47)
Extended-duration home use test	(46, 47)
Soap chamber test	(48)
Chamber scarification test	(49)

the erythema reactions on a five-point ordinal scale. Scoring was performed blinded, with the identity of the subjects concealed. Each site was scored twice by the same physician to evaluate intra-observer agreement. Scores from different physicians allowed evaluation of inter-observer variation. Observers were asked to determine the irradiation dose that produced barely perceptible erythema (+), and erythema with a well-defined border (≥ 1). The variability for determining barely perceptible erythema was less than the variability for determining well-defined erythema. For both inter-observer and intra-observer variability, the agreement at the lowest dose was high (83.5% and 90.5%, respectively), with κ -values consistent with good agreement (0.68 and 0.80, respectively). Less agreement was observed as the doses of UV increased. At the highest two dose increments (i.e. the fourth and fifth), agreement was moderate. Inter-observer agreement was 57.1% ($\kappa = 0.45$) and 52.9% ($\kappa = 0.40$), respectively, and intra-observer agreement was 66.7% for both doses ($\kappa = 0.53$ and $\kappa = 0.57$, respectively).

Griffiths et al. (54) conducted an intra-laboratory comparison of results from patch testing of volunteer human subjects using eight substances and 20% SLS as a positive irritant control. Patches of test substances were applied progressively for up to 4-hr to avoid unacceptably high levels of irritation. All laboratories scored reactions using the same four-point scale; however, any degree of erythema (i.e. a reaction of 1+ or greater) was an indication of a positive reaction for irritation. Laboratories were located in the UK, Germany, the United States, and China, and studies were conducted over a 2-year period (1994–1996). Specific protocols varied in the total overall duration of each patch application, and the site of reapplication (i.e. the same site or a different site). In spite of variations in the protocols, there was agreement in approximately 80% of the evaluations between the various laboratories with regard to the classification of materials as irritants for the purposes of labelling.

All of the above investigations lead to an overall conclusion that visual scales are reproducible with regard to classifying reactions as negative or positive, meaning that it is relatively easy to evaluate a reaction for the absence or presence of erythema. Evaluating the intensity of the erythema appears to be somewhat less reproducible. This would suggest that simplified grading scales may be the best approach when comparing results from different investigators or different studies. However, Basketter et al. (55) conducted a controlled evaluation of the reproducibility of patch test scoring in the testing of irritants, using an expanded grading scale. Graders were trained with a photographic index of possible skin reactions, and in side-by-side blinded assessment with experienced

graders. Human volunteers were patch tested (24 and 48 hr) with test materials, including an irritant control (0.3% SLS), and a negative control (water). A total of five observers was used for the studies. In each study, pairs of observers conducted scoring, each working independently. Observers were blinded with regard to the specific test or control sample patch tested at the individual test sites. Skin reactions were scored on a 10-point scoring scale, where 0 was no reaction and 9 was an intense, strong erythema that may spread beyond the exposure site. In this scoring scale, the bottom half of the 10-point scale (scores of 0–5) focused on minimal or slight reactions. The top half (scores of 6–9) was comparable to the more traditional scoring scale in the description of reactions, such as that shown in Table 2c. Thus, the scoring scale was designed to differentiate between mild and medium reactions. Paired scoring was conducted for a total of 1445 individual assessment points over the course of three experiments. In just over half (50.31%), the two scorers gave an identical grade (Table 4). In 1275 cases (88.24%), scores were within one numerical grade of agreement, and in 1411 cases (97.64%), scores were within two numerical grades. In only 34 of the cases (2.35%) was the discrepancy three or more numerical grades out of the 10-point scale. This study suggests that relatively subtle degrees of erythema could be reliably discriminated by observers trained to recognize reactions of various intensity using photographic images.

Visual Scoring versus Bioengineering Methods

Visual scoring can be standardized through appropriate training and experience. However, this method of data

Table 4. Comparison of visual scoring in human patch testing: Inter-observer agreement on a 0–9 scale (55)

Difference between observers ^a	Frequency over studies	
	Number	Percentage
0	727	50.31
1	548	37.92
2	136	9.41
3	16	1.11
4	10	0.69
5	3	0.21
6	3	0.21
7	2	0.14
0 or 1	1275	88.24
0, 1, or 2	1411	97.64
≥ 3	34	2.35

^aThe difference represents the degree of disagreement on the numerical scale of 0–9 between two observers scoring independently.

collection still involves an element of human judgement. A common question is whether a purely objective measure would increase the robustness of the protocols, and enable the differentiation of more subtle skin reactions. Although bioengineering methods have come into more common use for the evaluation of skin reactions and disease states, there is little evidence that the use of such bioengineering methods improves the sensitivity and overall quality of the test result. Furthermore, a potential limitation of bioengineering methods is that each measure evaluates only one aspect of the overall reaction. For example, increases in colour intensity (redness) can be measured with spectroradiometers or chromometers. Increases in blood flow can be measured by laser Doppler flowmetry (LDF). Barrier disruption can be measured with evaporimeters. Increases in skin thickness resulting from oedema and vasodilatation can be measured with ultrasound scanners. However, only visual assessment can evaluate all aspects of the reaction simultaneously. A number of investigators have conducted comparative studies between visual scoring and the results obtained from various bioengineering methods. Some of these comparison studies are reviewed below.

In a study conducted by Willis et al. (56), 10 subjects were patch tested with a series of seven irritants and two controls. Reactions were scored visually on a five-point scale, and evaluated by LDF. The mean LDF readings correlated well with visual scores, producing a Spearman rank correlation coefficient (r) of 0.91 ($p < 0.001$). In addition, the means from LDF measures that corresponded to an incremental increase in the visual scores were significantly different from each other; that is, the mean LDF reading for the reactions that were scored 0.5 on the visual scale was significantly different from the mean reading for sites that were scored 0 and for sites that were scored 1 on the visual scale. However, when LDF measures for individual sites were considered rather than the means, there were some notable outliers that were not consistent with the visual scores, indicating that LDF measures may not always be reliable for individual assessments. Also, the minimum time needed to measure blood flow at 10 patch sites was 45 min, as compared with 5 min for a visual assessment. In addition, the subject was required to lie in a relaxed position in quiet surroundings, in order to obtain steady LDF readings. The authors concluded that LDF was not suitable for routine patch testing.

Seidenari and Belletti (57) compared both colorimetric and echographic procedures with visual scoring. These investigators patch tested 120 nickel-sensitive subjects with 40 mg of 5% nickel sulfate, and evaluated the reactions at baseline (i.e. before patching) and at 72 hr post-patching, using visual, echographic and colorimetric assessments. They found that echographic

evaluations showed significant increases in value for reactions scored visually as +, ++, and + + +. Colorimetric values increased with higher visual scores. However, the intense reactions (+ + +) did not differ from the more moderate reactions (++) .

Agner and Serup (58) applied increasing concentrations of SLS (0.12–1%) via patch tests on the volar surfaces of the forearms of 12 volunteer subjects. Reactions were evaluated at 24- and 48-hr on a visual scale, and by several bioengineering measures [transepidermal water loss (TEWL), colorimetry, LDF, and skin thickness using an ultrasound scanner]. As expected, the mean of the visual scores showed a linear dose–response relationship with increasing concentrations of SLS (statistical analysis was not reported). All bioengineering measuring methods showed a significant difference between normal skin and 0.12% SLS, and a significant linear dose–response relationship.

Lahti et al. (59) conducted a comparative study of visual assessment (four-point scale), LDF and three reflectance erythema meters for measuring erythema reactions. Skin reactions in 15 subjects were caused by UVB irradiation, or two well-known substances known to elicit non-immunological immediate contact reactions in the skin (benzoic acid and methyl nicotinate). The correlation between visual assessment and the measured values for the benzoic acid, methyl nicotinate and UVB reactions was good with all of the instruments, with the Spearman's rank order correlation coefficient varying between 0.54 and 0.79 ($p < 0.001$).

The use of LDF was evaluated by Wahlberg (60). Eleven common solvents were applied undiluted to the volar surface of the forearm of 1 individual, using two different techniques: a 5-min application inside a 20-mm-diameter glass ring, or application of 0.1 ml, which was allowed to spread on the surface and evaporate for 5 min. The results showed a similar rank order of potency to visual scoring. The author concluded that LDF may provide additional information in the case of a marginal irritant; however, it cannot replace the traditional visual scoring.

In a study by Babulak et al. (61), the irritation caused by six different treatments with soap and detergent bar solutions applied to human volar forearms was assessed by three measures: visual scoring, Minolta™ Chroma Meter measurements of redness, and TEWL. All three methods of assessment were in agreement, and two treatments, B and D, were significantly ($p \leq 0.01$) less irritating than the other four treatments (A, C, E, and F). The correlation coefficients between visual scoring and both instrumental measures were significant: $r = 0.97$ for the chroma meter measurement ($p \leq 0.003$), and $r = 0.86$ for TEWL ($p \leq 0.03$).

Zuang et al. (62) undertook an investigation with the goal of determining which of three non-invasive methods (LDF, evaporimetry, and colorimetry) was the best predictor of the visual scores. Fifty-one patients (Caucasian males and females), sensitized to various chemicals, participated in the study. A standard patch test series containing different test allergens was adapted to each patient's medical history, and applied to the skin of the patients' backs. In total, 133 patch test reactions were evaluated on the 51 patients in the study. All three instrumental methods were useful in discriminating between any positive reaction and a negative one. However, with regard to the severity of reactions based on the visual scale of 1+ to 4+, each of the three instrumental methods misclassified skin reactions on known sensitized subjects in 36–38% of cases. The percentage of misclassifications was reduced to 21% when all three methods were used concurrently.

Fluhr et al. (63) evaluated eight bioengineering methods in the assessment of irritant reactions. The goal was to determine which of these methods could detect significant skin changes induced by different types of irritation over time. Irritation reactions were induced with five methods: 2% SLS (6-hr patch), tape stripping (tape applied 15 times), 0.05% tretinoin (2-hr patch), 0.5% dithranol (2-hr patch), and UV exposure (6-second exposure to 30 W/m² UVB and 95 W/m² UVA). Measurements were conducted prior to irritation induction, and at 24, 48 and 72 hr. The results showed that visual assessment was more discriminating in detecting skin changes over time than colorimetric measurement (red–green and black–white measures), capacitance measures, and measures of TEWL. Visual assessment performed similarly to Mexameter[®] haemoglobin measures and LDF. The only method that was more discriminating over time than visual assessment was LDF.

Magnusson and Koskinen (64) evaluated the effects of topical capsaicin treatment, using several methods to determine whether some were more sensitive than others in detecting skin irritation. The methods included impedance, visual assessment with measurement of the flare, panelists' descriptions of cutaneous sensations, LDF measures of cutaneous blood flow changes, and impedance monitor measurement of changes in the skin. Visual assessment, descriptions of sensations and LDF measurements produced similar results, and showed a maximum effect at approximately 30 min, with resolution after 3 hr. Electrical impedance did not show any significant differences when the values at baseline were compared with those after capsaicin application.

Ollmar et al. (65) compared bioengineering methods using irritation induced by 24-hr patch tests with

SLS at concentrations ranging from 0.1% to 5.0%. Skin sites were evaluated prior to treatment (baseline) and on days 1, 2, 7 and 14 post-treatment. They found close agreement between visual scoring, electrical impedance, and TEWL. Electrical capacitance and LDF produced results inconsistent with the other measures, and results obtained with these methods showed fewer significant changes from baseline values.

Fullerton et al. (66) investigated laser Doppler perfusion imaging in scoring the responses of skin sites treated with three irritant materials at increasing concentrations (SLS, retinoic acid, and nonaioic acid). Visual scoring was conducted with a 0–4 grading scale. Laser Doppler perfusion imaging was conducted after a 15-min acclimation of subjects in a temperature-controlled and humidity-controlled environment. A dose–response relationship was observed with both the visual scoring and LDF, indicating similar discrimination for these two assessment methods.

Wigger-Alberti et al. (67) evaluated the irritancy of three metal-working fluids, using visual scoring, TEWL, and chromometry. Two exposure methods were used to produce irritation: a 24-hr patch test, and a cumulative irritation test (6 hr of exposure each day for 13 days, except on the 2-day weekend). All methods of evaluation, that is, visual scoring, TEWL, and chromometry, were in agreement, and produced an identical rank order of irritation of the three test materials in both exposure models.

Held et al. (68) investigated visual scoring with the Derma Spectrometer[™] in detecting contact sensitization in allergic subjects. The Derma Spectrometer[™] measures an erythema index based on blood flow. The study was designed to determine which scoring method better differentiated between degrees of reactivity: visual scoring or measured erythema index. Patients with formaldehyde allergy were patch tested with a concentration series (0–10 000 ppm formaldehyde), and reactions were evaluated. A correlation coefficient was calculated for the concentration of formaldehyde and either the visual score or the erythema index. The resulting visual scores showed a higher correlation coefficient ($r = 0.60$) with the applied concentrations than the spectrometer measure of erythema index ($r = 0.35$), demonstrating that the visual scoring was better able to discriminate between degrees of reactivity.

Andersen and Straberg (69) evaluated the utility of measuring perfusion (by LDF) and swelling (via skin fold thickness) in allergic contact dermatitis reactions in guinea pigs. Contact allergy to chlorocresol was induced in guinea pigs with a maximization method, and animals

were challenged with two concentrations of the substance (1% and 0.1%). Visual scores that were not indicative of sensitization were 0 and 1 (no and slight reactions, respectively). Scores indicative of sensitization were 2 and 3 (moderate and intense reactions, respectively). Measures of perfusion discriminated between animals that were not sensitized (i.e. scores of 0 or 1) and animals that were sensitized (i.e. scores of 2 or 3). Measures of swelling discriminated between the severity of the challenge response (i.e. between scores of 2 and scores of 3). Both measures discriminated between the challenge concentrations of 1% and 0.1% in sensitized animals. These authors concluded that quantitative methods were useful supplemental measures in evaluating inter-observer and inter-laboratory differences.

Comparing the Benefits of Visual Scoring and of Bioengineering Methods

Clearly, a key benefit of bioengineering methods over visual scoring is the fact that bioengineering methods are purely objective, with no element of human judgement. They can directly measure the specific physiological changes that may occur during skin reactions. Quantitative measures of barrier function, hydration, changes in blood flow, etc. can provide insights into biochemical changes and mechanisms of action that occur over the time-course of the reaction. In addition, most bioengineering methods provide results on a linear, continuous scale, making statistics easier to apply. Visual assessment cannot directly measure these specific physiological changes, and nor are the visual scales continuous, or necessarily linear.

Conversely, there are some clear disadvantages to the use of bioengineering methods. Some of these parameters, such as blood flow, skin thickness, and skin pigmentation, can vary between individuals or between different regions of the body. For bioengineering measures that rely on optics, such as LDF, care must be taken to avoid applying samples to test sites that may have uneven pigmentation, such as blemishes or freckles (66). Scaling resulting from dryness, a common condition with some irritants, may also produce optical interference (66). Sample application on the back may not be appropriate, as movement resulting from normal respiration may create artefacts (66). Andersen and Straberg observed effects on measures of perfusion and swelling caused by tape and shaving irritation at the test sites (69). This is only a partial list of some of the controls and considerations necessary when bioengineering measures are used. A more complete discussion can be found in Breternitz et al. (70).

The use of bioengineering methods adds considerable complexity and cost to the testing protocols. For example, many of these measures can be performed effectively only after the subject has acclimatized in a temperature-controlled and humidity-controlled environment. This requirement not only adds steps to the study, but also adds to the time commitment required of volunteers considering participation in the study. Furthermore, the time required to make the measurements is almost invariably longer than that required for a trained, visual grader to make an assessment. It may be necessary to control ambient light and temperature for some bioengineering measures. As a result, the grader must develop an additional set of skills to control and monitor these variables.

Current test protocols for skin reactions have been optimized over decades to provide results that can be assessed via visual scoring. These protocols continue to serve us well in the practical evaluation of irritation or sensitization reactions. There is little evidence that bioengineering measures improve sensitivity or overall quality in comparison with current testing methods, as illustrated by the preceding discussion of comparative studies. However, the rapidly evolving bioengineering techniques will probably prove more useful in providing information that will deepen our understanding of the processes that occur in inflammation and skin reactions. Physiological changes that occur early in the process of irritation, such as changes in blood flow, moisture content, and pH, would be expected to occur before any reaction is visible. In fact, such changes undoubtedly occur in reactions that never present with visible changes, that is, subclinical reactions. An ability to measure these physiological changes may allow us to detect subtle skin effects (71). We have recently reported the use of cross-polarized light to allow visualization of subsurface erythema; this increased the ability to distinguish between subtle skin effects caused by two very similar products following application of the materials in a behind-the-knee study (72). In order to fully exploit this new technology, it is likely that specific protocols will need to be developed with this objective in mind.

The usefulness of any scoring method is directly related to the purpose of the experiment. When evaluating the presence or absence of irritation or sensitization as the result of exposure to a material, bioengineering methods tend to provide information that is not necessary, without providing a better measure of these endpoints. After all, there is no need to use a sledgehammer to drive a nail, or a calliper to measure a kilometre. Basketter et al. (55) pointed out that, when a rapid visual assessment is all that is necessary to meet the purpose of the experiment, such

an assessment can be performed reliably in the hands of a sufficiently trained grader. When benefits and cost are weighed in the balance, the visual assessment scales popularized by Draize and others remain an effective, practical method of evaluating skin irritation and contact sensitization reactions.

Acknowledgements

The authors are grateful to Drs F. Ajayi, K. W. Miller, M. Simone, R. Bartolo and M. Lafranconi for technical review of this manuscript, and to Terresa L. Nusair of the Health and Environmental Safety Alliance (Cincinnati, Ohio) for technical input.

References

- Movat H Z. The acute inflammatory process. In: *Inflammation, Immunity and Hypersensitivity*, 2nd edition, Movat H Z (ed.): Hagerstown, MD, Harper & Row, 1979: pp. 1–162.
- Devos S A, van der Valk P G. Epicutaneous patch testing. *Eur J Dermatol* 2002; **12**: 506–513.
- Lachapelle J M. *Giant Steps in Patch Testing: A Historical Memoir*: Phoenix, AZ, Smart Practice, 2010: 169 pp.
- Sulzberger M B, Wise F. The contact or patch test in dermatology: its uses, advantages and limitations. *Arch Derm Syphilol* 1931; **23**: 519–531.
- Graves J M. The patch test – its use in dermatology – report of cases. *Cal West Med* 1932; **36**: 157–160.
- Mayer R L. *Das Gewerbeekzem. Schriften aus dem Gesamtgebiet der Gewerbehygiene*: Berlin, Verlag von Julius Springer, 1930, Neue Folge, Heft 30: pp. 1–89.
- Blumenthal F, Jaffé K. *Ekzem und Idiosynkrasie*: Berlin, Verlag von S. Karger, 1933: pp. 1–188.
- Bonnevie P. 1939 *Aetiologie und Pathogenese de Ekzemkrankheiten. Klinische Studien über die Ursachen der Ekzeme unter besonderer Berücksichtigung des Diagnostischen Wertes der Eksamproben*: Copenhagen/Barth, Letapzig, Busch.
- Schwartz L, Peck S M. The patch test in contact dermatitis. *Public Health Rep* 1944; **59**: 546–557.
- Draize J H, Woodard G, Calvery H O. Methods for the study of irritation and toxicity of substances applied topically to the skin and mucous membranes. *J Pharmacol Exp Ther* 1944; **82**: 377–390.
- Guild T B. Window patch test. *Arch Derm Syphilol* 1939; **39**: 807–810.
- Rokstad I. Patch test, new modification of (the chamber method). *Arch Dermat Syphilol* 1940; **41**: 649–653.
- Cronin E. Technique of patch testing. In: *Contact Dermatitis*, Cronin E (ed.): Edinburgh, Churchill Livingstone, 1980: pp. 1–19.
- Fisher A A. The role of patch testing. In: *Contact Dermatitis*, 3rd edition, Fisher A A (ed.): Philadelphia, PA, Lea & Febiger, 1986: pp. 9–29.
- Wilkinson D S, Fregert S, Magnusson B et al. Terminology of contact dermatitis. *Acta Derm Venereol* 1970; **50**: 287–292.
- National Academy of Sciences. *Dermal and eye toxicity tests. Principles and Procedures for Evaluating the Toxicity of Household Substances*: Washington, DC, National Academy of Sciences, National Research Council, 1977: pp. 23–59.
- Consumer Product Safety Commission CPSC. Consumer Product Safety Commission CPSC History, 2010. Available at: <http://www.usrecallnews.com/2008/05/us-consumer-product-safety-commission-cpsc.html> (last accessed 24 March 2010).
- U.S. Environmental Protection Agency, OPPTS harmonized test guidelines series 870 – health effects test guidelines. OPPTS 870.2500 Acute Dermal Irritation, EPA712-C-98-196, August 1998. Available at: http://www.epa.gov/oppts/pubs/frs/publications/Test_Guidelines/series870.htm (last accessed 6 April 2010).
- OECD. OECD guideline for the testing of chemicals 404. 24th April 2002. Acute toxicity: Dermal Irritation/Corrosion, 2002. Available at: <http://ecb.jrc.ec.europa.eu/testing-methods/annex5/> (last accessed 2 April 2010).
- Title 16: Commercial Practices PART 1500 – hazardous substances and articles; administration and enforcement regulations § 1500.41 Method of testing primary irritant substances, 2010. http://www.access.gpo.gov/nara/cfr/waisidx_00/16cfr1500_00.html (last accessed 2 April 2010) p442.
- ASTM F 719 – 81 (Reapproved 2007). *Standard Practice for Testing Biomaterials in Rabbits for Primary Skin Irritation*, 2007. Available at: <http://www.astm.org/SEARCH/sitesearch.html?query=ASTM+F+719+-+81> (last accessed 2 April 2010).
- U.S. Consumer Product Safety Commission. Office of compliance requirements under the Federal Hazardous Substances Act: labeling and banning requirements for chemicals and other hazardous substances. 15 U.S.C. § 1261 and 16 C.F.R. Part 1500, August 2002. Available at: <http://www.cpsc.gov/businfo/regsumfhhsa.pdf> (last accessed 2 April 2010).
- Guidance to regulation (EC) No. 1272/2008 on classification, labeling and packaging of substances and mixtures. 27 July 2009. 3.2.2.1.2.5 Testing methods: in vivo data. 2009: p. 222. Available at: http://ecb.jrc.ec.europa.eu/documents/Classification-Labeling/CLP_Guidance_to_Regulation.pdf (last accessed 2 April 2010).
- OECD. OECD guidelines for the testing of chemicals. Tests No. 430, 431 and 435, 2010. Available at: <http://titania.sourceoecd.org/vl=2183769/cl=13/nw=1/rpsv/cw/vhosts/oecdjournals/1607310x/v1n4/contp1-1.htm> (last accessed 26 May 2010).
- Marzulli F N, Maibach H I. The rabbit as a model for evaluating skin irritants: a comparison of results obtained on animals and man using repeated skin exposures. *Food Cosmet Toxicol* 1975; **13**: 533–540.
- Buehler E V. Delayed contact hypersensitivity in the guinea pig. *Arch Dermatol* 1965; **91**: 171–175.
- Ritz H L, Buehler E V. Planning, conduct, and interpretation of guinea pig sensitization patch tests. In: *Current Concepts in Cutaneous Toxicity*, Drill V A, Maibach H I (eds): New York, Academic Press, 1980: pp. 25–40.
- Maguire H C Jr. The bioassay of contact allergens in the guinea pig. *J Soc Cosmet Chem* 1973; **24**: 151–162.
- Kligman A M. The identification of contact allergens by human assay. III. The maximization test: a procedure for screening and rating contact sensitizers. *J Invest Dermatol* 1966; **47**: 393–409.
- Shelanski M V, Gabriel L. *Cutaneous Toxicity Evaluation of Air Force Development Materials – IV*. ASD Technical Report 61-77: Dayton, OH, McGregor & Werner, Inc., 1961. Contract No. AF 33(616)-6962, Project No. 7165, Task No. 71836.
- U.S. Environmental Protection Agency. OPPTS harmonized test guidelines series 870 – health effects test guidelines.

- OPPTS 870.2600. Skin Sensitization. EPA 712-C-03-197, March 2003. Available at: http://www.epa.gov/oppts/pubs/frs/publications/Test_Guidelines/series870.htm (last accessed 6 April 2010).
- 32 OECD. OECD guideline for the testing of chemicals 406. 17th July 1992. Skin Sensitization, 1992. Available at: <http://ecb.jrc.ec.europa.eu/testing-methods/annex5/> (last accessed 2 April 2010).
 - 33 Active Standard: D 6355 – 07. *Standard Test Method for Human Repeat Insult Patch Testing of Medical Gloves*. ASTM International.
 - 34 Klecak G. Identification of contact allergens: predictive tests in animals. In: *Dermatotoxicology*, 3rd edition, Marzulli F N, Maibach H I (eds): New York, Hemisphere Publishing, 1987: pp. 227–276.
 - 35 Andersen K E, Maibach H I. Guinea pig sensitization assays: an overview. In: *Contact Allergy: Predictive Tests in Guinea Pigs*, Vol. 14, Andersen K E, Maibach H I (eds): Karger, Basel, Current Problems in Dermatology, 1985: pp. 263–290.
 - 36 Marzulli F N, Maibach H I. Contact allergy: predictive testing in humans. In: *Dermatotoxicology*, 3rd edition, Marzulli F N, Maibach H I (eds): New York, Hemisphere Publishing, 1987: pp. 319–340. (Alternative ref. Marzulli 1976.)
 - 37 Hjorth N. Diagnostic patch testing. In: *Dermatotoxicology*, 3rd edition, Marzulli F N, Maibach H I (eds): New York, Hemisphere Publishing, 1987: pp. 307–317.
 - 38 McNamee P M, Api A M, Basketter D A, Gerberick G F, Gilpin D A, Hall B M, Jowsey I, Robinson M K. A review of critical factors in the conduct and interpretation of the human repeat insult patch test. *Reg Tox Pharm* 2008; **54**: 24–34.
 - 39 OECD. OECD guideline for the testing of chemicals 429, 22nd July 2010. Skin Sensitization: Local Lymph Node Assay, 2010. Available at: <http://ecb.jrc.ec.europa.eu/testing-methods/annex5/> (last accessed 29 October 2010).
 - 40 Farage M A. Development of a modified forearm controlled application test method for evaluating the skin mildness of disposable wipe products. *J Cosmet Sci* 2000; **51**: 153–167.
 - 41 Lukacovic M F, Dunlap F E, Michaels S E, Visscher M O, Watson D D. Forearm wash test to evaluate the mildness of cleansing products. *J Soc Cosmet Chem* 1988; **3**: 355–366.
 - 42 Finkelstein P, Laden K, Meichowski W. New methods for evaluating cosmetic irritancy. *J Invest Dermatol* 1963; **40**: 11–14.
 - 43 Hannuksela M, Heikki S. The repeated open application test (ROAT). *Contact Dermatitis* 1986; **14**: 221–227.
 - 44 Andersen K E. Reproducibility of the chamber scarification test. *Contact Dermatitis* 1996; **43**: 181–184.
 - 45 Farage M A. The behind-the-knee test: an efficient model for evaluating mechanical and chemical irritation. *Skin Res Technol* 2006; **12**: 73–82.
 - 46 Bannan E A, Griffith J F, Nusair T L, Sauers L J. Skin testing of laundered fabrics in the dermal safety assessment of enzyme-containing detergents. *J Toxicol Cut Ocul Toxicol* 1992; **11**: 327–339.
 - 47 Rodriguez C, Calvin G, Lally C, Lachapelle J M. Skin effects associated with wearing fabrics washed in commercial laundry detergents. *J Toxicol Cut Ocul Toxicol* 1994; **13**: 39–45.
 - 48 Frosch P J, Kligman A M. The soap chamber test; a new method for assessing the irritancy of soaps. *J Am Acad Dermatol* 1979; **1**: 35–41.
 - 49 Frosch P J, Kligman A M. The chamber-scarification test for irritancy. *Contact Dermatitis* 2; 314–324.
 - 50 Falk M, Anderson C. Reliability of self-assessed reading of skin tests: a possible approach in research and clinical practice? *Dermatol Online J* 2010; **16**: 4. Available at: http://dermatology.cdlib.org/1602/originals/skin_tests/falk.html (last accessed 2 April 2010).
 - 51 Ivens U, Serup J, O'goshi K. Allergy patch test reading from photographic images: disagreement on ICDRG grading but agreement on simplified tripartite reading. *Skin Res Technol* 2007; **13**: 110–113.
 - 52 Uter W, Becker D, Schnuch A, Gefeller O, Frosch P J. The validity of rating patch test reactions based on digital images. *Contact Dermatitis* 2007; **57**: 337–342.
 - 53 Lock-Andersen J, Wulf H C. Threshold level for measurement of UV sensitivity: reproducibility of phototest. *Photodermatol Photoimmunol Photomed* 1996; **12**: 154–161.
 - 54 Griffiths H A, Wilhelm K-P, Robinson M K, Wang X M, McFadden J, York M, Basketter D A. Interlaboratory evaluation of a human patch test for the identification of skin irritation potential/hazard. *Food Chem Toxicol* 1997; **35**: 255–260.
 - 55 Basketter D, Reynolds F, Rowson M, Talbot C, Whittle E. Visual assessment of human skin irritation: a sensitive and reproducible tool. *Contact Dermatitis* 1997; **37**: 218–220.
 - 56 Willis C M, Stephens C J, Wilkinson J D. Assessment of erythema in irritant contact dermatitis. Comparison between visual scoring and laser Doppler flowmetry. *Contact Dermatitis* 1988; **18**: 138–142.
 - 57 Seidenari S, Belletti B. The quantification of patch test responses: a comparison between echographic and colorimetric methods. *Acta Derm Venereol* 1998; **78**: 364–366.
 - 58 Agner T, Serup J. Sodium lauryl sulphate for irritant patch testing – a dose–response study using bioengineering methods for determination of skin irritation. *J Invest Dermatol* 1990; **95**: 543–547.
 - 59 Lahti A, Kopola H, Harila A, Myllylä R, Hannuksela M. Assessment of skin erythema by eye, laser Doppler flowmeter, spectroradiometer, two-channel erythema meter and Minolta chroma meter. *Arch Dermatol Res* 1993; **285**: 278–282.
 - 60 Wahlberg J E. Erythema-inducing effects of solvents following epicutaneous administration to man – studied by laser Doppler flowmetry. *Scand J Work Environ Health* 1984; **10**: 159–162.
 - 61 Babulak S W, Rhein L D, Scala D D, Simion F A, Grove G L. Quantification of erythema in a soap chamber test using the Minolta chroma (reflectance) meter: comparison of instrumental results with visual assessments. *J Soc Cosmet Chem* 1986; **37**: 475–479.
 - 62 Zuang V, Archer G, Rona C, Vignini M, Mosca M, Berardesca E. Predicting visual assessment of allergic patch test reactions by non-invasive measurements. *Skin Pharmacol Appl Skin Physiol* 2000; **13**: 39–51.
 - 63 Fluhr J W, Kuss O, Diepgen T, Lazzerini S, Pelosi A, Gloor M, Berardesca E. Testing for irritation with a multifactorial approach: comparison of eight non-invasive measuring techniques on five different irritation types. *Br J Dermatol* 2001; **145**: 696–703.
 - 64 Magnusson B M, Koskinen L D. Effects of topical application of capsaicin to human skin: a comparison of effects evaluated by visual assessment, sensation registration, skin blood flow and cutaneous impedance measurements. *Acta Derm Venereol* 1996; **76**: 129–132.
 - 65 Ollmar S, Nyren M, Nicander I, Emtestam L. Electrical impedance compared with other non-invasive bioengineering techniques and visual scoring for detection of irritation in human skin. *Br J Dermatol* 1994; **130**: 29–36.
 - 66 Fullerton A, Rode B, Serup J. Skin irritation typing and grading based on

- laser-Doppler perfusion imaging. *Skin Res Technol* 2002; **8**: 23–31.
- 67 Wigger-Alberti W, Hinnen U, Elsner P. Predictive testing of metalworking fluids. A comparison of 2 cumulative human irritation models and correlation with epidemiological data. *Contact Dermatitis* 1997; **36**: 14–20.
- 68 Held E, Lorentzen H, Agner T, Menné T. Comparison between visual score and erythema index (DermaSpectrometer) in evaluation of allergic patch tests. *Skin Res Technol* 1998; **4**: 188–191.
- 69 Andersen K E, Straberg B. Quantification of contact allergy in guinea pigs by measuring changes in skin blood flow and skin fold thickness. *Acta Derm Venereol* 1985; **65**: 37–42.
- 70 Breternitz M, Fluhr J W, Berardesca E. Technical bases of biophysical instruments used in sensitive skin testing. In: *Sensitive Skin Syndrome*, Berardesca E, Fluhr J W, Maibach H I (eds): New York, Taylor & Francis, 2006: pp. 75–106.
- 71 Farage M A. Are we reaching the limits or our ability to detect skin effects with our current testing and measuring methods for consumer products? *Contact Dermatitis* 2005; **52**: 297–303.
- 72 Farage M A. Enhancement of visual scoring of skin irritant reactions using cross-polarized light and parallel-polarized light. *Contact Dermatitis* 2008; **58**: 147–155.