Isolated Rabbit Eye Test Method

[This Page Intentionally Left Blank]

I. ISOLATED RABBIT EYE TEST METHOD

1.0 IRE TEST METHOD RATIONALE

1.1 Scientific Basis for the IRE Test Method

The Isolated Rabbit Eye (IRE) test method, an *in vitro* alternative to the Draize rabbit eye test, is an organotypic model in which effects on the cornea are measured, while effects on the iris and conjunctiva are not determined. Moreover, the IRE is a short-term test. Therefore, in contrast to the *in vivo* rabbit eye test, reversible effects cannot be determined over a period of up to 21 days.

1.1.1 Mechanistic Basis of the IRE Test Method

Although corrosive, irritant, and non-irritant responses are described in the IRE Background Review Document (BRD), the emphasis is on the manifestation of the injury rather than the mechanism(s) by which injury is caused. For example, a corrosive is defined as a "substance that causes visible destruction or irreversible alteration in the tissue at the site of contact." However, the mechanism(s) responsible for the destruction are not described. Such a description could include what happens at the cellular level. For example, if damage is caused by cell death, the mechanism for such cell death (necrosis, apoptosis, or both) could be described. The BRD should be updated to reflect the fact that the basis of the IRE is not mechanistic but rather a correlation of descriptive observations of toxicity. The IRE test is conducted using the same organ from the same animal as the *in vivo* test, and therefore defining a mechanistic basis may not be necessary. The accumulated IRE data have been compared to the *in vivo* rabbit eye test data by correlative methods; precedent exists for using such comparisons for validation of toxicological test methods. This is an important point with applicability not just to the IRE, but also to the three other *in vitro* test methods for ocular damage under consideration.

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

The differences in endpoints between IRE and the *in vivo* rabbit eye test are described. There is some discussion of the various kinds of responses in different parts of the eye that occur *in vivo*. For example, the IRE BRD indicates that development of slight corneal opacity can result from the destruction of superficial epithelial cells and consequent swelling in the remaining cells (epithelial edema), but the cellular response mechanisms producing these epithelial cell changes are not described. In some instances, corneal changes that appear to have the same endpoint might arise from different mechanisms (e.g., direct epithelial cell damage versus endothelial cell damage leading to changes in the corneal cells and loss of corneal clarity). In the *in vivo* rabbit eye test, the manifestations of corneal injury involve an inflammatory response. Some discussion of the role of resident and/or migrating inflammatory cells, their products (e.g., cytokines which are early responders anytime the cornea and/or conjunctiva are perturbed), and potential ocular effects should be included in the BRD. The consequence of the loss of vascular perfusion on ocular responses in the *in vitro* test should also be discussed. Furthermore, extrapolation of the effect of not having responding cells and their products would be another topic for consideration when the *in vivo* and *in vitro* tests are compared. This discussion may be useful in providing groundwork for future research efforts and also to contrast differences between the *in vivo* and *in vitro* responses, which will possibly help to delineate limitations of the IRE test method compared to the *in vivo* rabbit eye test.

1.1.3 Similarities and Differences of Mechanisms/Modes of Action and Target Tissues between the IRE Test Method and Humans and Rabbits

As noted above, the mechanisms by which cellular damage in the eye could be caused by various agents are not considered in the IRE BRD. If there is published information on the response of cells to corrosive and irritating agents (from *in vivo* and/or *in vitro* studies), this information could be used to compare and contrast the responses of the different types of corneal cells from different species to various types of irritants. While the basis for the IRE is correlative between results obtained in the same organ from the same animal *in vivo* versus *in vitro*, further consideration of mechanisms may be warranted. More robust discussion of possible mechanisms may highlight specific needs for further research either before or during standardization or validation studies. Thus, it may be useful to propose additional methods (e.g., microscopy, immunohistochemistry) and to perform mechanistic assays (e.g., apoptosis, necrosis) to develop a better understanding of the mechanisms of corneal damage in response to severe irritants from different chemical classes. There is a good description of differences in the anatomy of the eye between humans and rabbits in this section of the BRD.

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

As discussed in the preceding section, additional considerations of mechanisms of cellular damage by different classes of irritants are needed. Also, additional side-by-side comparisons of various classes of substances in the *in vivo* and *in vitro* tests (the same substance in both tests) would strengthen the case for the use of the IRE test. Historical published results are presented in later sections of the IRE BRD, but inclusion of parallel *in vivo* and *in vitro* test results might also be useful in this section to strengthen the rationale.

1.2 Regulatory Rationale and Applicability

The IRE test method is designed to identify substances that are severely irritating/corrosive to the cornea. Since corneal effects are given the greatest weight in the Draize rabbit eye test (73% of the total score), the endpoints measured in the IRE test focus on the most important endpoint used in the *in vivo* test.

1.2.1 Similarities and Differences in the Endpoints Measured in the IRE Test Method and the *In Vivo* Rabbit Eye Test Method

The similarities and differences in endpoints between the *in vivo* and the *in vitro* test are covered quite thoroughly. The limitations of the IRE test method in terms of not being able to detect effects on the iris, conjunctiva (including the limbus), or systemic damage are also well described as is the difference in time it takes for either assay to be conducted (up to 21 days *in vivo* compared to four hours *in vitro*). It is also noted that the IRE test does not evaluate the reversibility of corneal effects.

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

The United Nations (UN) Globally Harmonised System (GHS) of Classification and Labelling of Chemicals tiered testing strategy (UN 2003) is described in the IRE BRD in Figure 1-2. While the situations in which severe eye damage is caused should not be difficult to evaluate using this

strategy, the effect of the non-corrosive or mildly irritating substances will be more difficult to judge using only macroscopic criteria and slit lamp examination. In the case where damage is not observed or the observation is equivocal, microscopic evaluation of the cornea could be used to determine whether any non-corrosive or non-irritating substance caused changes in any or all of the corneal layers that could not be observed by eye or with the slit lamp. By analogy, histopathology has been reported to improve the sensitivity of the Bovine Corneal Opacity and Permeability (BCOP) test method (see BCOP BRD). It is recommended that histopathology or microscopy be considered to evaluate early markers of ocular effects and identify transient versus progressive changes. A limited number of apparently non-corrosive or non-irritating substances that caused changes at the microscopic level could be tested *in vivo* to determine if the changes were transient or perhaps would progress and cause additional damage to the cornea; effects that could not be assessed in a short-term (hours) *in vitro* assay. Although the IRE test method as described is intended only for corrosives and severe ocular irritants, assessing the validity of this *in vitro* test against a broader range of irritants (e.g., mild and/or moderate) would be useful.

2.0 TEST METHOD PROTOCOL COMPONENTS

It is well known that a proposal for an optimized, new protocol based on other existing but non-optimal protocols represents a compromise protocol that has never been directly assessed in any laboratory. This has to be kept in mind because the results that will be obtained with the new protocol may differ significantly from the results obtained using the individual protocols in previous validation exercises. For example, the proposed standardized protocol for the IRE test method was provided by SafePharm Laboratories (Derby, United Kingdom) and was used by Guerriero et al. (2004) to provide data described in the IRE BRD. However, the data set generated using this protocol was limited to 36 substances classifiable by the GHS classification system (UN 2003). Furthermore, this protocol has not been used in other laboratories.

While the proposed standardized protocol provided in Appendix A of the IRE BRD adequately describes the decision criteria used in IRE test method, the protocol does not include a description of the biostatistically-based algorithm used to justify the decision criteria for identifying a corrosive or severely irritating response. Decision criteria based on a biostatistically-derived algorithm are an essential part of every toxicity test, as outlined in the current documents on the validation of *in vitro* toxicity tests published by the Organisation of Economic Co-operation and Development (OECD), the European Centre for the Evaluation of Alternative Methods (ECVAM), and the Interagency Coordinating Committee for the Validation of Alternative Methods (ICCVAM) (OECD 2002; ECVAM 2005; ICCVAM 2003). Another weakness in the existing IRE test method protocols is the lack of established reference substances (negative and positive controls, benchmarks). These are needed as part of the decision criteria for identifying ocular corrosives and severe irritants. Thus, acceptable reference substances from a validated reference list should be identified in the standardized protocol provided in Appendix A of the IRE BRD. Also, additional *in vitro* data obtained using a set of test substances for which high quality *in vivo* data are available are needed. With such a data set, simple biostatistical approaches (e.g., discriminant analysis) can be used to identify a cut-off score to distinguish between test substances that are positive and those that are negative for the endpoints that are evaluated.

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

The protocol components are thoroughly described along with background information, a recommendation, and a rationale for each recommendation. In the IRE test method, the following endpoints should be measured on the cornea: opacity, thickness (swelling), and fluorescein penetration. Identification of reference substances that are part of the performance standards should be developed for the validated test method. New tests should be conducted according to Good Laboratory Practice (GLP) guidelines. The numerical data obtained for each endpoint by subjective or objective evaluation will allow a determination, for a series of test substances, of the variability of the endpoint values, the calculation of scores, and a comparison with the *in vivo* rabbit eye scoring system.

2.1.1 Materials, Equipment, and Supplies

The IRE BRD is not clear in regard to the position of the rabbit eyes during the test (i.e., vertical or horizontal or vertical pre- and post- and horizontal during the application of the test substance). The reference materials (i.e., publications, submitted reports) were also not very clear on the position of the eyes during treatment and it appeared that different protocols might have used different positions. The inclusion in the protocol in Appendix A of the BRD of a diagram or picture of the superfusion chamber used for the studies would improve clarity since readers might not have ready access to the Burton et al. (1981) reference that describes this equipment. Furthermore, the commercial availability of this apparatus should be addressed. If not available commercially, the feasibility for custom-building this apparatus should be discussed.

The New Zealand White is a common strain of rabbit used in many laboratories, and IRE test method studies have been performed primarily using eyes from these rabbits, although some data have been obtained using eyes from non-specified albino strains. However, there was no comparison in the IRE BRD of results based on which rabbit strain was used as a source for eyes. Use of a different type of rabbit would be an area of concern only (a) if there are significant differences in corneal characteristics between different types of rabbits, and, if (b) the supplier provided eyes from rabbits of different strains without informing the laboratory that was going to be doing the *in vitro* testing. Thus, guidance should be provided in the protocol regarding the appropriate strain(s) of rabbit that may be used in the IRE test.

In the test method protocol, another section could be added to Section 3.1 of Appendix A of the IRE BRD to describe the evaluation of the eyes after removal but prior to shipment to the testing laboratory. The protocol should indicate whether use of both eyes from a single rabbit can appropriately be used in the same test, and if a concern, how to prevent bias (e.g., through randomization).

Section 6.2 of Appendix A of the IRE BRD discusses the evaluation of eyes once they have reached the testing laboratory. Additional guidance is needed on storage/transport conditions for enucleated eyes (i.e., optimum temperature and buffer conditions, maximum storage times, etc.) prior to and during shipment to the testing facility.

2.1.2 Dose-Selection Procedures

This section of the IRE BRD adequately describes dose-selection procedures.

2.1.3 Endpoint(s) Measured

Additional methods that could be used in the IRE test method include confocal microscopy or fixation, sectioning, and staining of corneal sections with a variety of stains to detect cellular changes. As noted earlier in this report, such additional tests might be used if the results of an *in vitro* test were equivocal. Use of a histological approach in which all layers of the cornea are examined microscopically might also provide information about whether eyes undergoing treatment with a mild irritant (which would not be detected by the *in vitro* studies) would be predictive for a response that took longer than four hours to develop. These studies would require histopathological results from eyes that were apparently normal after four hours of *in vitro* testing to be compared with microscopic and macroscopic results from *in vivo* tests of substances for which signs of ocular damage did not appear until later in the study (>four hours to days).

2.1.4 Duration of Exposure

This section of the IRE BRD adequately describes exposure duration.

2.1.5 Known Limits of Use

Some information on known limits of use is provided in Sections 1.2.3 and 2.2.5 of the IRE BRD. However, no mention is made of specific considerations that would contradict use of this test. If such information is available, it should be included at the beginning of the proposed standardized protocol provided in Appendix A and in these two BRD sections.

2.1.6 Nature of the Response(s) Assessed

IRE test method users should evaluate if there is a way to quantify the extent of fluorescein penetration (for example, by microscopy and assessment of pixel intensity of fluorescein stains or measurement of the amount of fluorescein after extraction from the cornea).

2.1.7 Appropriate Controls and the Basis for Their Selection

In addition to the negative control, inclusion of a positive control and, when appropriate, benchmark and solvent/vehicle controls is an important addition to the IRE protocol and is appropriately stressed in several sections of the IRE BRD.

2.1.8 Acceptable Range of Control Responses

This topic is minimally defined in the IRE BRD. The use of control charts to monitor responses to control substances over time and across laboratories is an effective means of monitoring the "range" of responses and for updating test acceptance criteria.

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

This section of the IRE BRD adequately describes the nature of the data collected and the methods used for data collection.

2.1.10 Type of Media in Which Data are Stored

While not defined in the IRE BRD, GLP or equivalent standards should apply.

2.1.11 Measures of Variability

The IRE BRD describes the summary statistics associated with the quantitative endpoints and the possible use of additional subjective measurement of variability. Clearly, some use could be made of these quantitative data to assess inter- and intra-laboratory variability (which is suggested later in the BRD). The quantitative and semi-quantitative data described in Table A-3 (BRD Appendix A) on maximum fluorescein uptake, corneal opacity, and corneal swelling (which are used to derive an overall score for evaluation) could be used to obtain quantitative estimates of intra- and inter-laboratory variation. However, as the individual eye data are combined to give an overall assessment, such data may not be easy to extract in a standard format from previous studies using other versions of the IRE protocol. The fact that there is currently no widely accepted standardized IRE test method protocol may further complicate this task.

2.1.12 Statistical or Nonstatistical Methods Used to Analyze the Resulting Data

This section describes the decision criteria used for identifying a severe irritant. These criteria are based on one or more of four ocular parameters exceeding a predefined cutoff. Clearly, a test substance could be classified as a severe irritant based upon different patterns of response in these four measures. In this sense, the criteria are not based on any formal statistical assessment of the data. Thus, it might be reasonable to more carefully evaluate the possible patterns of results. For example, data on substances falling just below the decision criteria cutoff values for one or more endpoints could be evaluated to see whether such substances could be realistically referred to as non-severe irritants. This evaluation would presumably have to rely on direct statistical comparison with *in vivo* rabbit eye data for test substances given a comparable severe or nonsevere irritant classification. It should also be recognized that any change to the IRE test method protocol, such as increasing or decreasing the number of eyes used per test substance, might have an appreciable effect on the decision criteria.

Information on the individual scores should be used to calculate descriptive statistics for corneal opacity, corneal swelling, and fluorescein penetration.

2.1.13 Decision Criteria and the Basis for the Algorithm Used

The IRE BRD does not currently identify the rationale or statistical algorithm used for the development of the decision criteria to identify an ocular corrosive or severe irritant, as described in Appendix A and Section 2.0, and does not identify appropriate reference substances (negative and positive controls, benchmarks). Thus, the BRD needs to be revised accordingly.

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

This section of the IRE BRD appears adequate. Exhibits (examples) of standard forms used for collection and transmission of data provided by laboratories using the assay would be helpful.

2.2 Adequacy of the Basis for Selection of the Test Method System

The use of the IRE as a screening method to identify ocular corrosive or severely irritating substances is well presented. The relationship of the IRE model to the *in vivo* rabbit eye test that has been the basis for ocular safety testing for many years is apparent.

2.3 Identification of Proprietary Components

The Panel agrees that no proprietary components are used in the IRE test method.

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

Within the context laid out in the ICCVAM Submission Guidelines (ICCVAM 2003), the statistical methods used to assess the data seem appropriate for these complex endpoints and provide a firm basis for further considerations across these data sets (see Sections 6.0 and 7.0 of the IRE BRD). The conclusions relating to test method reliability (IRE BRD Section 7.4) drawn from the analyses in Section 7.0 of the documents based upon these analyses seem basically sound.

2.5 Study Acceptance Criteria for the IRE Test Method

An individual test result is acceptable if an appropriate response is obtained for the negative and positive controls and, if used, a benchmark substance. The appropriate response could be a quantitative response or an acceptable range of responses relative to historical data (control chart analysis) for control substances. Compliance with GLP guidelines is not in itself a required or sufficient acceptance criterion.

2.6 Basis for any Modifications made to the Original IRE Test Method Protocol

The basis for the recommended protocol has been adequately described. However, any additional revisions (e.g., to add potential enhancements) must be supported by specific written technical rationale.

2.7 Adequacy of the Recommended Standardized Protocol Components for the IRE Test Method

This section is appropriately covered in the IRE BRD with the following two exceptions. First, as already described in **Section I - 1.1.2** of this Panel report, the protocol should include the potential application of histopathology, which would require that a standardized histopathology scoring system be implemented with visual aids and that the conditions for the use of histopathology in the IRE be clearly defined. Second, reference substances (negative and positive controls, benchmarks) need to be identified; the description of reference substances in Section 5.0 of Appendix A of the IRE BRD does not meet the standard of the most recent OECD Test Guidelines (TGs), in which guidance is given on appropriate reference substances (i.e., those that are supported by high quality *in vivo* and *in vitro* data). For example, tables of reference chemicals to be used as positive and negative controls and as benchmarks are provided

in TG 431, *in vitro* skin corrosion test (OECD 2004a) and in TG 432, 3T3 NRU *in vitro* phototoxicity test (OECD 2004b). The standardized protocol should be revised to identify appropriate reference substances from the list of recommended Reference Substances provided by the Expert Panel Reference Substance Subgroup.

3.0 SUBSTANCES USED FOR PREVIOUS VALIDATION STUDIES OF THE IRE TEST METHOD

3.1 Substances/Products Used for Prior Validation Studies of the IRE Test Method

The types and numbers of substances/products used in prior studies appear to be adequate to the extent that the IRE protocol has progressed to its current status. However, the types and number of substances/products to be used for any further standardization/validation studies need to be identified.

3.2 Coding Procedures Used in the Validation Studies

Coding with respect to the IRE test method validation studies appears to have been adequate and no specific concerns have been identified.

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 in full detail.

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 in the cited 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* ocular test 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 Records

In the case of the IRE test method, sanitized copies of such records were available for the Guerriero et al. (2004) data. 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 is adequate, 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 Availability of GLP Compliance

The Balls et al. (1995) European Commission/Home Office (EC/HO) validation study included criteria that *in vivo* data be submitted from GLP compliant post-1981 studies. The *in vivo* rabbit eye test data used in the Gettings et al. (1996) Cosmetic, Toiletries, and Fragrance Association (CTFA) alternatives evaluation study was also GLP compliant. Most of the *in vivo* data from the Guerriero et al. (2004) study was GLP compliant (Guest R, personal communication). 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 IRE BRD may not be necessary. According to the current European Union (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 criteria to exclude *in vivo* or *in vitro* data from the evaluation of the performance of a toxicity test. Verification of data quality can be difficult but is essentially similar whether the study was GLP or non-GLP. In either case, laboratory/data inspection could be required. This may be determined, subjectively, to be unnecessary, particularly if further standardization/validation studies are pending that will be carefully controlled and managed to current standards and expectations.

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 method. 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. Controlled human studies are ethically initiated only after careful *in vivo* animal tests and involve essentially non-irritating materials. Non-irritants have little or no discriminating power with regard to agent, test method, or laboratory. There needs to be a 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

The Draize rabbit eye irritation test has never gone through a formalized validation process. 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 evaluations of this assay conducted by Kaneko (1996) and Ohno et al. (1999). Using a fixed protocol and a single supply of chemical agents tested in 25 laboratories, Weil and Scala (1971) 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 non-irritants provided little useful information on laboratory performance. The discordance in Maximum Average Score (MAS)

values calculated for the same substance among different laboratories in this study has been reviewed by Spielmann (1996), who noted that three of the ten substances tested were classified anywhere from non-irritant (MAS < 20) to irritant (MAS > 60) when tested in 24 different laboratories. 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. It is also well documented that the Draize eye test has a very 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 coefficient of variation [CV] and standard deviation [SD] values for such substances in Balls et al. [1995]).

In the development of alternative methods to intact animal testing, the question always arises regarding the quality of reference *in vivo* data used to evaluate or validate the newer *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", that is, 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 intact rabbit eye. Consequently, the evaluation (or acceptability) of an alternative method is unavoidably biased by the selection of the *in vivo* data used in that evaluation. Thus, there should be more discussion in the IRE BRD of the variability of the *in vivo* rabbit eye test data. This is particularly important in the determination of the accuracy of an *in vitro* test method. While there are often multiple study results for each *in vitro* determination of irritation potential, there generally is only one *in vivo* test result. Because of the known variability in the rabbit test, it is not possible from the data presented to determine if the inconsistencies between the two tests are due to "failure" of the *in vitro* test method or a misclassification by the single *in vivo* result provided. When interpreting the *in vitro* test data, these differences in reproducibility/variability of the *in vivo* Draize eye test data have to be taken into account.

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

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 I - 12.3**).

5.0 IRE TEST METHOD DATA AND RESULTS

5.1 IRE Test Method Protocols Used to Generate Data Considered in the BRD

The recommended test method protocol includes additional parameters that enhance the accuracy of the IRE test method (Guerriero et al 2004).

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

Although the IRE BRD considered all of the comparative data sets produced with the IRE test method that were available for this evaluation, National Toxicology Program (NTP) Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) should make additional efforts to obtain comparative data from testing laboratories and other private sources.

5.3 Statistical and Nonstatistical Approaches Used to Evaluate IRE Data in the BRD

Within the context described in the ICCVAM Submission Guidelines (2003), the statistical methods used to assess the data seem appropriate for these complex endpoints and provide a firm basis for further considerations across these data sets (IRE BRD Sections 6.0 and 7.0). The conclusions relating to test method reliability (Section 7.4) drawn from the analyses in BRD Section 7.0 based upon these analyses seem basically sound.

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

Documentation of data quality is adequate. Only two studies (Balls et al. 1995; Getting et al. 1996) were described as GLP compliant in the IRE BRD. One of the remaining two studies (Guerriero et al. 2004) was also GLP-compliant and this should be stated in the BRD. As noted previously in this report, the absence of 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, when the basic requirements of the GLP procedure have been implemented in a study.

5.5 Lot-to-Lot" Consistency and Time Frame of the Various Studies

This point is adequately covered in Section 5.6 of the IRE BRD. Substances were tested only once in each study, and therefore, lot-to-lot consistency was not applicable. However, lot consistency was controlled and described in three of the four studies (Balls et al. 1995; Gettings et al. 1996; CEC 1991).

6.0 IRE TEST METHOD ACCURACY

As outlined in prior sections, the IRE BRD does not adequately discuss the high variability of the Draize eye test *in vivo* as has been described by Weil and Scala (1971), Balls et al. (1995), Spielmann (1997), Kaneko (1996), and Ohno et al. (1999). Moreover, a biostatistical concept on how to include this variability into calculating the performance of the IRE has not been

presented. Thus, the biostatistical evaluation in the current study is limited and may be inadequate.

6.1 Accuracy Evaluation of the IRE Test Method for Identifying Ocular Corrosives and Severe Irritants

The variability of the *in vivo* rabbit eye test method is not considered in this evaluation. Some discussion of this is warranted, particularly as to its performance with severe irritants and corrosives, and therefore, its basis as a standard for comparison for the IRE test method. However, the results given in Section 6.1 of the IRE BRD, in particular the results summarized in Tables 6-1, 6-2, and 6-3, provide a correct overview of the performance of the IRE test as reported in the studies. The description of discordant results obtained among the four studies, as presented in IRE BRD Section 6.2, is also correct.

There are several weaknesses in the evaluation of the accuracy of the IRE test. These include:

- The lack of a common protocol in the different IRE studies. The relevant studies were conducted over a period of 10 years, and during this time the decision criteria changed. In earlier studies, corneal swelling and opacity only were evaluated. Most recent studies measured maximal corneal opacity, maximal corneal swelling, and fluorescein penetration, and conducted a slit-lamp assessment of epithelial integrity over time. It is encouraging that, for the most part, the protocol used in the later study (i.e., Guerriero et al. [2004]), upon which the recommended protocol is based, improved both the sensitivity and specificity of the test method for the substances tested.
- The lack of individual *in vivo* rabbit test data. All three regulatory classification systems utilize individual rabbit data and these data were not consistently available in the publications considered for this evaluation.
- The limited database. The evaluation is based on a relatively small number of substances; more data are being requested and additional data mining may permit a more robust evaluation.

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 IRE Test Method

The text in Section 6.3 of the IRE BRD gives the wrong impression about the timing of various IRE comparative studies. The Commission of the European Communities (CEC) study was published in 1991 while the EC/HO study (Balls et al. 1995) was started in 1992. In a similar manner, the CTFA study was published by Gettings et al. (1996) and was, therefore, most probably conducted after the CEC study

The source/reference for the individual *in vivo* and *in vitro* test results in Tables 6-4 and 6-5 of the IRE BRD need to be provided, as does whether the test results represent individual chemicals or products from a single study or from several studies. Moreover, the criteria used for compiling the data included in these tables need to be described and the experts who compiled the tables need to be identified. Furthermore, the tables need to indicate which *in vitro* data set was used to calculate the IRE classifications. Thus, the tables should be appropriately titled and referenced; otherwise it is unclear whether the recommendations based on Tables 6-4 and 6-5 of the IRE BRD are justified.

Additional testing appears to be needed. While existing data would suggest that the IRE test method overpredicts some substance classes, the number of substances tested in these categories of chemicals is very small. More testing might provide for a better analysis of strengths and weaknesses. In addition to the analyses conducted, a comparative ranking assessment, based on severity both for the IRE and the *in vivo* rabbit eye test methods, should be conducted.

6.3 IRE Test Method Data Interpretation

The discussion in the IRE BRD of the value of including all of the proposed endpoints appears to be thorough. However, rather than using the "weight of evidence" approach appropriately and taking into account both the limitations of the results of the Draize eye test in rabbits *in vivo* and of the IRE test *in vitro*, the BRD focuses only on the limitations of the *in vitro* data sets produced with the IRE method. When drawing conclusions about strengths and limitations of an *in vitro* test, the strengths and limitations of the standard test method against which the alternative test is being measured must also be considered. For example, issues regarding data quality in the Draize eye test have been discussed (Balls et al. 1995). Furthermore, Weil and Scala (1971), Kaneko (1996), and Ohno et al. (1999) demonstrated intra- and inter-laboratory variability in the Draize test. There appears to be a lack of data in the BRD to either refute or confirm their observations. Clearly, variability in the reference test method would confound attempts to demonstrate consistency of the alternative test method. This being the case, issues related to test interpretation, and the strengths and limitations of the *in vivo* rabbit eye test should be included in the IRE BRD. However, it is important to remember that the variability of the Draize test for severe irritants and corrosives may not occur to the same extent as for moderate irritants, and the IRE test method seems to err more toward false positives than false negatives.

7.0 IRE TEST METHOD RELIABILITY (REPEATABILITY/ REPRODUCIBILITY)

The IRE BRD indicates that the reliability of the IRE could not be evaluated. Since this problem was encountered in previous prevalidation and validation studies that were conducted in Europe under the auspices of ECVAM, three documents have been provided to NICEATM in which the problem is discussed in more detail. The information in these documents should be included in Section 7.0 of the IRE BRD.

- The first contribution is the classical statistical publication by Bland and Altman (1986). The authors describe the problem being faced in the current evaluation in the first paragraph of the section on "Repeatability" as follows: " Repeatability is relevant to the study of method comparison because the repeatability of the two methods of measurement limit the amount of agreement which is possible. If one method has poor repeatability (i.e. there is considerable variation in repeated measurements on the same subject), the agreement between the two methods is bound to be poor too. When the old method is the more variable one, even a new method that is perfect will not agree with it. If both methods have poor repeatability, the problem is even worse." As a consequence, from a scientific perspective, if the repeatability of the IRE and the *in vivo* rabbit eye test methods are determined to both be unacceptably low, then the correlation between these tests can not be expected to either be high or reliable.
- The second document is entitled "ECVAM Skin Irritation Pre-Validation Study - Repeatability and Reproducibility Analysis" (Spielmann H, personal communication) that provides equations to calculate CVs for repeatability and/or reproducibility from a small number of laboratories and small number of replicates at each of the three phases of prevalidation defined by ECVAM (Curren et al. 1995).
- The third document is entitled "Detailed Variability Analysis", which was drafted by Dr. Sebastian Hofmann (ECVAM) for the on-going ECVAM validation study of *in vitro* skin irritation tests (Spielmann H, personal communication). In this document, Dr. Hofmann compares SD and CV values for two skin models. A comparable analysis of SD and CV values is missing in the present evaluation of the reproducibility of *in vitro* methods for eye irritation testing. More importantly, a strategy to evaluate reliability in any further standardization or validation testing must be developed and implemented.

7.1 Selection Rationale for the Substances Used in the IRE Test Method Reliability Assessment

This section is appropriately covered in the IRE BRD.

7.2 Intralaboratory Repeatability and Intra- and Inter-laboratory Reproducibility of the IRE Test Method

The IRE BRD appropriately states that an evaluation of intra-laboratory repeatability and reproducibility could not be carried out because of a lack of quantitative IRE data of replicate

experiments within an individual laboratory. Estimates of interlaboratory CV values for the various endpoint measures were described as 'moderate' (with numbers such as 40% and 84% quoted), leading to the statement that 'efforts to increase the interlaboratory reproducibility of the test method might be warranted'. As a consequence, the conclusions in IRE BRD Section 7.4, and particularly in the final paragraph of this section, seem appropriate for the analysis carried out.

7.3 Availability of Historical Control Data

There appears to be no historical positive control data available because positive controls are not typically included in the studies. The reports considered in the BRD state that negative controls are always included, but the results are not available. Thus, there is insufficient information to evaluate control data.

7.4 Effect of Minor Protocol Changes on Transferability of the IRE Test Method

Improved transparency of the IRE BRD can be achieved by specifically noting that the protocol used by Guerriero et al. (2004) was essentially identical to the protocol provided by SafePharm, as described in Appendix A of the IRE BRD. The main difference in the standardized protocol described in Appendix A is the inclusion of concurrent positive control and (where useful) benchmark substances. Any other differences in the protocol from that provided, or any future protocol revisions, should be specifically justified. It may be useful to contrast the IRE test results obtained in each of the four studies using the SafePharm decision criteria versus the original study decision criteria; good agreement with *in* vivo data would suggest that all existing data from all protocols can be used as validation data.

It would appear that the recommended version of the IRE test is likely to be insensitive to minor protocol changes and to be readily transferable. If the BCOP quantitative assessment of corneal opacity could be incorporated into the IRE test method, it should add objectivity to the test and improve its inter-laboratory reproducibility.

8.0 TEST METHOD DATA QUALITY

8.1 Impact of GLP Noncompliance and Lack of Coded Chemical Use

Review of the BRD supports the conclusion that only Balls et al. (1995) appears to have conducted IRE studies in compliance with GLP guidelines. While the methods in the other studies are explained in detail, there is no way to determine whether the quality of the data generated was impacted by the failure to follow GLP procedures. However, according to the current EU and OECD documents on the validation of toxicity tests 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, when the basic requirements of the GLP procedure have been implemented in a study. The reviewed data appear to be of satisfactory quality.

8.2 Results of Data Quality Audits

No evidence was presented that the original published data were verified for their accuracy against the original experimental data. Such verification may be beyond the scope of the IRE assessment. This section is appropriately covered in the IRE BRD.

8.3 Impact of GLP Detected in Data Quality Audits

Lacking the original test data from the studies conducted to evaluate the IRE, the accuracy of the study results cannot be evaluated. Noncompliance with GLPs is not a mandatory exclusion criterion. All laboratories performing the studies were reputable.

8.4 Availability of Original Records for an Independent Audit

Original raw *in vitro* data for all studies were not available for review; availability and review of raw data would improve the confidence in the data. However, doing retrospective GLP-like audits may not be needed and would be difficult to conduct. 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 (ICCVAM 2003) is reasonable and should be supported.

9.0 OTHER SCIENTIFIC REPORTS AND REVIEWS

9.1 Other Published or Unpublished Studies Conducted Using the IRE Test Method

This section is appropriately covered in the IRE BRD.

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

This section is appropriately covered in the IRE BRD.

9.3 Approaches to Expedite the Acquisition of Additional Data

This section is appropriately covered in the IRE BRD. A *Federal Register* (*FR*) notice (Vol. 69, No. 57, pp. 13859-13861, March 24, 2004) requesting data was published. In addition, authors of published IRE studies were contacted to request original IRE data and *in vivo* reference data.

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

10.1 Extent to Which the IRE Test Method Refines, Reduces, or Replaces Animal Use

The discussion of animal welfare considerations is accurate, and may well be sufficient. The reason for hesitation in drawing a final conclusion about this statement is that the ultimate focus of this effort (i.e., to find a replacement for the Draize test) has a special significance for many individuals and organizations. It is well known that, on a regular basis, rabbits have chemicals

applied to, what we might assume from our own experience, is the most sensitive area of their exterior body surface. The IRE and other alternative tests have the potential to eliminate any distress and discomfort that may arise in the *in vivo* test, and therefore are consistent with the objectives of the 3Rs (i.e., reduction, refinement, or replacement of animal studies).

There is also a separate question which, depending on the answer, could affect animal welfare considerations. This is related to the availability of rabbit eyes from the meat industry and other research/testing applications. If the IRE test progresses in a way that allows it to be considered a valid test method and for it to be widely applied, will there be sufficient "secondary use eyes" available, or is it likely that rabbits would have to be raised simply to provide the organs for this test? Current regulatory standards, such as those promulgated by the U.S. Environmental Protection Agency (EPA), may preclude the use of eyes from rabbits used for other experimental (e.g., toxicological) purposes. Thus, additional information in the IRE BRD about the availability of rabbits used for studies that have no effect on the eye or that are killed for food would be useful. Regardless, rabbits should not be raised and killed specifically for use in this test. In addition, NICEATM should define in the IRE BRD the current policy of U.S. regulatory agencies or GLP impacts regarding the use of eyes from rabbits used for other scientific purposes.

11.0 PRACTICAL CONSIDERATIONS

It appears that with sufficient training and attention to detail that a standardized IRE test protocol could be developed that would be relatively straightforward to use in multiple laboratories and would be expected to produce similar results. Information could be added to the IRE BRD about how inter-laboratory agreement would be verified. This could be general information about what type of materials would be tested and how inter-laboratory variation would be assessed. Although costs of *in vivo* and *in vitro* testing are provided, a more detailed itemization of costs for each test would be useful. The rest of this section in the IRE BRD addresses practical considerations in appropriate detail.

11.1 IRE Test Method Transferability

11.1.1 Facilities and Major Fixed Equipment Needed to Conduct the IRE Test Method
This section is appropriately covered in the IRE BRD with one exception. The BRD should indicate that the perfusion apparatus may not be readily available for purchase and may need to be custom built.

11.1.2 General Availability of Other Necessary Equipment and Supplies
This section is appropriately covered in the IRE BRD.

11.2 IRE Test Method Training

11.2.1 Required Training to Conduct the IRE Test Method
This section is appropriately covered in the IRE BRD. However, in addition, a training video and other visual media on the technical aspects of the assay is recommended, as well as the development and implementation of other approaches in the application of this test method.

11.2.2 Training Requirements Needed to Demonstrate Proficiency

This section is appropriately covered in the IRE BRD.

11.3 Relative Cost of the IRE Test Method

The BRD compares costs between the United States (*in vivo*) and the United Kingdom (*in vitro*); this is inappropriate as costs in the United States are typically greater depending on the current exchange rate. A more appropriate comparison would be between the *in vivo* and *in vitro* costs from a single laboratory or a single country. The BRD should be revised to reflect this concern.

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

This section is appropriately covered in the IRE BRD, except that the BRD should note that the *in vivo* rabbit eye test may be ended in a few hours if the test substance is a severe irritant or corrosive.

12.0 PROPOSED TEST METHOD RECOMMENDATIONS

12.1 Recommended Version of the IRE Test Method

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

The most appropriate version of the IRE test method, which included an assessment of fluorescein staining and epithelial integrity as well as of corneal thickness and opacity, has been identified. However, this version of the IRE has only been conducted in one laboratory (SafePharm, based on Guerriero et al. [2004]), and the available data that were generated using this version are too limited (36 substances classifiable to GHS) to allow an adequate judgment of its accuracy and reliability. Thus, this test method has not yet fully met the ICCVAM criteria for validation (ICCVAM 2003).

However, the Panel concludes that the recommended version of the IRE test method appears to be capable of identifying ocular corrosives/severe irritants in a tiered testing strategy (e.g., GHS). Substances with less acute toxicity or substances that cause damage by slower cellular responses will not be detected by the proposed IRE methodology so some potentially damaging substances might be missed until an *in vivo* test is performed. However, the GHS tiered testing strategy largely obviates this concern.

12.2 Recommended Standardized IRE Test Method Protocol

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

The Panel agrees with the proposed standardized IRE test method protocol in Appendix A of the IRE BRD, with the following comments and suggestions:

- The appropriate sources of rabbit eyes need to be defined. The current policy of some U.S. regulatory agencies (e.g., EPA) in regard to use of eyes from rabbits

used for other scientific studies should be reviewed and updated. The protocol should explicitly state that rabbits should not be raised and killed specifically for use in this test.

- The rationale for the decision criteria included in Appendix A, Table A-3 of the IRE BRD needs to be provided, and its application should be discussed in Appendix A, Sections 7.0-9.0. In addition, appropriate reference substances (positive and negative controls, benchmarks) should be identified, based on the Panel recommendations in regard to the proposed Reference Substances List in the IRE BRD.

Experience with this recommended protocol will help to evaluate its ability to reduce the false negative rate and could guide decisions regarding the need for optimization.

12.1.2 Other Endpoints that Should be Incorporated into the IRE Test Method

First, it is important that an analysis be made of the extent to which leading-edge veterinary and human ophthalmology research and medical practice techniques can be applied to the measurement of corneal damage in the IRE test system.

Second, given the sophistication and variety of currently available methods for the assessment of cellular damage and death, the lack of inclusion of these methods into the IRE test method may be problematic. Validation of this or any other *in vitro* test may require inclusion of additional methods to detect cellular damage, at least in the early stages of test validation.

Third, histopathology, including determining the nature and depth of corneal injury, should be considered when the standard IRE endpoints (i.e., corneal opacity, swelling, and fluorescein retention; epithelial integrity) 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.

Fourth, to maximize the likelihood of obtaining reproducible results, reference photographs for all subjective endpoints (i.e., corneal opacity, fluorescein retention, and histopathology) should be made readily available.

Finally, personnel handling tissue using the proposed IRE test method protocol should be aware of the risk from potential zoonoses and take appropriate protective measures.

12.3 Recommended Optimization and Validation Studies

12.1.1 Recommended Optimization Studies to Improve Performance of the IRE Test Method Protocol

As stated in **Section I - 12.1**, the recommended IRE test method appears to be capable of identifying ocular corrosives/sever irritants in a tiered testing strategy. However, as the relevant IRE test database is so small (36 substances classifiable to GHS) and because there is a lack of data on reproducibility, additional data needs to be considered before an appropriate evaluation of the IRE test for regulatory classification can be conducted. These data may be obtainable from application of the BRD recommended protocol decision criteria (Table A-3 in Appendix A

of the IRE BRD) to data obtained in studies that did not include all aspects of the recommended protocol.

The existing data with the recommended version of the IRE test method indicate a relatively high false positive rate of 33% (8/24) and a very low false negative rate of 0% (0/12). Although the numbers of substances included in these evaluations are very few, these data are encouraging. If additional analyses are needed to corroborate these findings, then the IRE decision criteria should be optimized to reduce the false positive rate without unacceptably increasing the false negative rate within the context of a tiered testing strategy. Also, consideration should be given to exploring the use of a battery of the *in vitro* tests compared in Table 12-2 of the IRE BRD. A battery of tests could be applied based on their individual strengths and weaknesses to improve overall predictability.

Any optimization and validation studies should use existing *in vivo* rabbit eye 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 information obtained (e.g., wound healing) and to minimize the number of animals used.

From a scientific point of view, there is no need to conduct optimization or validation studies until the IRE data that are available in the IRE BRD have been analyzed more thoroughly. Before planning any laboratory studies, the following points should be taken into account:

1. A statistical concept to take into account the variability of the *in vivo* Draize eye test data should be developed. As suggested by Dr. Leon Bruner (Bruner et al., 1996), the CV values for the *in vivo* Draize eye test data should be calculated. High quality *in vivo* data of the Draize eye test will allow a determination of the probability of correct classification when the test is conducted in three rabbits. This calculation has to take into account the relatively low variability at the high and low ends of the Draize scale and the higher variability in the medium range.
2. The repeatability of results obtained with positive and negative and reference substances should be determined both for the Draize rabbit eye test and for the IRE. Thus, a high quality database of *in vivo* and *in vitro* data of reference substances should be established from the existing literature.
3. Decision criteria may be improved by applying advanced statistical methods (e.g., discriminant analysis) to identify the most predictive endpoints and to establish cut off values for classification purposes; this approach has yet to be used for any of the four studies used to evaluate performance of the IRE test method. From a comparison of the decision criteria identified for these studies, a more general set of decision criteria might be derived, which will allow the identification of severely irritating substances when using the recommended IRE protocol.
4. The practical consideration of whether sufficient eyes are available for use in the test (i.e., appropriate sources of rabbit eyes must be identified if further optimization and validation is to proceed).

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.

12.1.2 Recommended Validation Studies to Evaluate Performance of the Optimized IRE Test Method Protocol

Validation of test repeatability and reproducibility with an appropriate range of chemicals is important to the eventual acceptance of the IRE test method in a tiered testing strategy or as a Draize test replacement. A critical aspect of this validation effort is comparing the IRE test results with those obtained *in vivo* in the Draize test, a test that has limitations that have not been completely characterized. The magnitude of these limitations and how to apply this information to *in vitro* validation efforts is unclear and the IRE BRD would benefit from a discussion on this matter.

12.4 Proposed Reference Substances for Validation Studies

See **Section V**.

13.0 IRE BRD REFERENCES

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

Information in two additional references need to be included in of the IRE BRD; these are Bland and Altman (1986), which is a detailed analysis of the variability of EPISKIN™, and an ECVAM prevalidation report on skin irritation repeatability and reproducibility (Spielmann H, personal communication).

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.

Bland JM, Altman DG. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1:307-310.

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

Burton ABG, York M, Lawrence RS. 1981. The *in vitro* assessment of severe eye irritants. Food Cosmet Toxicol 19:471-480.

CEC. 1991. Collaborative study on the evaluation of alternative methods to the eye irritation test. Doc. XI/632/91/V/E/1/131/91 Part I and II.

Curren RD, Southee JA, Spielmann H, Leibsch M, Fentem JH, Balls M. 1995. The role of prevalidation in the development, validation and acceptance of alternative methods. ECVAM Prevalidation Task Force Report 1. ATLA 23:211-217.

ECVAM. 2005. General guidelines for submitting a proposal to ECVAM for the evaluation of the readiness of a test method to enter the ECVAM prevalidation and/or validation process. Available: https://ecvam.jrc.it [accessed 07 February 2005].

Gettings SD, Lordo RA, Hintze KL, Bagley DM, Casterton PL, Chudkowski M., Curren RD, Demetrulias JL, Dipasquale LC, Earl LK, Feder PI, Galli CL, Glaza SM, Gordon VC, Janus MG, Tedeschi JP, Zyracki J. 1996. The CTFA evaluation of alternatives program: An evaluation of *in vitro* alternatives to the Draize primary rabbit eye irritation test. (Phase III) Surfactant-based formulations. Food Chem Toxicol 34:79-117.

Guerriero F, Seaman CW, Olson MJ, Guest RJ, Whittingham A. 2004. Retrospective assessment of the rabbit enucleated eye test (REET) as a screen to refine worker safety studies [Abstract]. Toxicologist 78(S-1):263.

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. 2002. Report of the Stockholm Conference on Validation and Regulatory Acceptance of New and Updated Methods in Hazard Assessment. Paris, France: Organisation for Economic Co-operation and Development.

OECD. 2004a. *In Vitro* Skin Corrosion: Human Skin Model Test. Test Guideline 431. (adopted 13 April 2004). Paris, France: Organisation for Economic Co-operation and Development.

OECD. 2004b. *In Vitro* 3T3 NRU Phototoxicity Test. Test Guideline 432. (adopted 13 April 2004). 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.

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.

[This Page Intentionally Left Blank]