

## **Isolated Chicken Eye Test Method**

***[This Page Intentionally Left Blank]***

## II. ISOLATED CHICKEN EYE TEST METHOD

### 1.0 ICE TEST METHOD RATIONALE

The Isolated Chicken Eye (ICE) test method is being evaluated for its ability to identify ocular corrosives and severe irritants as defined by the GHS (UN 2003), the EPA (1996), and the EU (2001) classification systems. Dose selection is not relevant to the assay as the test substance is typically applied neat in either liquid or solid (pulverized) form. Three measurements are made during the course of the test: one objective measurement (corneal thickness/swelling) and two subjective measurements (corneal opacity, fluorescein dye retention). Corneal opacity is the only common endpoint shared between the ICE test and the *in vivo* rabbit eye test.

### 1.1 SCIENTIFIC BASIS FOR THE ICE TEST METHOD

#### 1.1.1 Mechanistic Basis of the ICE Test Method

The ICE is an organotypic model that provides short-term (4 hours) maintenance of the whole eye. The ICE was developed as a modification of the IRE test method and was intended as a screening assay to identify the ocular corrosive and severe irritation potential of products, product components, individual chemicals, or substances. Substances that are predicted by ICE as corrosives or severe irritants could be classified as GHS Category 1, EU R41, or EPA Category 1 eye irritants without the need for animal testing. Substances that are negative in ICE would undergo further testing to confirm that they are not false negatives or to determine if they are mild to moderate ocular irritants. The ICE test method may also be useful as one of several tests in a battery of *in vitro* eye irritation methods that collectively predicts the eye irritation potential of a substance *in vivo*.

The mechanistic basis for ocular irritation in the ICE is not known, and it is unclear if similar effects occur in the chicken relative to the rabbit (or human). Essentially, the ICE test method was designed by manipulating a number of free parameters, such as rate, time, and amount of test chemical exposure so that the outcome matches that of the *in vivo* rabbit eye test system. Because the primary concern is an accurate correlation to the ocular irritancy classification of a test substance, the ICE test does not necessarily have to be mechanistically based. Therefore, a clear understanding of the mechanistic basis of the assay may not be required prior to using the ICE test. However, the ICE BRD should contain a discussion of cellular mechanisms of corrosion and severe irritation and their relevance to *in vitro* testing.

#### 1.1.2 Advantages and Limitations of Mechanisms/Modes of Action of the ICE Test Method

The endpoints in the ICE test measure:

- integrity of the epithelial and (to a lesser extent) endothelial barrier function, which on the corneal surface is maintained primarily by the intercellular junctions of the most superficial layer of surface epithelial cells, by measuring corneal thickness and fluorescein penetrability of the stroma; and
- stromal edema and/or physical alteration of epithelial cells, stromal keratocytes, collagen, or extracellular matrix that alter transparency.

These endpoints correspond to the nonspecific opacification of the cornea utilized in the Draize rabbit eye test. The Draize test provides data on the conjunctival, anterior chamber, and iris responses (including the vascular response) that are not accounted for in the ICE test method. Very importantly, the ICE (and other *in vitro* organotypic ocular irritation test methods) does not include the tear film, and tears are an essential component of normal surface physiology and protection. A common limitation to all ocular irritancy test methods is that they do not allow definition of the mechanism of corneal opacification (i.e., edema versus coagulation versus infiltration).

Corneal swelling is an endpoint measured in the ICE test method, but the ICE BRD fails to state that corneal swelling can result from two sources: damage to the endothelium and damage to the epithelium. While it has been shown that epithelial damage induces corneal swelling very rapidly in the rabbit, damage to the endothelium is likely to take longer. However, swelling due to mild epithelial damage is not serious and after several hours to a day may resolve. Therefore, this measurement does not provide much information as to actual damage because of the short-term observation duration (4 hours) of the model.

The conjunctiva of the mammalian eye is generally similar across species in that it is a delicate supporting epithelium comprising most of the ocular surface; the cornea cannot survive without the conjunctiva. The conjunctiva, as compared to the cornea, is more permeable. The vascular bed is a major site of the release of immune function cells that can participate in ensuing inflammation. Moreover, these effects may be expected on a longer time scale and the four-hour observation time for ICE may be too short to observe the maximal effects of substances that act through mediators. This would suggest another wide departure from the *in vivo* rabbit eye as inflammation of the ocular surface and loss of conjunctival support would result in additional stress on the cornea and therefore increase the likelihood of adverse effects.

#### 1.1.3 Similarities and Differences of Mechanisms/Modes of Action and Target Tissues Between the ICE Test Method and Humans and Rabbits

The short discussion in the ICE BRD of the mammalian eye includes a section about the differences between the human and rabbit eye. *In vivo*, the rabbit eye is more sensitive to some irritants, while the reverse is true for other irritants. While much is known about the anatomy of the human and rabbit eye, the relationship between species differences in eye anatomy and physiology and the sensitivity to ocular irritants has not been clearly established. However, historical use of the rabbit eye test in regulatory applications has made the Draize rabbit eye test a suitable animal model for the evaluation of irritation potential of substances in the human eye.

The chicken eye has not been studied as intensively as the rabbit eye, but it is clear that the basic anatomy and structure of the chicken eye is markedly different from the human, although the structure of the cornea is relatively similar. Little is known as to the biochemistry of the cornea of the chicken and the comparison with the mammalian cornea. It is also a concern that the human and rabbit cornea differ in their structure. The ICE BRD needs to point out that the cornea has two important properties for vision: 1) that it is transparent; and 2) that, as the major refracting element in the optical path, it needs to have a smooth anterior surface and an appropriate index of refraction.

While some of the species differences are mentioned in the BRD, they are not well related to the problems at hand. Bowman's layer, found in the human eye just under the epithelium, is also found in the chicken eye, but not in the rabbit eye. Descemet's layer is mentioned but probably has little to do with the chemical response. Both young and old rabbits have the ability to regenerate the endothelium, a property seen in most species (with the exception of primates). Differences in the types of collagen found in the stroma in the rabbit and human may be a source of concern. Certainly, mechanically, the corneas of rabbits and humans are different, but this is not known for the chicken. The two types and sources of edema (i.e., epithelial and endothelial damage) are not mentioned in the ICE BRD, nor is it possible to find information on the time course for edema in the rabbit eye. This could be revealing information as it could suggest that the residual protective tear film is more easily washed off the isolated chicken eye, while the rabbit blinks less than the human and probably has a tear film more resistant to evaporation. Once the tear film is removed (as the constant drip of isotonic saline will probably do), the epithelium will become more vulnerable to chemicals.

The BRD does point out that the four-hour study duration may be a limitation of ICE and that solid or adherent chemicals may not be reliably tested. However, the contribution of the conjunctiva to corneal viability, and corneal effects associated with conjunctival damage, are not fully realized in the ICE test method. *In vivo*, the rabbit, as well as the human, also has intraocular damage, inflammation, and iridial effects measured, but none of these measurements are possible with the ICE model.

#### 1.1.4 Mechanistic Similarities and Differences Between the ICE Test Method, the *In Vivo* Rabbit Eye Test Method, and/or Human Chemically-Induced Eye Injuries

There are many data gaps between the ICE test method and the current *in vivo* rabbit eye test (also in regard to human chemically induced eye injuries). The ICE test method is being evaluated for its ability to identify ocular corrosives or severe irritants, as required for hazard classification according to the EPA (1996), EU (2001), and GHS (UN 2003) classification systems. As such, its use has the potential to refine or reduce animal use in eye irritation testing and to spare animals from the extreme pain caused by the placement of corrosive agents onto the eyes. Because the accuracy of the ICE test method and limitations for predicting specific chemical and/or product classes are not known due to the lack of comparative data with humans, the potential of this method to improve prediction of adverse health effects in humans is unknown.

## 1.2 **Regulatory Rationale and Applicability**

### 1.2.1 Similarities and Differences Between Endpoints Measured in the ICE Test Method and the *In Vivo* Rabbit Eye Test Method

Differences between the chicken and mammalian eye are discussed. The differences between the ICE test method and the *in vivo* rabbit eye test include:

- ICE evaluates only corneal effects and does not account for effects on the iris and conjunctiva, including the limbal stem cell population.
- ICE does not account for the reversibility of corneal effects.
- ICE does not account for systemic effects.
- ICE is a short-term test and many not identify slow-acting irritants.

In addition, the current *in vivo* test method observes rabbits for up to 21 days after treatment to assess the reversibility of observed endpoints or persistence of damage. The ICE can only observe effects for four hours after treatment. Therefore, the potential reversibility of the affected endpoint beyond four hours or an effect with a delayed onset cannot be adequately evaluated with the ICE test.

### 1.2.2 Suggestions Regarding Other Evidence that Might be Used in a Tiered Testing Strategy

Information on pH, concentration, osmolality, and chemical structure and its correlation to available *in vivo* results could be used in a weight of evidence approach to provide some degree of predictability of irritancy potential.

## 2.0 TEST METHOD PROTOCOL COMPONENTS

### 2.1 Description and Rationale of the Components for the Recommended ICE Test Method Protocol

#### 2.1.1 Materials, Equipment, and Supplies

This procedure has been modified only slightly since its inception and seems to have been used in very few laboratories. The extent of damage to the isolated chicken eye following exposure to a chemical substance is measured by corneal swelling (as determined optically), corneal opacity (also determined with a slit-lamp examination using the area of the cornea most densely opacified), and fluorescein retention. The latter two measurements are subjective.

Seven-week-old spring chickens are the source of the eyes in the ICE test. The facility should be located in proximity to the laboratory such that the chicken heads can be transferred and processed within two hours after the birds are killed. Because baseline fluorescein retention and corneal thickness measurements are conducted to verify the integrity of the test eyes, longer transport times could be evaluated for feasibility for inclusion in the protocol.

Intact heads are transported to the laboratory at ambient temperature in plastic boxes humidified with tissues moistened with isotonic saline or water. The number of heads needed for a single assay should be determined by the historical rate of rejection of eyes for the ICE test (8% to 45% based on six to ten heads necessary to obtain 11 useable eyes [Prinsen M, personal communication]) and number of samples to be tested (i.e., at minimum, one test substance, one positive control, and one negative control tested in triplicate, or nine eyes).

The details for inspection of each eye and further dissection of the eye are adequately described. Each accepted eye is positioned in a clamp and transferred to the superfusion apparatus. The entire cornea is supplied with isotonic saline at a rate of 2-3 drops/minute at  $32 \pm 1.5^\circ\text{C}$ . Consideration might be given to other “bathing” solutions and rate of superfusion to determine if these factors would improve the overall performance of the method (See **Section II - 2.1.3**). After placement into the apparatus, the corneas are again examined with the slit-lamp to ensure no corneal damage during dissection. The basis of rejection or replacement of eyes is described. The eyes are equilibrated prior to dosing for 45 to 60 minutes. An attempt should be made to randomize the selection of eyes for the test. Alternating the position of the eye in the apparatus

(similar to what has been described [Prinsen M, personal communication]) seems to be a reasonable approach (i.e., Sample # 1: positions 1, 4, and 7; Sample #2: positions 2, 5, and 8; Sample #3: positions 3, 6, and 9).

Two major obstacles appear in the conduct of the ICE test: 1) differences in slit-lamp systems (including examiners) to measure corneal swelling; and 2) the limitations of the custom-built stainless steel eye clamps for the superfusion apparatus in terms of the maximum number of eyes that can be evaluated at the same time (i.e., 11 eyes). Corneal swelling values for test substances may vary based on differences in the slit-lamp system used. In order to compare ICE test data from different laboratories, a “correction factor” may be required to compensate for these differences (i.e., ranking of substances according to corneal swelling figures should be similar, regardless of the apparatus). The potential impact of this issue has not been resolved to date and should be the focus of a pre-validation study. The ability to test only 11 eyes at the same time severely limits the number of samples tested concurrently. Given that three replicate eyes for each treatment group (test substance, positive control, negative control) are needed for an experiment, nine eyes would be required. If the apparatus could be modified to 12 clamps, another test substance or a benchmark substance could then be included in the experiment. As recommended in the ICE BRD, the basic protocol should include a provision to repeat each test (e.g., when equivocal test results are obtained) and clarify how these additional data would be used for classification.

There are some additional concerns:

- The temperature is not well controlled which could adversely affect cell metabolism, and the drip system is very difficult to adjust to ensure that the whole cornea is superfused properly
- The number of replicate eyes is small ( $n = 3$ ), making meaningful statistical analyses unlikely. However, it is not known if including additional eyes would result in enhanced performance of the ICE test because a formal evaluation of the optimum number of eyes for inclusion has not been performed.
- It is suggested that the chambers be moved to a horizontal position, which would ensure that the whole cornea is superfused adequately and allow the test substances to be applied without removing the eyes from the apparatus. This could also improve the consistency of data collected by allowing for a more accurate approximation of exposure time (e.g., the potential variability resulting from removing and returning the eyes from the apparatus during dosing is significant, as a precise 10-second exposure would be very difficult under these conditions).
- Reference substances (negative and positive controls, benchmarks) that are part of the performance standards developed for the validated test method should be identified.

### 2.1.2 Dose-Selection Procedures

Dose selection procedures are not relevant to the ICE test as a liquid substance is applied neat at 0.03 mL and a solid is applied at 0.03 g after grinding it into a fine powder.

### 2.1.3 Endpoint(s) Measured

Control and test eyes are examined pre-treatment and at 30, 75, 120, 180 and 240 minutes after a 10-second treatment, using corneal opacity, swelling, fluorescein retention, and morphology (on a case-by-case basis) as endpoints. Subjective measurements such as corneal opacity and fluorescein retention can vary from scorer to scorer and therefore, within a study, one individual would need to perform all of the measurements. Sufficient training is needed to acquire these measurement skills. The term “fluorescein retention” seems inappropriate as once the fluorescein moves into the cornea, it continues to diffuse into the anterior chamber of the eye. Fluorescein penetration would be facilitated by the isotonic drip as the pH is different from physiological values (i.e., isotonic saline is slightly acidic). Furthermore, the lack of divalent ions in isotonic saline can disrupt cell-cell adhesion by opening up tight junctions, causing the cells to increase in permeability or slough off of the corneal surface. Therefore, a balanced salt solution (e.g., Hank’s Balanced Salt Solution; Ringer’s Solution) would be more appropriate as an assay medium. The fluorescein measurements would be aided by the use of an automated mechanical system (e.g., sensor system) that could detect variations in fluorescein staining more accurately and quantitatively than the naked eye.

### 2.1.4 Duration of Exposure

The test substance is applied for 10 seconds and subsequently rinsed from the eye with 20 mL isotonic saline at ambient temperature. However, because of the required manipulation of the eyes prior to dosing, the 10-second application time appears to be just an estimate of the true contact time. Details of this procedure are described in the ICE BRD. The time of application was chosen based on the IRE study design to discriminate between irritant and non-irritant substances. This brief exposure time appears adequate based on use in a limited number of laboratories, but it may be unsatisfactory if a larger number of laboratories conduct the assay. Some consideration for extended exposure times, where extremes in variability among laboratories could be reduced, could be useful.

### 2.1.5 Known Limits of Use

Studies indicate that the ICE test method is amenable to use with a broad range of solid and liquid substances with a few limitations. However, substances that are poorly soluble or those materials that run off corneal surfaces may not be compatible with this test. Test limitations are described for hydrophobic compounds (inadequate contact with cornea) and solids that adhere to the corneal surface. Modifications to the basic protocol would require optimization to ensure accurate results for such test substances. Previous studies have shown that a number of surfactants or formulations containing surfactants, along with some solid substances, appear to be underpredicted by the ICE test method while some alcohols may be overpredicted. These limitations may place restrictions on the applicability of the method across chemical classes.

### 2.1.6 Nature of the Response(s) Assessed

The data collected in this assay are both qualitative and quantitative. If morphological and histopathological examinations are performed, descriptive data would be included. The focus on corneal effects in the ICE test appears to limit its application to predicting corrosives and severe irritants only.

### 2.1.7 Appropriate Controls and the Basis for Their Selection

Negative controls (usually isotonic saline, distilled water, or appropriate solvent) should be run concurrently with the positive control and the test substance. The positive control is used to test the limits of the experiment and help to develop a historical database. None of the published ICE protocols recommend the use of a concurrent positive control. However, a substance classified as a GHS Category 1 (UN 2003) (e.g., 10% acetic acid) should be included in each experiment, with three eyes tested. A positive control will demonstrate the functional adequacy of the test method and the consistency of laboratory operations in accurately identifying ocular corrosives and severe irritants. Benchmark controls should be included when testing chemicals of a specific class with consideration of structural and functional similarity. It would be useful to have a system where the eyes used for the controls were spread throughout the superfusion apparatus such that the replicate eyes are randomly placed so that order effects in dosing would be less likely.

### 2.1.8 Acceptable Range of Control Responses

The negative and/or solvent control should produce an irritancy classification that falls within the nonirritating classification. If not, the experiment may need to be discarded or an alternative solvent (i.e., one that would produce a nonirritating classification) used. The positive control test substance should produce an irritancy classification that corresponds to the anticipated irritancy response (i.e., ocular corrosive/severe irritant), based on the known classification of the test substance in the *in vivo* rabbit eye test. Benchmark controls should produce an irritation response that is within acceptable limits and may be useful for demonstrating that the test method is functioning properly for detecting the ocular irritating potential of chemicals within a specific class.

### 2.1.9 Nature of the Data to be Collected and the Methods Used for Data Collection

The data collected include: 1) measurement of corneal swelling with a slit-lamp microscope and expressed as a percentage ( $[\text{corneal thickness at time } t - \text{corneal thickness at time } 0] / \text{corneal thickness at time } 0 \times 100$ ); 2) corneal opacity using the area of the cornea most densely opacified for scoring (scores ranging from 0 to 4); and 3) fluorescein retention calculated for the 30 minute observation time point only (scores ranging from 0 to 3). Morphological effects may also be examined on a case-by-case basis and could include pitting of epithelial cells, loosening of the epithelium, and roughening of the corneal surface. Corneal thickness is an objective measurement that requires either a slit-lamp microscope equipped with an optical pachymeter or an ultrasonic pachymeter. The severity of each endpoint, indicative of corneal damage, should be documented at each time point (except fluorescein retention) with a slit-lamp microscope.

### 2.1.10 Type of Media in Which Data are Stored

There are no concerns with regard to this section of the ICE BRD.

### 2.1.11 Measures of Variability

There are no concerns with regard to this section of the ICE BRD.

#### 2.1.12 Statistical or Nonstatistical Methods Used to Analyze the Resulting Data

The level of severity for each study endpoint (corneal swelling, opacity, and fluorescein retention) recorded at each time point can be used to calculate the maximum mean score<sup>2</sup> for each endpoint from which an irritation index can be determined. This index, along with the individual maximum mean scores for each ICE test method endpoint, can be used in a comparison to a numerical *in vivo* score. However, there does not appear to be a rationale for the current method employed for normalizing the data when calculating the Irritation Index. Rather than multiplying the maximum opacity and fluorescein retention measurements by the historical equalizing value of 20, one could simply adjust the current data to cover the same range.

While the irritation index has been used to correlate ICE results to various *in vivo* endpoints/scores, only the ICE categorization scheme (described in Section 2.2.13 of the ICE BRD) has been used as a predictive tool to assign an irritancy classification.

#### 2.1.13 Decision Criteria and the Basis for the Algorithm Used

In defining the irritancy classification, various combinations of the endpoint scores (i.e., the ICE categorization scheme) are considered. This scheme has been correlated to the EU regulatory classification system for comparison to *in vivo* results. Although this approach may correlate with the rabbit *in vivo* data, it is not clear if there are any real tissue change parallels between the ICE test and *in vivo* rabbit eye test data. Histopathology may be warranted in order to discriminate between effects that are on the borderline of severe and moderate irritation.

#### 2.1.14 Information and Data that Will Be Included in the Study Report

Conduct of the ICE test should follow GLP guidelines for recognized rules designed to ensure high-quality laboratory records. Individual measurements should be reported using the sample scoring sheet provided in Figure 2-4 of the ICE BRD. The raw values are most likely asymmetric and therefore standard deviations are of limited value in characterizing their distribution.

### **2.2 Basis for Selection of the Test Method System**

There are no concerns with regard to this section of the ICE BRD.

### **2.3 Identification of Proprietary Components**

There are no concerns with regard to this section of the ICE BRD.

### **2.4 Numbers of Replicate and/or Repeat Experiments for Each Test**

Historically, only a single negative control eye has been used in each test. In Balls et al. (1995), the number of chicken eyes evaluated per test substance was reduced from five to three, which was purported to have no effect on accuracy (Prinsen M, personal communication). However, such a small number provides little information on between eye response variability, and the predictive value of the test may be diminished by using only three eyes to detect a severe

---

<sup>2</sup> ICE endpoint measurements are averaged at each time point across the three test eyes. The mean value for each endpoint that is the greatest at any time point (maximum mean value) is used for categorization.

reaction. Since the most appropriate number of eyes that would result in optimum performance is not known, it would appear suitable to use known irritants to examine the effect of the number of eyes on prediction consistency and accuracy. Some basic probability estimates of the tradeoffs involved with multiple eyes will provide useful information.

Indirectly related to the number of eyes is the variability that would be inherent to the somewhat uncontrolled methodology by which the eyes are harvested and utilized.

## **2.5 Study Acceptance Criteria for the ICE Test Method**

Currently, the single criterion for an acceptable test is that the negative control gives an irritancy classification that falls within the nonirritating classification. If a modified ICE test method protocol is proposed to include concurrent positive and negative control responses (as is recommended in the ICE BRD), the positive control should also be included in the criteria for an acceptable test. Inclusion of these controls could also provide an indication as to the adequacy of the number of eyes that are included for each test substance.

## **2.6 Basis for any Modifications made to the Original ICE Test Method Protocol**

There does not appear to have been a formal evaluation performed on the effects of reducing the number of eyes per test substance from five to three. It is not clear if such a reduction adversely affects the performance of the ICE test.

## **2.7 Adequacy of the Recommended Standardized Protocol Components for the ICE Test Method**

The proposed ICE protocol provided in Appendix A of the ICE BRD deviates very little from the original protocol with the exception that a concurrent positive control substance and, if appropriate, a benchmark substance is to be included in each test, with three eyes to be used for each treatment group (test substance; negative and positive controls; benchmarks, if included).

However, before the recommended protocol is adopted, several aspects of the test should be considered for optimization of the method. Some of these issues are addressed in the ICE test method protocol components. The following questions should be addressed in future optimization studies:

- How can the different corneal swelling values for test substances from different laboratories be resolved to avoid applying a correction factor to compare results?
- Can the custom superfusion apparatus be modified to accommodate at least 12 eyes in order to test two test substances (or one test substance plus a benchmark) along with negative and positive controls simultaneously without adversely affecting results? For example, given the additional time requirements that would be required by adding additional eyes, could all of the necessary measurements with 12 eyes be made? Furthermore, would the time required to harvest 12 eyes as opposed to only 10 eyes (as is current practice) adversely affect the integrity of the eyes?

- The specifics of how the eyes will be randomized in the clamps should be identified. Alternating the position of the eye in the apparatus seems to be a reasonable approach (i.e., Sample #1: positions 1, 4, and 7 in the superfusion apparatus; Sample #2: positions 2, 5, and 8; Sample #3: positions 3, 6, and 9; similar to current practice [Prinsen M, personal communication]).
- What effect, if any, does the bathing solution or rate of drip have on the system? Would a solution containing electrolytes be better than isotonic saline (see **Section II - 2.1.3**)?

In addition, the protocol must make it clear that a minimum test includes a test substance and positive and negative controls, each performed using three eyes. Records should be kept for the rate of rejection of eyes for each test. Histopathology, including determination of the depth of injury, may be considered when the standard ICE endpoints (i.e., corneal opacity, swelling, and fluorescein retention) produce borderline results. The selection of a positive control substance should be based on the best historical control data in terms of the magnitude of the severe response desired. If a benchmark substance is used, the reason for its use should be specified.

The ICE test method has limitations but it appears to successfully identify many ocular corrosives and severe irritants that would eliminate subsequent testing in a live animal.

### **3.0 SUBSTANCES USED FOR PREVIOUS VALIDATION STUDIES OF THE ICE TEST METHOD**

#### **3.1 Substances/Products Used for Prior Validation Studies of the ICE Test Method**

The three ICE validation studies considered in the BRD utilized a spectrum of organic and inorganic substances that adequately covered the range of irritancy responses. Among these studies, 121 substances were evaluated which likewise is a reasonable number for assessing the validation status of this test method; the ICE methodology used was similar among the three studies although one study (Balls et al. 1995) incorporated results obtained in four different laboratories.

#### **3.2 Coding Procedures Used in the Validation Studies**

Balls et al. (1995) was the only study that made reference to the use of coded substances. Use of coding eliminates bias especially where subjective interpretation is involved (e.g., scoring effects in the Draize test; grading opacification in the ICE test). However, for the purposes of a retrospective evaluation, lack of coding does not appear to be justification for rejecting the data.

### **4.0 *IN VIVO* REFERENCE DATA USED FOR AN ASSESSMENT OF TEST METHOD ACCURACY**

This section provided a detailed analysis of the published *in vivo* methods used to evaluate ocular irritancy and/or corrosivity. The regulatory schemes for interpreting such *in vivo* data were provided.

#### **4.1 In Vivo Rabbit Eye Test Method Protocol(s) Used to Generate Reference Data**

The *in vivo* rabbit eye test method protocol(s) used to generate the reference data considered in the three validation studies were appropriate.

#### **4.2 Interpretation of the Results of the In Vivo Rabbit Eye Tests**

The interpretation of the results of the *in vivo* rabbit eye tests was correct. The *in vivo* methods described have been judged by the agencies using these methods as suitable for their regulatory needs. The concern can reasonably be raised that these regulatory classification methods may be less than adequate for use in evaluating or making distinctions between *in vitro* methods and their suitability for chemical or product class evaluations.

#### **4.3 In Vivo Rabbit Eye Test Data Quality with Respect to Availability of Original Study Records**

In the case of the ICE test method, original study records were not available for any of the reports evaluated. However, a lack of original study records does not necessarily raise concerns about a study. As long as an evaluation of the results can be made and the quality of the study otherwise appears to be adequate (as is the case for the studies evaluated in the ICE BRD), the study should be used. Future validation studies should be conducted under GLP compliance and original study records should be readily available.

#### **4.4 In Vivo Rabbit Eye Test Data Quality with Respect to GLP Compliance**

The criteria used in selecting substances in two of the three validation studies for the ICE test method cited in the BRD were not specified. The Balls et al. (1995) study included the criterion that the *in vivo* data were from GLP-compliant, post-1981 studies, and were conducted in accordance with OECD TG 405 (OECD 1987).

However, as the GLP regulations do not deal with the actual performance of the tests as much as with background documentation, a distinction in the weight given to GLP-compliant versus non-GLP-compliant studies in the ICE BRD may not be necessary. According to the current EU and OECD documents on the validation of toxicity tests, when the basic requirements of the GLP procedure (the “spirit” of GLPs) have been implemented in a study, lack of complete/formal GLP compliance is not an adequate criterion to exclude *in vivo* or *in vitro* data from the evaluation of the performance of a toxicity test.

#### **4.5 Availability of Relevant Human Ocular Toxicity Information**

The small set of human data, whether from accident reports or controlled human studies is of little value in examining the performance of an *in vitro* test. Appropriately, the discussion of this topic is quite limited. Very little human ocular injury data exist and most of the available information originates from accidental exposure for which the dose and exposure period were not clearly documented. Accidental exposures have no measure of dose and typically, even if the individual is seen in a clinical setting, there is no “scoring” or time course data. However,

there still needs to be greater effort to obtain and consider information on human topical ocular chemical injury.

#### 4.6 Accuracy and Reliability of the *In Vivo* Rabbit Eye Test

There should be more discussion in the ICE BRD of the variability of the rabbit data. This is particularly important in the determination of the accuracy of an *in vitro* test method. While there are often multiple results for each *in vitro* determination of irritation potential, there is generally only one *in vivo* test result. Because of the known variability in the rabbit eye test, it is not possible from the data presented to determine if the inconsistencies between ICE and the *in vivo* rabbit eye tests are due to “failure” of the *in vitro* test method or a misclassification by the single *in vivo* result provided.

However, data on the reproducibility or reliability of the *in vivo* rabbit eye test do exist in the literature, most notably the intra- and inter-laboratory study published by Weil and Scala (1971), as well as Kaneko (1996) and Ohno et al. (1999). Using a fixed protocol and a single supply of chemical agents tested in 25 laboratories, these investigators identified “good” laboratories as those that had the lowest variance in ranking of irritancy using a sum of ranks statistical measure. They also found that nonirritants provided little useful information on laboratory performance. GLP regulations were not in place at the time of this study, but are not thought to be critical in the evaluation of the data.

In the development of alternative methods to intact animal testing, the question always arises regarding the quality of reference *in vivo* test data used to evaluate or validate the newer, alternative *in vitro* test method. These questions typically center on two major concepts. The first is the availability of a “gold standard” for measuring the intended effect. The second is the reliability (intralaboratory repeatability and reproducibility; interlaboratory reproducibility) of the *in vivo* test. With respect to ocular injury (irritation or corrosion), there is no “gold standard” (i.e., there is no set of substances that have been shown, regularly and reproducibly, in any competent laboratory, to produce a particular degree of irritancy or damage in the *in vivo* rabbit eye test). Consequently, the evaluation (or acceptability) of an alternative test method is unavoidably biased by the selection of the *in vivo* reference data used in that evaluation.

While any repeat performance of *in vivo* rabbit eye irritancy testings or testing of known corrosives or severe irritants should be discouraged, it is important to have available multiple *in vivo* rabbit eye test data that demonstrate reproducible results. Any optimization and validation studies should use existing animal data, if available. Additional animal studies should only be conducted if important data gaps are identified and such studies should be carefully designed to maximize the amount of pathophysiological (e.g., wound healing) information obtained.

The discordance in MAS scores calculated for the same substance among different laboratories has been documented (Spielmann 1996). Based on data in the Weil and Scala (1971) intra- and inter-laboratory study, Spielmann (1996) noted that three of the ten substances tested were classified anywhere from non-irritant (MAS scores < 20) to irritant (MAS scores > 60) when tested in 24 different laboratories.

It is well documented that the Draize eye test has a low variability at both ends of the MAS scale (e.g., the low end in the range of non-irritating chemicals and at the upper end of the scale in the range of severely eye irritating materials) (Kaneko 1996; Ohno et al. 1999). However, in the middle range, the variability is very high (as indicated by the high CV and SD values for such substances in Balls et al. [1995]). Nevertheless, this range of variability may be considered insignificant for the purposes of this evaluation, since it is focused only on the detection of severe irritants.

When evaluating the performance of the ICE test method, the reliability of the Draize rabbit eye test data has to be considered. Therefore, how this aspect of the Draize eye test will be considered when attempting to determine the predictive value of the *in vitro* alternative needs to be defined prior to any evaluation. This important aspect has been cited as a reason why the replacement of the Draize eye test by *in vitro* tests has failed in the past. Although this has been well documented in the scientific literature (e.g., Figure 1 in Balls et al. [1995], in a review by Spielmann [1997]), additional discussion in the ICE BRD is warranted.

Not all substances evaluated in the BRD were tested concurrently in both the ICE test method and in the *in vivo* rabbit eye test. In addition, none of the substances were identified as having been tested in the *in vivo* rabbit eye test in multiple laboratories. It would seem that the entire effort to develop alternatives to intact animal testing for ocular effects would benefit from some attention to providing an approximation of a “gold standard”.

#### Minority Opinion

This section was approved by consensus of the Panel with a minority opinion from Dr. Martin Stephens that sufficient animal data are available for further optimization/validation studies and no further animal testing should be conducted (See Minority Opinion from Dr. Stephens in **Section II - 12.3**).

## **5.0 ICE TEST METHOD DATA AND RESULTS**

### **5.1 ICE Test Method Protocols Used to Generate Data Considered in the BRD**

The ICE test method protocols used in each of the published validation studies are described and are straightforward. Training is clearly required, as a great deal of operator evaluation is required for determination of fluorescein retention and corneal opacity, along with operation of the slit-lamp microscope for corneal thickness measurements. The preparation of the eyes also requires adequate training. Chemical contact with the eye and possible limitations with certain types of substances are discussed. Types of measurements are all described. The protocol used for each study is described and tables of the chemicals used in the studies are provided.

### **5.2 Comparative ICE Test Method–*In Vivo* Rabbit Eye Test Data Not Considered in the BRD**

The three reports that meet the requirements for inclusion in the ICE BRD provide limited rabbit comparisons. Additional comparative ICE - *in vivo* data do not appear to be available.

### **5.3 Statistical and Nonstatistical Approaches Used to Evaluate ICE Data in the BRD**

The approaches used to evaluate the ICE test method data appear to adequately describe its accuracy and reliability. However, given the unavailability of original ICE data, a definitive statement regarding the adequacy of these approaches is not feasible.

### **5.4 Use of Coded Substances, Blinded Studies, and Adherence to GLP Guidelines**

Although GLP conditions were used in each of the three validation studies, the details are vague. Coding of test substances was carried out in only Balls et al. (1995). However, as indicated in **Section II - 3.2**, the absence of coding is not an adequate justification for rejecting the data from these studies.

### **5.5 “Lot-to-Lot” Consistency of the Test Substances and Time Frame of the Various Studies**

The concentration tested was indicated in all three validation studies. The substances in Prinsen (1996) were presumed undiluted unless otherwise specified (e.g., as in Table 2 of Prinsen [1996]). The test substances and the concentrations used were adequately described in the ICE BRD. Based on the selection criteria for Balls et al. (1995), the chemicals used were of known high consistency and purity. However, given the lack of specifically cited selection criteria in Prinsen and Koëter (1993) and Prinsen (1996), an accurate assessment of lot-to-lot consistency was not feasible. Prinsen (1996) did indicate that the same batch of each test substance was used in both the ICE and *in vivo* test methods.

## **6.0 ICE TEST METHOD ACCURACY**

### **6.1 Accuracy Evaluation of the ICE Test Method for Identifying Ocular Corrosives and Severe Irritants**

Based on the three validation studies considered in the ICE BRD, the accuracy (concordance) of the ICE test was variable (71% to 100% with an overall rate of 82%, according to the GHS classification system). Likewise the false positive and negative rates were variable. However, comparisons between studies were difficult as the original data were not available and the studies were not designed for these later comparisons.

A false positive rate of 10% (0-18%, Tables 6-1 to 6-3 of the ICE BRD) would appear to be acceptable. It is not clear if using additional eyes per substance would further reduce this rate. With regard to hazard evaluation, the consequences of a false negative result (up to 40% in some studies) will be resolved because *in vivo* tests will then be conducted in a tiered testing approach. It also is important to know if additional eyes per test group (or any other methodological improvements) would reduce the false negative rate and thereby further reduce the number of animals tested.

The method appears to perform equally well for the three ocular irritancy classification systems. Similarities likewise occur in discordant substances.

Although the assessment of test method accuracy is an essential element of validation, it often cannot be assessed directly, in that human data are lacking. Consequently accuracy is assessed indirectly by comparison to data from the *in vivo* rabbit eye test. The use of terms such as “false negative” and “false positive” should be preceded by a discussion of the difference between a true reference standard (in this case human data) and a default reference standard (in this case animal data).

A comprehensive accuracy assessment in the absence of suitable human data should take into account the variability in the Draize test itself. Specifically, Draize test data should be analyzed to see how well the test predicts itself. Any test yields variable results, and Bruner et al. (1996) have shown that the Draize test has considerable variability, although this variability is least pronounced at the extremes of the irritation range (i.e., severe irritants/corrosives and nonirritants). Consequently, a chemical’s “true” Draize score can be thought of as a moving target, and it is in this light that the accuracy of ICE test and other potential alternatives should be judged. The ICE BRD mentions that a reliability analysis of the *in vivo* rabbit eye test is planned and will be distributed when completed. The absence of such an analysis in the BRD is a major stumbling block to a proper assessment of the ICE test method.

In addition to the analyses conducted, the Panel suggests an assessment based on ranking of experimental data for severity for both the *in vivo* rabbit eye test and the ICE test method using the proposed reference substances listed in Section 12.4 of the ICE BRD.

#### Minority Opinion

Drs. Martin Stephens and Peter Theran note that the term “accuracy” is used throughout the four BRDs and this Panel Report to address the degree of consistency between the *in vivo* rabbit (Draize) test and each of the four *in vitro* alternative test methods being evaluated.

It is well documented that there is a significant degree of variability in the data produced by the *in vivo* rabbit eye test when it is compared with itself, which raises the question as to the accuracy of the *in vivo* test to predict the human experience. Given this variability and the fact that no data demonstrating the ability of the *in vivo* test to predict the human experience was presented to the Panel, Drs. Stephens and Theran feel it should be recognized that this test is an imperfect standard against which the new tests are being measured.

Drs. Stephens and Theran are filing a minority report because they believe that the term “accuracy” is inappropriately used, and that it is more appropriate to use the term “consistency with *in vivo* data” when comparing test results.

## **6.2 Strengths and Limitations of the ICE Test Method**

Discordant results in the ICE test relative to the *in vivo* classification most often were attributed to either surfactants (57% [4/7] false negatives) or alcohols (50% [5/10] false positives). Such instances of discordance with regard to specific chemical classes may reflect some systematic error with the chicken eye or in standardizing the procedures. However, although the ICE BRD analysis attempts to relate failures of classification concordance to chemical class, the lack of concordance should not be attributed solely to such a simple explanation as the variability is too

broad, affecting some chemicals from many classes and their lack of agreement with one or more *in vivo* classification systems. The workers in this field are hampered by historical precedent and the lack of understanding about the cornea as a living tissue.

### **6.3 ICE Test Method Data Interpretation**

There are adequate explanations regarding tissue measurements and endpoints. However, because alcohols are often solvents, and solvents fall into specific chemical classes, they should not be discussed when interpreting accuracy as if they are mutually exclusive designations for a test substance. Mixing product types with chemical nature only confuses the overall conclusions.

## **7.0 ICE TEST METHOD RELIABILITY (REPEATABILITY/REPRODUCIBILITY)**

A major concern with the ICE test method is the number of *in vivo* rabbit eye corrosive/irritants it underclassified. However, if it is part of a tiered testing strategy, this may not be a problem with regard to hazard classification (i.e., if the test is negative, then the substance would be evaluated in the animal test).

### **7.1 Selection Rationale for the Substances Used in the ICE Test Method Reliability Assessment**

Information related to interlaboratory reproducibility is available only from the Balls et al. (1995) study. Sixty substances were evaluated for performance and reproducibility in the ICE test method. One substance was eliminated during testing because of its extreme toxicity (all treated rabbits died). The substances tested covered a broad range of products and ocular irritation responses, and included both solids and liquids as well as polar and non-polar substances. Selection was based, at least initially, on the availability of quality *in vivo* rabbit eye test data. The rationale and the extent to which the substances represented the range of possible test outcomes appear appropriate.

### **7.2 Intralaboratory Repeatability and Intra- and Inter-laboratory Reproducibility of the ICE Test Method**

The analysis and conclusions regarding intralaboratory repeatability and intra- and inter-laboratory reproducibility were appropriate. Both qualitative and quantitative evaluations of ICE interlaboratory variability were conducted appropriately. No intralaboratory repeatability and reproducibility analyses of the ICE test method were conducted because of a lack of appropriate information.

Based on a correlation analysis of ICE results obtained by the four laboratories testing the same set of substance, some endpoints were highly variable (Balls et al. 1995). For example, a correlation coefficient of 0.21 was obtained for corneal swelling when testing water insoluble substances; the consistency among laboratories for this data set is not adequate.

No evaluation has been conducted of ICE interlaboratory reproducibility or repeatability; this is an important data gap for this test method.

It is not surprising that variability among observations increases as the mean value increases, and it is not clear if CV values would be reduced if more eyes per substance (or any other methodological changes) were used. In evaluating the intralaboratory repeatability and intra- and inter-laboratory reproducibility of the ICE test method, the following observations were made:

- The mean/median CV values substantiate the observation of increased interlaboratory variability of corneal swelling relative to the other measures.
- The variation in the CV values among substances covers over two orders of magnitude (e.g., Captan 90 concentrate has fluorescein retention CV=158.7 while 1-naphthalene acetic acid, Na salt has fluorescein retention CV =0). Zero values are only reasonably obtained with very small sample sizes. The rationale for including these in the calculations of the means across substances is unclear. Indeed, it raises the question (which cannot be answered without additional data) of how much of this variation is due to the substances and how much is due to the small sample sizes. Undoubtedly, some of both are involved.
- Box plot summaries of these data (Table 7-4 of the ICE BRD) would provide more of a sense of the distributional aspects of these data, particularly, given that there is so much variation between substances.

There are no criticisms of the statistical methods, but a judgment of the importance of the results for the CV values or the correlations cannot be made. The analysis is thoughtful and sensible, but the conclusions that can be drawn from them are dependent on what is expected and acceptable.

### **7.3 Availability of Historical Control Data**

Historical negative and positive control data were not available. One eye is traditionally used as a negative/vehicle control but irritancy data for this control eye were not available. No analysis of historical negative control data was possible.

### **7.4 Effect of Minor Protocol Changes on Transferability of the ICE Test Method**

The recommended version of the *in vitro* ICE test method may be somewhat sensitive to protocol changes. Any validation study of this test, or any test for that matter, should use a standard test protocol that is not altered by the testers. The protocol should be readily transferable to properly equipped laboratories that are composed of properly staffed and trained personnel.

## 8.0 TEST METHOD DATA QUALITY

### 8.1 Impact of GLP Noncompliance and Lack of Coded Chemical Use

The extent of adherence to national and international GLP guidelines for the three studies reported in the ICE BRD is not adequately presented (see below). This is due to the failure of the reporting organizations to state in a definitive manner that the study (studies) was conducted under GLP. Coding of samples apparently was only employed in one of the three ICE validation studies. Without assurance of GLP guidance including sample coding, the quality of the data cannot be easily verified.

In the case of the Prinsen and Koëter (1993) report, the extent of compliance of the *in vivo* phase of the study with GLP guidelines is not stated. However, these same 21 chemicals when tested in the ICE test were reported to have followed GLP guidelines as outlined by OECD. No specific coding mechanism for the chemicals appeared to have been used.

In the case of the Balls et al (1995) study, 38 of 60 test substances were from the European Center for Ecotoxicology and Toxicology of Chemicals (ECETOC) Eye Irritation Reference Data Bank. The remaining 23 test substances were either from other sources of unpublished data that met the ECETOC selection criteria (nine substances) or were tested after the ICE test method studies had begun (14 substances). (This equals 61 test substances and not 60 test substances as indicated in the ICE BRD [page 8-1, section 8.1.2, first line]. The number of substances from other sources of unpublished data was actually eight, an error that should be corrected in the final version of the BRD). Although not specifically stated in the report, it is assumed by the ICE BRD that these studies were conducted according to GLP guidelines in order to meet the ECETOC selection criteria. A numeric coding of the test substances was used to blind the identities of the test substances or laboratory.

All tests (*in vivo* and *in vitro*) in the Prinsen (1996) study were reportedly conducted according to GLP guidelines as outlined by the OECD.

### 8.2 Results of Data Quality Audits

Since there was no quality assurance to verify the accuracy of the published data and the methods and data were presented in varying degrees of detail and completeness, caution must be exercised when evaluating the data supporting the ICE test method (see Sections 6.0 and 7.0 of the ICE BRD). No information regarding data quality audits was reported for any of the three ICE validation studies. No formal attempt was made to assess the quality of the *in vitro* ICE test method data included in the BRD or to obtain information about the data quality audits from the authors of the ICE test method study reports. The BRD states that raw data were not available for review and evaluation.

A number of limitations were revealed that complicates interpretation of the ICE test method data, including:

- Incomplete substance information such as the Chemical Abstracts Services Registry Number (CASRN).

- The purity and supplier of the test substances not being consistently reported, thereby making comparisons of data from different studies that evaluated the same test substance difficult because of possible differences in purity (this only applies to glycerol and toluene, both of which were tested in Prinsen and Koëter (1993) and Balls et al. (1995)).
- Incomplete data reporting including presenting only the mean ICE endpoint score (i.e., corneal opacity, swelling, fluorescein retention) with no standard deviation to indicate the extent of variability in the data.

### **8.3 Impact of GLP Deviations Detected in the Data Quality Audits**

The impact of deviations or absence from GLP guidelines or other noncompliance issues have been adequately summarized and there is no disagreement with the overall conclusion that “since no reports from data quality audits have been obtained, information on GLP deviations or their impact on the study results is not available”. In the absence of such information, the validation status of the ICE may be questioned.

### **8.4 Availability of Original Records for an Independent Audit**

The lack of available laboratory notebooks or other records of the raw data has been addressed adequately in the ICE BRD. No raw data were used in these evaluations and no records beyond those acquired through the published studies were available for review. The ICCVAM recommendation that all of the data supporting validation of a test method be available with the detailed protocol under which the data were produced is reasonable and should be supported (ICCVAM 2003). Access to the original *in vitro* and *in vivo* data would allow for a more complete retrospective evaluation of ICE. Any future validation studies on the ICE test should include coded test substances of known purity obtained from a common source and centrally distributed, appropriate controls, and be conducted under GLP guidelines.

## **9.0 OTHER SCIENTIFIC REPORTS AND REVIEWS**

### **9.1 Other Published or Unpublished Studies Conducted Using the ICE Test Method**

Information/data from two additional sources (Chamberlain et al. 1997; Procter & Gamble [unpublished data]) were obtained either in response to an ICCVAM *FR* notice (Procter & Gamble), or from the published literature (Chamberlain et al. 1997). In general, inadequate information on the substances tested (identity not specific) and/or on the results obtained from the *in vitro* or *in vivo* studies precluded an assessment of the performance characteristics of the ICE test method.

In addition, a synopsis of two correlation analyses provided in their respective publications (Balls et al. [1995] and Prinsen [1996]) of ICE test results to *in vivo* MAS scores were included in Section 9.0 of the ICE BRD.

Overall, the available information has been adequately considered.

## **9.2 Conclusions Published in Independent Peer-Reviewed Reports or Other Independent Scientific Reviews**

The conclusions have been adequately discussed and compared. The need for histopathological findings, as suggested by Procter & Gamble, appears to be a valuable addition to the routine ICE test method protocol. A public comment (Dr. John Harbell of Institute for *In Vitro* Sciences) was submitted with a similar recommendation for the BCOP test method.

## **9.3 Approaches to Expedite the Acquisition of Additional Data**

The use of an *FR* notice requesting information did not seem to be very productive, since only Procter & Gamble responded by providing additional ICE test data. Personal contacts by the agencies to which data have been submitted may be the best method to secure additional in-house data from the private sector. However, as discussed in **Section II - 4.6**, if such data are not received, additional *in vivo* rabbit studies may be necessary to compile an adequate reference database.

## **10.0 ANIMAL WELFARE CONSIDERATIONS (REFINEMENT, REDUCTION, AND REPLACEMENT)**

### **10.1 Extent to Which the ICE Test Method Refines, Reduces, or Replaces Animal Use**

The ICE test method is considered the first tier in a potential two-tiered battery, where *in vivo* testing is the second tier when the unknown test substance produces a negative result in the first tier. Therefore, live animals would be needed only to confirm the absence of a severe or corrosive outcome from the initial tier. While the ICE test both refines and reduces animal use, the test method is probably best characterized as a partial replacement under the 3Rs of refinement, reduction, and replacement.

Because chickens are used widely as a food animal species, access to chicken eyes can be readily obtained. There is no additional infliction of pain or distress to the animal as a result of the testing procedures. Substances that are identified as ocular corrosives or severe irritants in the ICE test would be excluded from *in vivo* testing, thus sparing rabbits from any pain. However, since mice, rats, birds, and farm animals do not come under the U.S. Animal Protection Act, there is still a need to ensure the humane treatment of chickens. Every effort should be made to ensure that the chickens that are used in the conduct of the ICE test are humanely killed by methods that minimize pain and distress (NOTE: the term “sacrificed” as used in the ICE BRD should be replaced by the more contemporary phrase, “humanely killed”).

## **11.0 PRACTICAL CONSIDERATIONS**

### **11.1 ICE Test Method Transferability**

#### **11.1.1 Facilities and Major Fixed Equipment Needed to Conduct the ICE Test Method**

Because the transferability of a test method affects its interlaboratory reproducibility, consideration must be given to the capital requirements to outfit a laboratory to perform the ICE

test. The location of the facility in the conduct of the test is flexible but should be conducted in a controlled temperature and humidity environment. The major investment in equipment would include a slit-lamp microscope equipped with a depth-measuring device and the superfusion apparatus with eye clamps. The superfusion apparatus and clamps must be custom-made from photographs and diagrams provided by the test method developer (detailed diagrams from which the apparatus could be reproduced should be made publicly available). Peristaltic and vacuum pumps are also needed. If histopathology is included as a component of the ICE method, tissue processing, sectioning, and staining equipment would be required at a significant additional cost. In contrast, the conduct of the *in vivo* rabbit eye test would require a functioning animal testing facility.

Training approaches in the application of this test method should be developed/implemented. A training video and other visual media on the technical aspects of the assay is recommended to ensure consistency.

#### 11.1.2 General Availability of Other Necessary Equipment and Supplies

There are no concerns with regard to this section of the ICE BRD.

### 11.2 ICE Test Method Training

#### 11.2.1 Required Training Needed to Conduct the ICE Test Method

The training required to conduct the ICE test is entirely dependent on the background and experience of the person. Good manual dexterity as well as knowledge of the anatomy of the eye will be required to provide consistent biological specimens with no damage. The ability to recognize an unacceptable specimen is critical. Evaluation of the results at the requisite time points must be addressed in the training, as timing is critical. The person to be trained must be instructed on the use of a slit-lamp to evaluate corneal thickness and the conduct of the subjective measurements. Knowledge of GLP requirements for data collection and storage as well as documentation of modifications in the protocol are also critical in the conduct of the ICE test.

#### 11.2.2 Training Requirements Needed to Demonstrate Proficiency

There are no concerns with regard to this section of the ICE BRD.

### 11.3 Relative Cost of the ICE Test Method

The cost of conducting the ICE test ranges from \$847 to \$1694 without the inclusion of a positive control. With the incorporation of additional eyes for the negative control and a positive control, the costs could double. If deemed necessary, adding histopathology would further increase the cost of the test. However, it would appear that the cost of conducting an ICE test with all of the necessary controls, in triplicate, would approximate the cost of conducting a 3 day/3 animal study.

## 11.4 Relative Time Needed to Conduct a Study Using the ICE Test Method

The ICE test would significantly reduce the time needed to assess the likelihood of a test substance to induce ocular corrosivity or severe irritancy. The ICE test is conducted in less than eight hours (accounting for time to collect material, dissect the eyes and equilibrate the system) as compared to the *in vivo* rabbit eye test that is carried out for a minimum of one to three days (and may continue up to 21 days). However, it is recognized that a corrosive or severe irritant may be detected within a few hours using a single rabbit.

## 12.0 PROPOSED TEST METHOD RECOMMENDATIONS

### 12.1 Recommended Version of the ICE Test Method

#### 12.1.1 Most Appropriate Version of the ICE Test Method for Use in a Tiered Testing Strategy to Detect Ocular Corrosives and Severe Irritants and/or for Optimization and Validation Studies

The ICCVAM criteria for validation (ICCVAM 2003) have not been fully met for the ICE test method based on the following deficiencies:

- The reliability of the ICE test method has not been adequately evaluated.
- The raw data from the three ICE studies included in this evaluation were not available for review.
- Detailed drawings/diagrams of the superfusion apparatus have not been made available to allow for transferability of the experimental setup.

However, the ICE test method can be used in the identification of ocular corrosives/severe irritants in a tiered testing strategy, with the following limitations:

- Alcohols tend to be overpredicted
- Surfactants tend to be underpredicted
- Solids and insoluble substances may be problematic as they may not come in adequate contact with the corneal surface (leading to underprediction)

The low overall false positive rate indicates that the ICE test can be used at present to screen for ocular corrosives/severe irritants. However, given the high false positive rates calculated for a small number of alcohols, caution should be observed when evaluating ICE test results with this class of substances.

### 12.2 Recommended Standardized ICE Test Method Protocol

#### 12.2.1 Appropriateness of the Recommended Standardized ICE Test Method Protocol and Suggested Modifications to Improve Performance

The recommended protocol is based on the original ICE test method protocol, which has changed only slightly since its development. However, it is unclear if the appropriate number of eyes (n=3) is being used to ensure optimum performance. The scientific basis for reducing the number of eyes from five to three has not been evaluated. Therefore, the potential effects on accuracy and reliability of the ICE test method should be the subject of a formal study. One possible approach would be analogous to previous studies performed to evaluate the effects of

reducing the number of animals in the *in vivo* rabbit eye test. During such an evaluation, random samples of five-, four-, or three-eye subsets could be extracted from a database of six-eye tests to simulate the results of using fewer eyes per test substance. It is also unclear if the use of maximum mean scores is the most appropriate scoring system to ensure optimum performance; this also should be formally evaluated.

The method for contact with the test substance has room for refinement since the eye is removed from the superfusion apparatus. The actual contact time may not be ten seconds as stated due to manipulation time. Some further evaluation of the chemical contact procedure should be examined, or the apparatus should be moved to a horizontal position to obviate the need for test eye removal during dosing.

Centering lights should be installed on the optical pachymeter to ensure consistent central corneal thickness measurements across laboratories.

The protocol must specify that universal safety precautions be observed when handling chemical and biological materials.

#### 12.2.2 Other Endpoints that Should be Incorporated into the ICE Test Method

Histopathology, including determining the nature and depth of corneal injury, should be considered when the standard ICE endpoints (i.e., corneal opacity, swelling, fluorescein retention) produce borderline results. A standardized scoring scheme should be defined using the formal language of pathology to describe any effects. The appropriate circumstances under which histopathology would be warranted should be more clearly defined. To maximize the likelihood of obtaining reproducible results, reference photographs for all subjective endpoints (i.e., corneal opacity, fluorescein retention, histopathology) should be readily available.

### 12.3 **Recommended Optimization and Validation Studies**

Any optimization and validation studies should use existing animal data, if available. Additional animal studies should only be conducted if important data gaps are identified, and such studies should be carefully designed to maximize the amount of pathophysiological (e.g., wound healing) information obtained and to minimize the number of animals used.

#### 12.3.1 Recommended Optimization Studies to Improve Performance of the Recommended ICE Test Method Protocol

Additional studies using the recommended ICE test method protocol are needed to better characterize the repeatability and the intra-and inter-laboratory reproducibility of the test method. However, if optimization studies are carried out, they should make maximum use of retrospective analyses to preclude the need for further, time-consuming studies. An evaluation of the impact of variations in the time between death and testing of the chicken eyes on assay performance should be included.

Reference substances should be identified that can be used as part of the performance standards developed for the validated test method. NICEATM/ICCVAM should facilitate the development of a histopathology scoring system for corneal damage (with visual aids as indicated above).

The combined score method has been published by Prinsen with comparison to the EU classification procedure. Some additional work has been carried out for comparisons with other *in vivo* schemes. Additional work is needed in this area with standardization across the method of scoring and chemicals with application to other *in vivo* data. It is also suggested that a more heterogeneous database be developed that includes as many chemical parameters (e.g., pH, functional groups etc.) as possible.

In addition, based on the excessive false negative rate of 40% (for the GHS classification system), using the current version of the ICE test method could result in a large number of ocular corrosives/severe irritants still undergoing testing in the *in vivo* rabbit. Therefore, studies designed to optimize the decision criteria used for classification should be conducted in an attempt to reduce this rate, without unacceptably increasing the current false positive rate. A multivariate analysis might be useful in optimizing the decision criteria. Finally, the impact of routinely performing replicate experiments on the performance of the ICE test method should also be evaluated.

#### 12.3.2 Recommended Validation Studies to Evaluate Performance of the Optimized ICE Test Method Protocol

Information on intra- and inter-laboratory reliability is important to know. The information that is available regarding interlaboratory reproducibility is encouraging. If further validation work is carried out, it should take full advantage of the new modular approach to validation that ECVAM is developing. According to this approach, “modules” of information could be populated with the available information for ICE, and deficient modules (e.g., interlaboratory reliability) could be the focus of additional studies. This activity would minimize the required resources by preventing the need for a full validation study.

To the extent that the recommended version of the ICE test method may be suitable for the testing of substances within certain chemical classes, additional testing of such substances to determine accuracy may not be necessary. However, given the small number of substances tested within each chemical class with the ICE test, such a conclusion may not be warranted at this time.

In addition, as part of any analysis of validation data, the Panel suggests an assessment based on the ranking of experimental data for severity for both the *in vivo* reference method and the *in vitro* test.

No matter what validation studies are deemed necessary, the BRD should discuss the pros and cons of the immediate implementation of the ICE test for the identification of ocular corrosives and severe irritants in a tiered-testing approach. This discussion should answer the question: What, if anything, is the downside of foregoing the proposed optimization and validation work and simply implementing the ICE Test in a tiered-testing approach?

#### Minority Opinion

According to Dr. Martin Stephens, **Section II – 12.3** recommends that additional optimization and/or validation studies be conducted, and the report leaves open the possibility of additional animal studies as part of this process. Dr. Stephens believes that no additional animal studies

should be conducted for such optimization or validation exercises. He cited several reasons for holding this view:

1. Draize testing of severely irritating or corrosive chemicals causes extremely high levels of animal suffering.
2. The intended purpose of the alternatives under review is narrow in scope (i.e., simply to serve as a positive screen for severely irritating or corrosive chemicals). Negative chemicals go on to be tested in animals.
3. The Panel learned that more animal and alternative data exist that are relevant to each of the alternative methods, and greater efforts should be made to procure these and any other existing data.
4. Some relevant animal data were dismissed from the analysis of each alternative method, and this dismissal should be reevaluated in light of any need for additional data.
5. Suggestions for further optimization and/or validation studies should be assessed critically, in light of the fact that only the most promising alternative method need be developed further, not necessarily all four methods, and that whatever alternative is selected for further development need be optimized only to the point at which it is at least as good as the Draize test.
6. A new modular approach to validation has been developed that could potentially reduce the number of chemicals needed to fulfill each module. Such an approach, if pursued, might be workable with the data already summarized in the BRDs.

#### **12.4 Proposed Reference Substances for Validation Studies**

See Section V.

#### **13.0 ICE BRD REFERENCES**

##### **13.1 Relevant Publications Referenced in the ICE BRD and any Additional References that Should Be Included**

There are no concerns with regard to this section of the ICE BRD.

#### **14.0 PANEL REPORT REFERENCES**

Balls M, Botham PA, Bruner LH, Spielmann H. 1995. The EC/HO international validation study on alternatives to the Draize eye irritation test. *Toxicol In Vitro* 9:871-929.

Bruner LH, Carr GJ, Chamberlain M, Curren RD. 1996. Validation of alternative methods for toxicity testing. *Toxicol In Vitro* 10:479-501.

Chamberlain M, Gad SC, Gautheron P, Prinsen MK. 1997. IRAG Working Group I: Organotypic models for the assessment/prediction of ocular irritation. *Food Chem Toxicol* 35:23-37.

EPA. 1996. Label Review Manual. 2<sup>nd</sup> Edition. EPA737-B-96-001. Washington, DC:U.S. Environmental Protection Agency.

EU. 2001. Commission Directive 2001/59/EC of 6 August 2001 adapting to technical progress for the 28th time Council Directive 67/548/EEC on the approximation of the laws, regulations and administrative provisions relating to the classification, packaging and labelling of dangerous substances. Official Journal of the European Communities L255:1-333.

ICCVAM. 2003. ICCVAM Guidelines for the Nomination and Submission of New, Revised, and Alternative Test Methods. NIH Publication No. 03-4508. Research Triangle Park, NC:National Institute of Environmental Health Sciences.

Kaneko T. 1996. The importance of re-evaluating existing methods before the validation of alternative methods – the Draize test (in Japanese). *The Tissue Culture* 22:207-218.

OECD. 1987. Acute Eye Irritation/Corrosion. Test Guideline 405. Paris, France: Organisation for Economic Co-operation and Development.

Ohno, Y, Kaneko T, Inoue T, Morikawa K, Yoshida T, Fuji A, Masuda M, Ohno T, Hayashi M, Momma J, Uchiyama T, Chiba K, Ikeda N, Imanashi Y, Itagaki H. 1999. Interlaboratory validation of the *in vitro* eye irritation tests for cosmetic ingredients. (1) Overview of the validation study and Draize scores for the evaluation of the tests. *Toxicol In Vitro* 13:73-98.

Prinsen MK. 1996. The chicken enucleated eye test (CEET): A practical (pre)screen for the assessment of eye irritation/corrosion potential of test materials. *Food Chem Toxicol* 34:291-296.

Prinsen MK, Koëter BWM. 1993. Justification of the enucleated eye test with eyes of slaughterhouse animals as an alternative to the Draize eye irritation test with rabbits. *Food Chem Toxicol* 31:69-76.

Spielmann H. 1996. Alternativen in der Toxikologie. In: Alternativen zu Tierexperimenten, Wissenschaftliche Herausforderung und Perspektiven (in German). (Gruber FP, Spielmann H, eds). Berlin/Heidelberg/Oxford:Spektrum Akademischer Verlag, 1006:108-126.

Spielmann H. 1997. Ocular Irritation. In: *In Vitro* Methods in Pharmaceutical Research. (Castell JV, Gómez-Lechón MJ, eds). London:Academic Press, 265–287.

UN. 2003. Globally Harmonised System of Classification and Labelling of Chemicals (GHS). New York & Geneva: United Nations.

Weil CS, Scala RA. 1971. Study of intra- and inter-laboratory variability in the results of rabbit eye and skin irritation tests. *Toxicol Appl Pharmacol* 19:276-360.