APPENDIX A1

EXPERT PANEL REPORT: EVALUATION OF THE CURRENT VALIDATION STATUS OF *IN VITRO* TEST METHODS FOR IDENTIFYING OCULAR CORROSIVES AND SEVERE IRRITANTS

[This Page Intentionally Left Blank]

Expert Panel Evaluation of the Current Validation Status of *In Vitro* Test Methods for Identifying Ocular Corrosives and Severe Irritants

Expert Panel Final Report

March 2005

Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM)

National Toxicology Program (NTP) Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM)

National Institute of Environmental Health Sciences
National Institutes of Health
U.S. Public Health Service
Department of Health and Human Services

This document is available electronically at:
http://iccvam.niehs.nih.gov/methods/ocudocs/EPreport/ocureport.htm

TABLE OF CONTENTS

IN VITRO OCULAR TEST METHOD EXPERT PANEL ROSTER

- **Robert Scala**, Ph.D. (Panel Chair), Consultant, Tucson, Arizona
- **Sally Atherton**, Ph.D., Professor, Medical College of Georgia, Augusta, Georgia
- **Roger Beuerman**, Ph.D., Professor, Louisiana State University, New Orleans, Louisiana
- **June Bradlaw**, Ph.D., International Foundation for Ethical Research, Rockville, Maryland
- **Ih Chu**, Ph.D., Health Canada, Ottawa, Canada
- **Henry Edelhauser**, Ph.D., Professor, Emory University, Atlanta, Georgia
- **Nancy Flournoy**, Ph.D., Professor, University of Missouri, Columbia, Missouri
- **Donald Fox**, Ph.D., Professor, University of Houston, Houston, Texas
- **James Freeman**, Ph.D., Section Head, ExxonMobil Biomedical Sciences, Inc., Annandale, New Jersey
- **Shayne Gad**, Ph.D., D.A.B.T., A.T.S., Consultant, Gad Consulting Services, Cary, North Carolina
- **Sidney Green**, Ph.D., A.T.S., Graduate Professor, Howard University, Washington, DC
- **Frederick Guerriero**, M.S., Senior Occupational Toxicologist, GlaxoSmithKline, King of Prussia, Pennsylvania
- **A. Wallace Hayes**, Ph.D., D.A.B.T., F.A.T.S., F.I.Biol., F.A.C.F.E., E.R.T., Scientist, Harvard School of Public Health, Andover, Massachusetts
- **Hiroshi Itagaki**, Ph.D., Principal Scientist, Shiseido Co., Ltd., Japan
- **David Lovell**, Ph.D., Reader in Medical Statistics, University of Surrey, United Kingdom
- **Yasuo Ohno**, Ph.D., D.J.S.T.S., Director of Japanese Society of Alternatives to Animal Experiments and Director of Division of Pharmacology, National Institute of Health Sciences, Japan
- **Robert Peiffer**, D.V.M., D.A.C.V.O., Senior Investigator, Merck Research Laboratories, West Point, Pennsylvania
- **Lionel Rubin**, V.M.D., D.A.C.V.O., Emeritus Professor of Ophthalmology, University of Pennsylvania, Philadelphia, Pennsylvania
- **Horst Spielmann**, Dr. Med., Director and Professor, ZEBET at the BfR, Germany
- **Martin Stephens**, Ph.D., Vice President for Animal Research, Humane Society of the United States, Washington DC
- **Katherine Stitzel**, D.V.M., Consultant, West Chester, Ohio
- **Peter Theran**, V.M.D., D.A.C.V.I.M., Vice President Animal Science, Massachusetts Society for the Prevention of Cruelty to Animals, Novato, California
- **Scheffer Tseng**, M.D., Ph.D., Director, Ocular Surface Research and Education Foundation, Miami, Florida
- **Philippe Vanparys**, Ph.D., Senior Research Fellow, Johnson and Johnson, Belgium

[This Page Intentionally Left Blank]

PREFACE

This is an independent report of the Expert Panel ("Panel") organized by the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) and the National Toxicology Program (NTP) Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM). The report summarizes discussions, conclusions, and recommendations of the public meeting of the Panel that was held at the National Institutes of Health in Bethesda, MD on January 11 and 12, 2005. The ICCVAM and the Ocular Toxicity Working Group (OTWG) will consider the report, along with public comments, to prepare test method recommendations for U.S. Federal agencies. ICCVAM test method recommendations will be forwarded to U.S. Federal agencies for consideration and action, in accordance with the ICCVAM Authorization Act of 2000 (P.L. 106-545).

NICEATM, in coordination with the OTWG and ICCVAM, prepared comprehensive draft background review documents (BRDs) reviewing the available data and information for four *in vitro* test methods: the Isolated Rabbit Eye (IRE), the Isolated Chicken Eye (ICE), the Bovine Corneal Opacity and Permeability (BCOP), and the Hen's Egg Test - Chorioallantoic Membrane (HET-CAM) assay. Each BRD was based on studies using the test method, and data and information submitted in response to a 2004 *Federal Register* (*FR*) request for submission of *in vitro* data for each of these test methods and for submission of high-quality *in vivo* rabbit eye test data (*FR* notice Vol. 69, No. 57, p. 13859-13861; March 24, 2004). All four draft BRDs were made publicly available on the ICCCVAM/NICEATM website (http://iccvam.niehs.gov) or from NICEATM on request.

NICEATM, in collaboration with the OTWG and ICCVAM, organized an independent Expert Panel review of the methods in January 2005. Comments from the public and scientific community were solicited and provided to the Panel for their consideration (*FR* notice Vol. 69, No. 212, p. 64081-2; November 3, 2004).

The Panel was charged with:

- Evaluating, for each of the four *in vitro* test methods, the extent and adequacy that each of the applicable ICCVAM validation and acceptance criteria[1]
  - have been addressed, based on available information and data, or
  - will be addressed in proposed studies for the purpose of identifying ocular corrosives and severe irritants in a tiered testing strategy.
- Developing, for each of the four *in vitro* test methods, conclusions and recommendations on:
  - current usefulness and limitations of each of the four test methods for identifying ocular corrosives and severe/irreversible irritants
  - the test method protocol that should be used for future testing and validation studies
  - the adequacy of proposed optimization and/or validation studies
  - the adequacy of reference substances proposed for future validation studies

[1] ICCVAM submission guidelines can be obtained at: http://iccvam.niehs.nih.gov/docs/guidelines/subguide.htm

During the public meeting in January 2005, the Panel discussed the current validation status of each of the four *in vitro* test methods. The Panel also provided formal comment on each of the BRDs and made recommendations for revisions to each document. In addition, the public were provided time at the public meeting to comment on the BRDs. The Panel then provided final endorsement regarding the validation status of each of the test methods.

EXECUTIVE SUMMARY

Introduction

This report describes the conclusions and recommendations of the Expert Panel ("Panel") regarding the validation status of four *in vitro* ocular toxicity test methods: the Isolated Rabbit Eye (IRE), the Isolated Chicken Eye (ICE), the Bovine Corneal Opacity and Permeability (BCOP), and the Hen's Egg Test - Chorioallantoic Membrane (HET-CAM) assays. Those areas of each background review document (BRD) not mentioned in this report were considered adequate and acceptably accurate by the Panel.

The Isolated Rabbit Eye Test Method

The Panel concluded that the IRE BRD proposed version of the IRE test method appears to be capable of identifying ocular corrosives/severe irritants in a tiered-testing strategy with the caveat that the accuracy of this test method be corroborated using a larger number of substances and that reliability analyses be conducted when additional data become available. This recommendation was based on the relatively small number of substances (n=36) tested using the proposed IRE test method version and because only one laboratory (SafePharm, Derby, United Kingdom) had experience using this test method protocol. The Panel agreed that the recommended standardized protocol described in the IRE BRD, which included fluorescein penetration and evaluation of epithelial integrity as endpoints, was appropriate and significantly improved accuracy when compared to other versions of the IRE test method.

With respect to IRE optimization and validation, the Panel recommended that additional data be requested from users of this test method and that analyses of additional data be conducted. The Panel also suggested, that as the IRE test method had a relatively high false positive rate of 33% (with a false negative rate of 0%), optimization of the decision criteria to minimize the false positive rate without appreciably increasing the false negative rate is needed. This may best be accomplished using statistical methods (e.g., discriminant analysis) to improve the decision criteria for the IRE. The Panel noted that any further optimization or validation should be conducted using existing data. Additional animal studies would only be conducted if important data gaps were identified and such studies would be carefully designed to maximize the amount of pathophysiological information obtained (e.g., wound healing). A minority opinion of one Panel member stated that no additional animals should be used for this purpose. The Panel also recommended that a high quality database of *in vivo* and *in vitro* data of reference substances be established from existing literature and new data.

The Panel proposed several modifications to the recommended standardized protocol. These include identification of an appropriate source of rabbits (e.g., an abattoir such as Pel-Freeze) to provide eyes to be used in the IRE, and inclusion of an explicit statement that that rabbits should not be bred and killed specifically for use in the IRE test method. The policies of the various U.S. regulatory agencies with respect to use of rabbits in the IRE that were used in previous tests or experiments needs to be reviewed and updated as it impacts the number of animals available for use in this test. The decision criteria used to identify ocular

corrosives/severe irritants should be clearly identified and a rationale provided for how it was developed. For any future studies, defined positive, negative, and benchmark substances need to be identified based on the proposed list of reference substances. In addition, the Panel proposed that the National Toxicology Program (NTP) Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) facilitate the development of a standardized histopathology scoring system for corneal damage, along with an appropriate atlas with visual aids. In addition, 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 should be developed (e.g., corneal opacity, fluorescein penetration, histopathology) to aid training and transferability. A discussion of the use of proper safety precautions when handling animals and isolated eyes and awareness of the risk of contamination with potential zoonoses should also be included in the IRE BRD.

The Isolated Chicken Eye Test Method

The Panel concluded that the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) criteria for validation (ICCVAM 2003) have not been fully met for the ICE test method. Cited deficiencies include: the intralaboratory 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; and detailed drawings/diagrams of the superfusion apparatus have not been made available to allow for transferability of the experimental setup. However, the Panel concluded that the ICE test method can be used in the identification of ocular corrosives/severe irritants in a tiered testing strategy, with specific limitations. Specifically, the Panel noted that alcohols tend to be overpredicted, while surfactants tend to be underpredicted. The Panel also recognized that solids and insoluble substances may be problematic in the ICE test method, since they may not come in adequate contact with the corneal surface, resulting in underprediction. Therefore, the Panel concluded that the low overall false positive rate (8% to 10%, depending on the regulatory classification scheme evaluated) indicates that the ICE test can be used at present to screen for severe eye irritants/corrosives. However, given the high false positive rates calculated for a small number of alcohols (50% [5/10]), the Panel noted that caution should be observed when evaluating ICE test results with this class of substances.

The Panel recognized that the recommended protocol is based on the original ICE protocol, which has changed only slightly since its development. However, there was concern expressed as to whether the appropriate number of eyes (n=3) is being used to ensure optimum performance. Therefore, the Panel recommended that the potential effects of using more than three eyes on the accuracy and reliability of the ICE test method be the subject of a formal study. The Panel also questioned the utility of using maximum mean scores, and thus to ensure optimum performance, recommended a formal evaluation of the most appropriate mathematical approach.

The Panel identified potential methodological areas of improvement to the protocol, including moving the superfusion apparatus to a horizontal position to obviate the need for test eye removal during dosing, adding centering lights to the optical pachymeter to ensure

consistent central corneal thickness measurements across laboratories, and inclusion of concurrent negative and positive control eyes (at least 3 per group). In addition, histopathology, including determining the nature and depth of corneal injury, was recommended for inclusion in the protocol when the standard ICE endpoints (i.e., corneal opacity, swelling, and fluorescein retention) produce borderline results. With this in mind, the development of a standardized scoring scheme using the formal language of pathology to describe any effects was advocated, along with defining the appropriate circumstances under which histopathology would be warranted. The Panel noted the need for reference photographs for all subjective endpoints (i.e., corneal opacity, fluorescein retention, and histopathology) to ensure consistency among laboratories.

Given the limited amount of ICE reliability data, additional studies using the recommended ICE test method protocol were suggested to better characterize the repeatability and the intra- and inter-laboratory reproducibility of the test method. The Panel recommended also optimization studies that were considered to be potentially useful for improving ICE test method performance. These studies included efforts to optimize the decision criteria used for identifying corrosives and severe irritants, an evaluation of the impact of routinely performing replicate experiments, and an evaluation of the impact of variations in the time between death and testing of the chicken eyes on test method performance.

The Panel specified that any optimization and validation studies should use existing animal data, if available, and that additional animal studies should only be conducted if important data gaps are identified. A minority opinion of one Panel member stated that no additional animals should be used for this purpose.

The Bovine Corneal Opacity and Permeability Test Method

The Panel concluded that the BCOP BRD proposed version of the test method has been shown to have adequate accuracy and reliability for detecting corrosive or severe eye irritants in the tiered testing scheme outlined in the BCOP BRD, with the following caveats:

- The test should not be used to identify corrosive or severely irritating ketones, alcohols, and solids. Further optimization and validation are necessary before these classes of materials can be assessed with this test.
- It needs to be confirmed that the BCOP test method can identify, as well as or better than the Draize test, those substances known to cause serious eye injury in humans. It appears from the list of chemicals tested that at least some of these substances have been tested in BCOP (e.g., floor strippers and heavy duty cleaners).
- A histopathological examination should be added to the test unless the test substance is from a class of materials known to be accurately predicted using only opacity and permeability in the BCOP assay.

The Panel concluded that the BRD proposed protocol for the BCOP test method is useful for identification of severe or corrosive ocular irritants in the tiered testing scheme outlined in the BCOP BRD, with the caveats noted above, as well as those noted below:

- 0.9% sodium chloride should be used instead of distilled water as the test substance diluent.
- Determination of osmolarity and pH of test solutions should be conducted.
- The optimum age range for cattle should be determined.
- Users should be aware of zoonoses, including the possibility of Bovine Spongiform Encephalopathy (BSE).
- Concurrent negative, positive, and benchmark controls should be used.

With respect to suggested modifications to improve performance (accuracy and reliability) of the recommended standardized protocol for the BCOP test method, the Panel recommended the following modifications:

- Use of the larger holder as suggested by Ubels et al. (2002, 2004).
- Re-examine the use of the calculated total score when the endpoint is severe injury only.
- Changes to the medium used to bathe the eyes, including a determination of whether fetal bovine serum is needed.

While the Panel believes these modifications are important, the Panel concluded that the data presented in the BCOP BRD support use of the BCOP assay in its current form for identifying ocular corrosives and severe irritants other than alcohols, ketones, and solids in a tiered testing strategy for regulatory hazard classification and labeling purposes.

The Panel also suggested that histopathological examination be added to the recommended test protocol unless the test substance is from a class of materials known to be accurately predicted using only opacity and permeability in the BCOP assay.

While actually a change to the BCOP method, the Panel suggested the possibility of using the porcine eye as a model for the human eye. The Panel recognizes that this change would require complete validation, but wants to be sure this possibility is considered for future work.

During a vote on Section 12.2 (Recommended Standardized Test Method Protocol) of the BCOP report at the Panel meeting, three panel members expressed minority opinions. Dr. Freeman abstained from voting on Section 12.2 because he believed the discussion on this section had not been satisfactorily resolved due to time constraints. Drs. Stephens and Theran did not agree with the final language presented for Section 12.2 because they believed the BCOP group members withdrew their original summary conclusion under undue pressure.

Regarding recommended optimization studies to improve performance (accuracy and reliability) of the recommended BCOP test method protocol, the Panel recommended using a larger holder similar to that suggested by Ubels et al. (2002), re-examining the use of the calculated total score when the endpoint is serious injury only, changing the medium used to bathe the eyes, using antibiotics if eyes are kept above 0 °C, and defining appropriate ages of donor animals. While the Panel feels these improvements are important, it believes the data presented in the BRD are sufficient for supporting use of the BCOP assay in identifying

ocular corrosives and severe irritants, except for alcohols, ketones and solids, in a tiered testing strategy for regulatory hazard classification and labeling purposes.

With respect to the recommended validation studies to evaluate performance of the optimized BCOP test method protocol, the Panel concluded validation studies, or submission of additional data supporting the three-minute exposure time suggested for volatile solvents, will be necessary before the BCOP test method can be recommended for use with alcohols and ketones. Validation studies or submission of additional data will be necessary before the BCOP test method is acceptable for solids. The Panel concluded the information in the BCOP BRD, along with the Panel's suggestions, is sufficient to support the use of this test method to identify severe irritants and corrosives, with the exception of alcohols, ketones and solids, in the tiered testing scheme described in the BRD.

The Panel concluded that an additional validation study is not necessary for the recommended additional histopathological examination to the BCOP test method. Although adding histology to the BCOP assay involves additional endpoints, current practice has not been to insist on validation of histopathological examination when it is added to an *in vivo* test method. A standardized histopathological scoring system was suggested by the Panel, but this should be arrived at by the experts in the field and will not require validation. NICEATM/ICCVAM should facilitate the development of a histopathological scoring system for corneal damage (with visual aids). Changes in the calculation method for the BCOP test score, or the use of the individual endpoint data instead of a calculated score also do not need to be validated.

When validation studies are conducted, the Panel believes the studies proposed in the BCOP BRD are appropriate but should be limited to the classes of test substances in question. Validation studies should be carefully planned. Tests should first be done to confirm that any modifications of the protocol do not decrease reliability. Once the inter- and intra-laboratory variability is defined, it will not be necessary to have a large number of laboratories test every chemical in the validation study. Validation should focus on the class of chemicals in question. The study should involve a very small number of experienced laboratories with only a limited number of duplicate samples at each laboratory.

Any validation or optimization 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 information obtained (e.g., wound healing) and to minimize the number of animals used.

With respect to Section 12.3 of the BCOP report, one Panel member, Dr. Stephens expressed a minority opinion. 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.

The Hen's Egg Test - Chorioallantoic Membrane Test Method

The Panel concluded that, for the purpose of detecting severe eye irritants in the tiered-testing strategy outlined in the HET-CAM BRD, the HET-CAM test has been shown to be useful for identification of severe or corrosive ocular irritants. The Panel stated that the high false positive rate was a limitation of the HET-CAM test method. It was proposed that positive results from the HET-CAM test method could be re-tested in a modified HET-CAM test method (e.g. using a lower concentration of test substance) to confirm the results. Alternatively, substances producing a positive result could be tested in a different *in vitro* test method (e.g., ICE, IRE, BCOP). Substances producing negative results (e.g., HET-CAM score defined as nonirritant, mild irritant, or moderate irritant) would follow the tiered-testing strategy.

It was agreed that the most appropriate version of the HET-CAM test method for use in a tiered-testing strategy is the test method protocol recommended in the HET-CAM BRD. The proposed HET-CAM standardized test method protocol is adapted from the one by Spielmann and Liebsch (INVITTOX 1992). The proposed standardized test method protocol contains negative controls, solvent control (if appropriate), positive controls and benchmark controls (if appropriate). The method also recommends using the time required for an endpoint to develop as the criteria for assessing irritation potential (IS(B) analysis method). The Panel stated that procedures for applying and removing solids from the chorioallantoic membrane (CAM), which may adhere to the CAM and demolish the CAM upon removal, should be included in the standardized test method protocol provided in the HET-CAM BRD.

Due to the numerous variations in the test method protocols and different analysis methods that have evolved since the development of the test method, the Panel stated that the use of a standardized test method protocol in future studies would allow for new data to be generated. These data would allow further evaluation of the usefulness and limitations of the recommended test method protocol.

With regard to optimization of the recommended standardized test method protocol, the Panel stated that a retrospective analysis should be conducted to determine if different decision criteria might enhance the accuracy and/or reliability of the test method for the detection of ocular corrosives and severe irritants, as defined by the European Union (EU 2001), United Nations Globally Harmonized System (UN 2003), and the U.S. Environmental Protection Agency (EPA 1996) classification systems. The Panel proposed the use of a modular approach to validation to identify needed validation modules (e.g., interlaboratory reliability) and focus on evaluating those modules.

The Panel stated that the recommendation to optimize and to use an optimized method should not minimize the value of data already obtained with the method of Spielmann and Liebsch (INVITTOX 1992). As some laboratories already apply the method of Spielmann and Liebsch (INVITTOX 1992), the data generated in these laboratories should still be valid and be used for labeling of ocular corrosives and severe irritants. The Panel proposed that an optimized test method may be used when a positive finding is obtained in the HET-CAM test

method of Spielmann and Liebsch (INVITTOX 1992); the substance could be re-tested in the optimized test method protocol.

The Panel further stated that inclusion of different endpoints (e.g., trypan blue absorption, antibody staining, membrane changes, etc.) for evaluation of irritancy potential may increase the accuracy of the HET-CAM test method. It was proposed that these additional endpoints may help reduce the number of false positives observed in the HET-CAM test. The Panel suggested that these endpoints could be included, but were not required, during optimization of the HET-CAM test method.

With respect to validation of the HET-CAM test method, the Panel agreed that if the test method were optimized and modifications made to the test method protocol had a major impact on the conduct of the study, a validation study should be conducted.

The Panel specified that any optimization and validation studies should use existing animal data, if available, and that additional animal studies should only be conducted if important data gaps are identified. A minority opinion of one Panel member stated that no additional animals should be used for this purpose.

The Panel further recommended that an evaluation be conducted to determine the relationship or predictability between the short-term effects observed in the HET-CAM and long-term effects observed in rabbits or humans be conducted. The Panel proposed that such an evaluation may provide additional support for the use of the HET-CAM method to assess the delayed and long-term effects of ocular corrosives and severe irritants.

Proposed List of Reference Substances for Optimization or Validation Studies and to Use in Establishing Performance Standards

The Panel reviewed the adequacy and completeness of the proposed list of reference substances and concluded that the list of proposed substances is comprehensive, the substances appear to be readily available and in acceptably pure form, and the range of possible ocular toxicity responses in terms of severity and types of lesions appears to be adequately represented. The Panel also concluded that, while it is recognized the selection of reference substances is in part limited by the availability of appropriate *in vivo* reference data, the current list has too many substances and is unwieldy, surfactants are over-represented and thus could be reduced in number, and more inorganic substances should be added, if feasible. The Panel also recommended that substances known to induce severe ocular lesions in humans should be included in the list, even in the absence of rabbit data. For all validation studies, Material Safety Data Sheets (MSDS) for the recommended substances should be provided (e.g., a coded MSDS); also prestudy safety briefings should be conducted routinely. Finally, the Panel recommended that an assessment based on the ranking of experimental data for severity for both the reference test method and the *in vitro* test, using the proposed reference substances, be conducted routinely.

For any future validation studies that are performed subsequent to protocol optimization, the Panel recommended that a two-staged approach be used to evaluate accuracy and reliability.

Accordingly, the first stage would evaluate test method reliability using a subset of substances that could be tested in multiple laboratories, followed by a second stage encompassing a larger number of substances to evaluate test method accuracy. The Panel suggested that the accuracy assessment include a statistical analysis of the ranking of experimental data for severity for both the *in vivo* reference method and the *in vitro* test.

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.

12.2.2 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

12.3.1 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.

12.3.2 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.

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 the 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 ± 1.5°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 ([corneal thickness at time t - corneal thickness at time 0/corneal thickness at time 0] X 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[2] 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] 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.

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 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 severre 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. 2nd 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.

[This Page Intentionally Left Blank]

Bovine Corneal Opacity and Permeability Test Method

[This Page Intentionally Left Blank]

III. BOVINE CORNEAL OPACITY AND PERMEABILITY TEST METHOD

1.0 BCOP TEST METHOD RATIONALE

1.1 Scientific Basis for the BCOP Test Method

1.1.1 Mechanistic Basis of the BCOP Test Method

This Section of the BRD discusses the mechanistic basis for current test methods (i.e., the *in vivo* rabbit eye test) and the BCOP test method that is proposed as the initial test in a battery of tests to evaluate the ocular irritancy of new substances. The use of viable corneal tissue provides similarity to the actual system of interest -- the human eye. Opacity is an important endpoint in both test methods (BCOP and the *in vivo* rabbit eye test) and the human eye, although the BCOP test system as outlined in the proposed protocol does not allow one to differentiate the mechanistic cause of the corneal opacity. The BRD mentions only one mechanism of corneal opacity, but it is recognized that opacity can occur either because of severe injury, possibly with protein denaturation of the epithelial layer, or by swelling of the epithelium and/or corneal stroma. The latter is usually due to loss of the barrier function of the epithelial layer. Histopathological examination of the cornea will provide information useful to identify these mechanisms. Permeability is a measure of the integrity of the corneal epithelium and adds important information on the degree of injury that would be predicted by the test.

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

The BCOP method differs from the *in vivo* method in that it only evaluates the potential of a test material to damage the cornea of the eye. Some materials can cause serious corneal injury without appearing to change opacity or permeability immediately. For instance, cell death (e.g., apoptosis, necrosis) can selectively be induced by some chemicals (such as mustard gas), and such death may take place in keratocytes and vascular endothelium. Previous Expert Panels have suggested that methods to determine the irritation potential of test materials via the ocular route need to consider both damage to the cornea and damage to the vasculature and stem cells that grow in to repair the cornea (Nussenblatt et al. 1998). These cells, which are located at the rim of the cornea within the sclera (Schermer et al. 1986), are not normally evaluated in either the *in vivo* or *in vitro* systems.

The BRD mentions that injury to the sclera is not assessed in the BCOP assay, but no information is presented on whether serious damage to the sclera, including the limbal stem cells, can occur without evidence of injury to the cornea. Maurer and Jester in their series of papers, which report on *in vivo* ocular irritation studies of 23 materials that caused minimal to severe eye irritation, did not identify any materials that injured limbal stem cells without causing histological changes elsewhere in the cornea (reviewed in Maurer et al. 2002). Agents such as mustard gas can produce this type of damage in humans. Damage to the remainder of the eye and/or systemic toxicity is not addressed by this assay.

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

Rabbit and bovine corneas both differ from human cornea. It is not known how these differences affect the ability of either the rabbit or bovine cornea to predict the response in the human, but the use of the *in vivo* rabbit test has apparently protected human populations from serious injury for many years.

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

The BCOP BRD does not include a discussion of the results of the studies by Maurer and Jester (reviewed in Maurer et al. 2002) in which they followed, using sequential *in vivo* confocal microscopy, the progression of eye lesions within the same animal over time. This extensive work was done on groups of rabbits exposed to 23 substances including surfactants, acids, alcohols, aldehydes, alkalis, bleaches, an aromatic amine, and a ketone. In addition to the sequential confocal examination of each animal, histopathological evaluations and live/dead staining studies were also done to confirm the results. These studies showed that "regardless of the process leading to tissue damage, extent of initial injury is the principal, mechanistic factor determining the outcome of the ocular irritation" (Maurer et al. 2002). These studies support the use of short-term assays to evaluate the long-term outcome of test substance exposure and should be discussed in the BCOP BRD. In addition, in human medicine, Hughes' classification is used to grade the severity of chemical injuries and predict the outcome based on initial injury. The classification includes the extent of corneal opacity (cloudiness) as judged by the visibility of the iris details, and the extent of limbal ischemia (based on the circumference involved) (Nussenblatt et al. 1998). The Draize and *in vitro* tests do not specifically examine limbal changes (Hughes 1946; McCulley 1987). More recent work supports the proposition that limbal stem cell injury predicts serious eye damage (Tseng and Sun 1999).

The BCOP BRD does not include a discussion of how protective mechanisms affect the outcome of the *in vivo* studies. Protective mechanisms are extremely important and are built into *in vivo* testing, but are absent in *in vitro* testing. The protective mechanisms include tearing and reflex blinking due to the activation of sensory trigeminal pathways, which in humans is interpreted as pain. However, note that for some test substances (e.g., solids), blinking can also induce mechanical damage *in vivo*, contributing to a higher degree of irritation. If an irritant not only causes cell/tissue damage, but also "denervates" the ocular nerve (sensory), this will alter the dynamics leading to more severe damage. This issue is not well covered in the BCOP BRD. The BCOP test proposed does not mimic these mechanisms. Consideration of the buffering effect of tears may be relevant to the apparent overprediction of injury by the BCOP for very dilute acids and bases.

The BCOP BRD reviews the important physiological and anatomical differences between the human eye and the rabbit eye, but provides little information with which to compare the bovine eye, other than the thickness of the corneal epithelium.

1.2 Regulatory Rationale and Applicability

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

The endpoint of corneal opacity is measured in both the BCOP and *in vivo* methods. However, the BCOP test method does not measure changes in the iris and conjunctiva, and does not identify substances systemically toxic via ocular exposure. The BRD states the BCOP does not assess reversibility without including a discussion of the work mentioned above (i.e., Maurer et al. 2002; Tseng and Sun 1999) that supports the concept that the final outcome of an eye injury can be predicted by the extent of the initial injury.

The BCOP BRD explains the current regulatory methods, including the differences between the three scoring systems (i.e., EPA 1996, EU 2001, UN 2003). The BRD points out clearly that there are no data comparing the results in the *in vivo* rabbit test to similar human exposure, except for very mild substances. Human ocular irritancy studies are not routinely conducted, and when they are only substances intended for use in or around the human eye (e.g., contact lens solutions, cosmetic formulations) are evaluated (Bruner et al. 1998; Cater et al. 2004). Historical experience indicates the rabbit test has protected human populations using existing scoring systems of the Federal Hazardous Substances Act (FHSA), EPA, and the EU.

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

In addition to data from the BCOP test method, all other data on the test substance should be considered in the hazard and risk assessment of eye exposure, including the systemic toxicity of the material, information on related chemicals, possibly a structure activity or structure property analysis, its physicochemical properties, and the results of dermal testing. As *in vitro* tests become available for specific endpoints, toxicologists in industry and government will need to rethink their testing strategies, as it is very unlikely that the *in vitro* tests will be able to replace the current animal tests on a one-for-one basis.

Based on the information presented in the BRD, the Panel believes a sufficient mechanistic basis for the BCOP test method has been established.

2.0 TEST METHOD PROTOCOL COMPONENTS

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

2.1.1 Materials, Equipment, and Supplies

The suggested protocol does provide a standard procedure for obtaining eyes. The optimum age range for cattle should be determined; however, until this is evaluated, eyes should be obtained from young adult animals of 18-48 months of age. The protocol states eyes should be collected in a suitable container in Hanks Balanced Salt Solution (HBSS) containing antibiotics, and the container then maintained on ice. Use of antibiotics is questioned since they are not effective at 4˚C and because of this there is no rationale for their use if the eyes

are adequately refrigerated. Eyes can probably be stored longer than the five hours stated in the protocol, possibly up to 12 hours, but this needs to be confirmed by careful examination of the eyes prior to testing. The single most important criterion for acceptance of eyes for use in the assay should be the careful examination of the eyes prior to dissection of the cornea and subsequent examination of the corneal preparation just prior to testing.

Eyes from animals that are sick or weakened should not be used because of concerns about zoonotic diseases, including Bovine Spongiform Encephalopathy (BSE). Standard laboratory precautions to protect against zoonotic diseases, such as use of gloves and eye protection, should be followed.

The Panel does not agree that sterile water is the preferred solvent for preparing solutions and suspensions; 0.9% NaCl is preferred. If solutions are diluted with distilled water, a distilled water control also needs to be evaluated. Distilled water itself can cause corneal damage and with edge damage from the corneal crush from the blocks, distilled water will further break down the epithelial barrier and cause corneal edema, as well as edema along the crush edge. Osmolarity and pH of the test solutions should be measured and recorded.

The BCOP assay should be optimized to decide which materials are used to bathe the cornea. It may not be necessary to add Fetal Bovine Serum (FBS), or even use Minimum Essential Medium (MEM). Balanced salt solutions designed for ophthalmic use may be more appropriate and may decrease cost as well.

The holder/clamp referenced in the BCOP BRD protocol does not maintain the bovine cornea with its natural curvature. The bovine cornea is oval in shape and has a radius of curvature. However, the blocks described in the BCOP BRD (Section 2.0) to mount the cornea are flat with round holes (17 mm); thus, when the cornea is clamped, the cornea surface can wrinkle, resulting in a loss of both epithelial and endothelial cells. Also, when the epithelium and endothelium wrinkle, there is loss of the corneal barrier function. The cornea needs to be mounted by clamping the sclera and the block needs to be designed with a radius of curvature appropriate for the bovine cornea.

Clamping directly on the cornea as described in the protocol leads to crush injury of the cornea. The crush zone, as well as the treatment area, are clearly seen in the picture on page 6 of the public comment letter dated November 18, 2004, from Drs. Harbell and Curren of the Institute for *In Vitro* Sciences (IIVS). The crushed area (edge damage) may have as much surface area as the treatment area. With edge damage, permeability of the sodium fluorescein will increase and the corneal response may be more severe as well as more variable. The use of the improved holder may also allow detection of limbal changes.

The papers by Ubels et al. (2002, 2004) referenced in the BCOP BRD and submitted as public comments (letter dated December 16, 2004, from Dr. Ubels) provide a good design of a holder large enough to clamp on the sclera and with the appropriate dimensions to maintain the natural curvature of the cornea.

2.1.2 Dose-selection Procedures

The BRD states dose-selection procedures are not relevant for the BCOP. However, there is discussion of various ways of dosing the eyes and dilution of the test materials in other sections.

2.1.3 Endpoint(s) Measured

Histopathological examination must be included unless the substance is from a class of materials known to be accurately predicted using only opacity and permeability in the BCOP assay.

A basic grading system that stresses utility needs to be established for the histopathological evaluation.

2.1.4 Duration of Exposure

The duration of exposure needs to be standardized (10 minutes - 4 hours) for certain types of test materials. In several places, the BCOP BRD discusses the fact that 10-minute exposure times cause volatile solvents to be overclassified by this method, but the protocol does not recommend a 3-minute exposure for these materials. This should be resolved before the protocol is finalized for volatile solvents.

The problem of the irritant potential of solids also needs to be defined more carefully. The very long exposures used are problematic, but since the application of solids to the conjunctival sac in Draize test rabbits also seems to be non-real-world, it is necessary to optimize the exposure time to solids in the BCOP assay. Perhaps further consideration should be given to the exposure method described by Casterton et al. (1996) for solid materials. Until these areas are optimized, the protocol does not appear to be appropriate for alcohols, ketones, and solids.

2.1.5 Known Limits of Use

The BCOP BRD discusses various known limitations. Based on information presented below (**Section III - 2.7**), the protocol outlined in the BRD, even with the additions described, is not appropriate for alcohols, ketones, and solids.

2.1.6 Nature of the Response(s) Assessed

Histopathological examination must be added unless the test substance is from a class of materials known to be accurately predicted using only opacity and permeability in the BCOP assay.

A basic grading system that stresses utility needs to be established for the histopathological examination.

2.1.7 Appropriate Controls and the Basis for Their Selection

As discussed in the BRD, every time a BCOP assay is run, a concurrent positive and a negative control needs to be included. A list of benchmark controls for common classes of chemicals should be suggested. Consideration should be given to the choice of a positive

control liquid that is not an alcohol. Identification of reference substances that are part of the performance standards developed for the validated test method must be added.

2.1.8 Acceptable Range of Control Responses

Historical values for each testing facility should be used to set an upper value for the negative control and the acceptable range of values for the positive control.

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

The discussion and evaluation in the BCOP BRD are appropriate.

2.1.10 Type of Media in Which Data are Stored

Storage of data should comply with current GLP guidelines.

2.1.11 Measures of Variability

The discussion and evaluation are appropriate in the BCOP BRD.

2.1.12 Statistical or Nonstatistical Methods Used to Analyze the Resulting Data

The discussion and evaluation are appropriate in the BCOP BRD.

2.1.13 Decision Criteria and the Basis for the Algorithm Used

Because the BCOP test method proposed by the BRD is specifically for identification of ocular corrosives or severe irritants, the use of the calculated endpoint score and its cutoff point (i.e., decision criteria) should be re-examined. It may be that in comparison with the GHS classification system, examination of the individual scores or a different cutoff point for the calculated score would improve the accuracy and/or reduce the variability of the test. Finally, the use of the permeability endpoint only for some surfactants, but not all, is problematic. It may be that all surfactants should be evaluated using at least permeability and histopathology (as appropriate).

2.1.14 Information and Data That Will be Included in the Study Report

The opacitometer and corneal holder need to be carefully described in the test report.

2.2 Basis for Selection of the Test Method System

The discussion and evaluation in the BCOP BRD are appropriate.

2.3 Identification of Proprietary Components

The corneal holder should be carefully described in the protocol. Specifications for the type and use of the opacitometer should also be included in the protocol.

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

The discussion and evaluation are appropriate in the BCOP BRD.

2.5 Study Acceptance Criteria for BCOP Test Method

The discussion and evaluation in the BCOP BRD are appropriate.

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

The discussion in the BCOP BRD is appropriate and the bases for the modifications are described adequately.

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

Solutions should be diluted in 0.9% NaCl whenever possible rather than in distilled water. With edge damage from the corneal crush from the holders, distilled water will further break down the epithelial barrier and cause corneal edema as well as edema along the crush edge. Distilled water itself can cause corneal damage. If solutions are diluted with distilled water, a distilled water control also needs to be evaluated.

The osmolarity and pH of test solutions should be measured and recorded. Solutions with osmolarity above 1000 are known to damage corneal epithelium.

Histopathological examination should be added to the recommended test protocol unless the test substance is known to be accurately predicted using only opacity and permeability.

Rinsing procedures should be optimized as a future improvement, particularly for viscous substances and solids.

With the addition of histopathology, the protocol as described in the BCOP BRD is appropriate for test materials other than alcohols, ketones and solids for the identification of corrosives and severe irritants in the test scheme described in the BRD. The Panel believes the other proposed changes could improve the test by reducing its variability and should be investigated as part of a continuing effort to improve the test.

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

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

Of the eight validation studies, three (Balls et al. 1995; Gautheron et al. 1994; Casterton et al. 1996) employed a broad range of chemical classes and products, and are considered adequate.

A total of 166 substances and formulations were evaluated in the eight studies. While the number of substances is considered adequate in the validation studies, methodological differences exist among these studies.

The Panel has encountered in human clinical practice materials that can cause severe eye damage without corneal opacity (Tseng S, personal communication). The Panel would like to be sure that representative types of these materials (e.g., heavy duty cleaning products for oven cleaning and drain cleaners) have been included in the prior validation studies. Materials known to be severe eye irritants in humans, if they have not already been evaluated in the BCOP assay, should be tested in the assay.

Better characterization of physicochemical data on all the test substances is needed.

3.2 Coding Procedures Used in the Validation Studies

Coding is important; if it is not used, it may affect the data quality. Without coding procedures, concern may be raised regarding potential bias and quality of the *in vitro* test data. Except for one study (Casterton et al., 1996), the other studies appeared to employ coded substances. The coding procedures for these studies were considered adequate.

In summary, the data reviewed from prior validation studies in the BCOP BRD are considered adequate.

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

This section of the BCOP BRD 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 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 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 according to the EPA (1996), EU (2001), and GHS (UN 2003) classification systems. These systems as 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 not be adequate for use in evaluating or making distinctions between *in vitro* methods and their suitability for chemical or product class evaluations. In addition to the analyses conducted in the BCOP BRD, the Panel suggests an assessment based on ranking of experimental data for severity for both the reference method and the *in vitro* test.

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

In the case of the BCOP BRD, original study records, such as laboratory notebooks and raw data entry sheets were not obtained 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 is adequate (as is the case for the studies evaluated in the BCOP BRD), the study should be used.

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

As far as the *in vivo* studies used for the accuracy analyses in Section 6.0 of the BCOP BRD, Balls et al. (1995) and Southee (1998) explicitly state GLP guidelines were followed. For the Bailey et al. (2004) report, about half of the *in vivo* studies were conducted according to GLP guidelines; for the other half, GLP compliance was not explicitly stated. For Gautheron et al. (1994), the *in vivo* studies were conducted according to European Economic Community (EEC) 1984 and 1991 test guidelines (predecessors of the current EU test guideline for eye irritation), but this information alone does not give enough information about GLP compliance. For the remaining reports (Swanson et al. 1995; Gettings et al. 1996; Casterton et al. 1996; Swanson and Harbell 2000), the extent of GLP compliance was not provided, so the extent of GLP compliance is not known.

4.5 Availability of Relevant Human Ocular Toxicity Information

ICCVAM should make an effort to obtain and consider information on human topical ocular chemical injury. It would seem worthwhile to determine if the current ocular hazard classification schemes are working correctly to protect workers and the public from severe eye injury by examining the injury databases maintained by the Poison Control Centers and the Department of Labor. The United States Eye Injury Registry (USEIR) may be another source of such information.

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

There should be more discussion 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 only one *in vivo* 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. Historical data show that between 10% and 15% of the time a single rabbit test will misclassify a compound (Weil and Scala 1971; Kaneko 1996; Ohno et al. 1999). If this is the case, then 10% of the *in vivo* results are misclassified. Unfortunately, there is no way to determine which results are correct and which are not. An effort to determine if the *in vivo* results are consistent with the known toxicity of these materials would be useful (e.g., as indicated in the Registry of Toxic Effects of Chemical Substances [RTECS] or the International Uniform Chemical Information Database [IUCLID] databases).

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, Weil and Scala (1971) identified "good" laboratories as those which 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. The data from all three papers should be discussed in the BRD.

It is 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 nonirritating chemicals and at the upper end of the scale in the range of severely irritating materials). However, in the middle range, the variability is very high (as indicated by the high CV and SD values in Balls et al. 1995, and Ohno et al. 1999).

When interpreting the *in vitro* test data, the differences in reproducibility/variability of the *in vivo* Draize eye test data have to be taken into account. Therefore, it has to be defined before data analysis is performed how this feature of the Draize eye test will be taken into account, when comparing it to results from *in vitro* tests and when attempting to determine the predictive value of the *in vitro* alternatives.

This important aspect has been cited as the main reason why the replacement of the Draize eye test by *in vitro* tests has failed in the past. As this view is well documented in the scientific literature (e.g., Balls et al. 1995), additional discussion in the BRD is warranted.

In summary, although the Panel believes there should be more consideration of the variability of the Draize data, the data are considered useful for evaluation of the BCOP assay.

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 III - 12.3**).

5.0 BCOP TEST METHOD DATA AND RESULTS

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

The Panel agrees with the BRD assessment of these data

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

The Panel is not aware of other data that include the raw scores for both tests.

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

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 (BCOP BRD Sections 6.0 and 7.0). The conclusions relating to test method reliability (BRD Section 7.4) drawn from the analyses in BRD Section 7.0 seem sound.

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

The Panel agrees with the BRD assessment of these data. The lack of GLP compliance should not *a priori* exclude data from evaluation.

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

The Panel agrees with the BRD assessment of these data. However, many of the substances used in the accuracy and reliability calculations are classified in Appendix E of the BCOP BRD not as 'liquid' or 'solid' but instead as 'not provided'. Since one of the issues for the BCOP is the problem with solids, it would be helpful to obtain physicochemical information on as many of these materials as possible. The use of 'volatile solvents' is described in the BRD as problematic with the 10-minute exposure time. The Panel evaluation of the data indicates that alcohols and ketones are the problematic substances, but additional physicochemical data are needed to refine this evaluation.

In summary, the *in vitro* data are sufficient and acceptable, but more data on the physicochemical characteristics of the test substances are needed.

6.0 BCOP TEST METHOD ACCURACY

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

The accuracy of the BCOP test method has been evaluated in comparison to the EPA (1996), EU (2001), and the GHS (UN 2003) ocular irritancy classification systems assuming the formula used to calculate the *in vitro* score currently used is optimal for identifying severe irritants. The discussion is very complete and the data are presented clearly.

Because the Panel does not have data that could give information on the variability in the *in vivo* test results, it is difficult to determine if the single rabbit test being used as the "reference standard" is in fact an "accurate" rabbit test. Combining all *in vitro* results on a substance into a single value minimizes the variability of the data and appears to be the best approach for obtaining an accurate *in vitro* number, realizing the variability has been defined during the inter- and intra-laboratory comparisons. However, without similar information on

the accuracy of the *in vivo* results, statistical comparisons are very one sided. As discussed previously, it can be assumed from past experience that 10% to 15% of the *in vivo* results from a single assay are 'wrong' (Weil and Scala 1971; Kaneko 1996; Ohno et al. 1999). The Panel is aware that NICEATM conducted an analysis of the variability of the *in vivo* test method and believes the final decision on what can be said about accuracy should be made after reviewing the results of the NICEATM study. In addition, the Panel recommends scanning other publicly available sources of eye irritation data (e.g., RTECS or IUCLID databases) to determine if the *in vivo* data used in these studies is comparable to the results now accepted for regulatory purposes.

The Panel has been asked to compare the data to three different regulatory standards. There are two sources of variability when comparing these results. First, the rabbit tests were evaluated in different ways and, secondly, different lists of substances could be evaluated for different regulatory standards. It is not clear if the Panel should suggest the use of the BCOP test method for one regulatory agency scheme but not another.

In addition, the use of single numbers for the various accuracy calculations is misleading. This approach gives the appearance that the *in vivo* tests used for comparison are 100% accurate and there is no possible source of variability around these numbers. The numbers should be clearly presented as concordances with a single Draize test result.

The Panel would like to point out that the scientific justification for the classification schemes for the *in vivo* data is not being examined in this review and this could well be a significant source of both variability in the *in vivo* test and the apparent lack of accuracy in the *in vitro* test as compared to the three regulatory classification schemes. This is particularly true for the two schemes that at least in part base their classification on the result of a single rabbit (i.e., EPA 1996; UN 2003), which would appear to increase the possibility of test-to-test variability as shown by Kaneko (1996), and for which there are no data on the variability of the *in vivo* results.

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 BCOP Test Method

The strengths and limitations identified within the confines of the substances tested are adequately discussed in the BCOP BRD with the exception of the effect of colored substances. Again, this determination is hampered by the lack of similar data obtained using the *in vivo* protocol. The exploration of the effects of physicochemical properties is limited. In the future, consideration should be given to exploring these effects further using a structure activity or structure property relationship program.

6.3 BCOP Test Method Data Interpretation

Issues of test data interpretation have been adequately addressed in the BCOP BRD. In addition to the analyses conducted, the Panel suggests an assessment based on ranking of experimental data for severity for both the reference method and the *in vitro* test.

In summary, the test method is accurate for identification of corrosive and severely irritating substances, except for alcohols, ketones, and solids, when used in the tiered testing scheme described in the BCOP BRD.

7.0 BCOP TEST METHOD RELIABILITY (REPEATABILITY/ REPRODUCIBILITY)

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

The Panel agrees with the BRD assessment of these data.

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

The BCOP BRD concludes, in Section 7.4, that while the intralaboratory repeatability and the intra- and inter-laboratory reproducibility of the BCOP test method appear sufficient for its general application to the detection of ocular corrosives and severe irritants, further work may be needed to reduce interlaboratory variability associated with alcohols, organic solvents and solids. After reviewing the data, the Panel agrees the intra- and inter-laboratory reproducibility of the test appear sufficient and that alcohols and solids need to be reviewed. From the data provided it is difficult to determine if it is organic solvents in general that are a problem. The data provided indicate that ketones also need to be reviewed.

CV values should be used with care with this data because the scores can range from 200 to less than 1. The median and mean CV data may not be informative because it will depend greatly on the scores of the individual tests used in the analysis; that is, comparing the means of the CVs of a set of results with predominantly high scores with a set of results with predominantly low scores is inappropriate.

The data from existing studies have been extensively reviewed and considered in the BCOP BRD. The impression from the summary and conclusions is that the test method showed acceptable levels of intralaboratory repeatability and reproducibility, and interlaboratory reproducibility. Note, though, that in Southee's interlaboratory comparison (Appendix F of the BCOP BRD), there are highly significant differences between the three laboratories in the values they obtained for the *in vitro* scores for ethanol, although variability between and within experiments in the same laboratory was low. The mean score for the three laboratories was 46.3 (SD = 9.7; CV = 21%). This indicates that even with good laboratories, a standard protocol, and a "simple" substance, significant differences in response can occur. It also supports the comment in the summary that further work may be needed to reduce interlaboratory variability.

7.3 Availability of Historical Control Data

The Panel agrees with the BRD assessment of these data.

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

The test method proposed is robust. Several additions to the currently used protocol have been proposed in the BCOP BRD to standardize current practice. Further suggestions have been made by this Panel to reduce variability within and between laboratories. Whether adopting these suggestions will actually reduce variability will need to be determined experimentally.

In addition, many of the suggestions for the protocol seem to come from IIVS. This is a good laboratory with a lot of experience, so their suggestions are important. On the other hand, it would be useful to determine if other laboratories believe the changes that have been suggested are possible within their constraints.

In summary, the inter- and intra-laboratory reproducibility of the method is acceptable.

8.0 TEST METHOD DATA QUALITY

8.1 Impact of GLP Noncompliance and Lack of Coded Chemical Use

The quality of the data used in the BCOP BRD is adequately described. Failure to use coded substances or to follow GLP guidelines significantly impacts on the quality of some data presented in the BRD. Coding was not used for one study but this study was not utilized in the accuracy analysis using pooled data from different studies. Coding should be used for all subsequent studies.

8.2 Results of Data Quality Audits

The Panel agrees with the BRD assessment of these data. Spot checks of data not part of the multilaboratory validation studies could be conducted; however, the Panel does not believe this is necessary.

8.3 Impact of GLP Deviations Detected in the Data Quality Audits

The BRD assessment of these data is appropriate.

8.4 Availability of Original Records for an Independent Audit

The availability of notebooks is described in the BCOP BRD. The lack of original notebook data for this review is of some concern but not sufficient to remove the data from consideration. Information presented at the January 11-12, 2005, meeting indicates that raw data may be available for many, if not all, of the studies included in this evaluation. The ICCVAM recommendation that all 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 1997, 2003).

In summary, the Panel believes the data quality is sufficient.

9.0 OTHER SCIENTIFIC REPORTS AND REVIEWS

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

Relevant data appear to be identified. The BCOP test bears direct biological relevance to the Draize test.

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

The Panel agrees with the BRD assessment of these data.

9.3 Approaches to Expedite the Acquisition of Additional Data

NICEATM has made every attempt to obtain available data. It is possible that more data could be obtained by working through trade associations, but much of the data in the BCOP BRD comes from these sorts of efforts, so whether more data could be obtained is unclear.

In summary, the additional data have been adequately reviewed.

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

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

The BCOP BRD adequately addresses these issues. Use of the BCOP test method will result in the use of fewer animals by classifying some substances without further animal tests and reduce the number of animals exposed to severe irritants.

In summary, the BCOP BRD adequately addresses animal welfare considerations.

11.0 PRACTICAL CONSIDERATIONS

11.1 BCOP Test Method Transferability

11.1.1 Facilities and Major Fixed Equipment Needed to Conduct the BCOP Test Method
The BCOP BRD addresses these considerations adequately.

11.1.2 General Availability of Other Necessary Equipment and Supplies
The BCOP BRD addresses these considerations adequately.

11.2 BCOP Test Method Training

11.2.1 Required Training Needed to Conduct the BCOP Test Method
The BCOP BRD addresses these considerations adequately.

11.2.2 Training Requirements Needed to Demonstrate Proficiency
The BCOP BRD addresses these considerations adequately with the exception that the description of training of technicians for the *in vivo* test may be improper -- the technicians essentially have to demonstrate proficiency in the *in vivo* test the same way as in the *in vitro* test.

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

11.3 Relative Cost of the BCOP Test Method

The BCOP BRD addresses these considerations but the discussion should be modified to reflect the public comments submitted by S.C. Johnson & Son, Inc. in December 2004 on the costs and time comparisons with the Draize test.

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

For very corrosive substances and some severe irritants, the evaluation may be completed within four hours in the *in vivo* test, since animals should be killed for humane reasons if severe lesions are seen.

In summary, the Panel sees no serious practical issues with the use of the BCOP test method.

12.0 PROPOSED TEST METHOD RECOMMENDATIONS

12.1 Recommended Version of the BCOP Test Method

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

For the purpose of identifying corrosive or severe eye irritants in the tiered testing scheme outlined in the BRD, the proposed version of the BCOP test method has been shown to have adequate accuracy and reliability for detecting corrosive or severe eye irritants, with the exception of the caveats described in **Section III - 12.2** of this report.

12.2 Recommended Standardized BCOP Test Method Protocol

For the purpose of detecting severe eye irritants in the tiered testing scheme outlined in the BRD, the proposed BCOP test method protocol is useful for identification of severe or corrosive ocular irritants with the following caveats:

- The test should not be used to identify corrosive or severely irritating ketones, alcohols, and solids. Further optimization and validation are necessary before these classes of materials can be assessed with this test.
- It needs to be confirmed that the BCOP test method can identify, as well as or better than the Draize test, those substances known to cause serious eye injury in humans. It appears from the list of chemicals tested that at least some of these substances have been tested in BCOP (e.g., floor strippers, heavy duty cleaners).
- Users should be aware of zoonoses, including the possibility of BSE.
- A histopathological examination should be added to the test unless the test substance is from a class of materials known to be accurately predicted using only opacity and permeability in the BCOP assay.
- Concurrent negative, positive, and benchmark controls should be used.
- 0.9% NaCl should be used instead of distilled water as the test substance diluent.
- Determination of osmolarity and pH of test solutions should be conducted.
- The optimum age range for cattle should be determined.

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

The following are recommended as modifications that might improve the accuracy and reliability (repeatability/reproducibility) of the BCOP test method:

- Use of the larger holder as suggested by Ubels et al. (2002, 2004)
- Re-examine the use of the calculated total score when the endpoint is serious injury only
- Changes to the medium used to bathe the eyes including a determination of whether FBS is needed

While these modifications are important, the data presented in the BRD support use of the BCOP assay in its current form for identifying ocular corrosives and severe irritants other than alcohols, ketones, and solids in a tiered testing strategy for regulatory hazard classification and labeling purposes.

12.2.2 Other Endpoints that Should be Incorporated into the BCOP Test Method

Histopathological examination should be added to the recommended test protocol unless the test substance is from a class of materials known to be accurately predicted using only opacity and permeability in the BCOP assay.

While actually a change to the BCOP method, the Panel calls attention to the possibility that porcine eyes might also be a useful model for human eyes. This change would require complete validation, but the Panel wants to be sure this possibility is considered for future work.

Minority Opinion

Dr. Freeman expressed no opinion as to whether the BCOP assay had met the validation criteria as set forth in Appendix D of the ICCVAM Submission Guidelines (2003). This is because the question of whether these validation criteria had been met never reached a conclusive decision by the Panel. This is the basis for his abstention from voting on the acceptance of **Section III - 12.2**.

The Panel raised the question as to whether the BCOP assay could be considered validated. This was determined to not be a function of the Panel; however, it was also determined that it was a function of the Panel to judge whether the validation criteria (as set forth in the ICCVAM guidelines cited above) had been met. Although the Panel report on the BRD addressed the validation criteria, during the discussion, it seemed that some Panel members were unclear as to whether they had been asked to specifically answer this question in a summary manner. Thus, no summary conclusion was reached on whether the validation criteria were fulfilled, and under time constraints to end the Panel review on schedule, the adopted language was that the assay "was useful" in the identification of severe irritants or corrosives to the eye.

The discussion regarding BCOP could have been resolved more definitively with a few minor changes to the process, as noted below:

- The Panel should have been clearly instructed and reminded as necessary that it was to conclude whether the available information on the assay fulfilled the validation criteria.
- When it became clear that there was confusion on the ultimate objective, the tasking should have been clarified and possibly a recess called to permit appropriate deliberation. Please keep in mind the extensive preparatory work (and cost) prior to the Panel meeting.

It is suggested that a pro forma checklist be developed as an aid to guide future Expert Panels to final resolution of their assigned tasks, e.g., determining the validation status, that is, whether validation criteria, have been met.

Minority Opinion

Drs. Theran and Stephens state that the chair of the BCOP group summarized the group's findings and conclusions on the afternoon of January 12th, during the plenary, public session of the full expert panel. The group's key conclusion was that the BCOP had satisfied ICCVAM's validation criteria, and therefore the validation status of the BCOP test method should be characterized as "valid" for the purpose of serving as a positive screen for severe or corrosive eye irritants. The BCOP group chair noted that as with all methods previously shown to be valid by ICCVAM, ECVAM, and others, the BCOP test method has particular strengths and limitations that should be taken into account when the method is used.

Drs. Theran and Stephens object to the pressure brought to bear on the BCOP group that ultimately led the members, under duress, to withdraw their summary conclusion that the test method was valid and to substitute the tepid and vague language from other group reports that the test method was "useful." They believe that ICCVAM personnel and panel members were incorrect in stating that the charge to the four groups did not include drawing conclusions about the validation status of the test methods under review. The very title of the 18-page charge to the panel was "Guidance to the Expert Panel for Evaluation of the Validation Status of the BCOP, ICE, IRE, and HET-CAM Test Methods for Identifying Ocular Corrosives and Severe Irritants" (emphasis added). After much heated discussion, the BCOP group was given the opportunity to make a statement on the validation status of the BCOP method, but the group had been subjected to such counter pressure by that point that they understandably decided against characterizing the method as valid.

An official effort to clarify the charge to the group on the final morning of our 4-day effort was helpful, but once again lead to heated discussion that muddied the waters.

This minority opinion was filed because Drs. Theran and Stephens believe the BCOP group was inappropriately pressured to withdraw its main scientific finding. The final report should have concluded that the BCOP has been found to be valid, within the identified limits, and that any further optimization or other studies should not be cause for delaying regulatory agency review for test method acceptance.

12.3 Recommended Optimization and Validation Studies

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

Future improvements to improve the accuracy and reliability (repeatability/reproducibility) are recommended including use of the larger holder similar to that suggested by Ubels et al. (2002), re-examining the use of the calculated total score when the endpoint is serious injury only, changes to the medium used to bathe the eyes, avoiding use of antibiotics, and appropriate ages of donor animals. While these improvements are important, the data presented in the BRD are sufficient for supporting use of the BCOP assay in identifying ocular corrosives and severe irritants, except for alcohols, ketones and solids, in a tiered testing strategy for regulatory hazard classification and labeling purposes.

The optimization study design recommended in the BCOP BRD is appropriate.

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

Validation studies, or submission of additional data supporting the three-minute exposure time suggested for volatile solvents, will be necessary before the BCOP test method can be recommended for use with alcohols and ketones. Validation studies or submission of additional data will be necessary before the BCOP test method is acceptable for solids.

The information in the BCOP BRD, along with the additions of our suggestions, is sufficient to support the use of this test method to identify severe irritants and corrosives, with the exception of alcohols, ketones, and solids, in the tiered testing scheme described in the BRD.

It is understood that adding histopathological examination to the test method involves additional endpoints, but current practice has not been to insist on validation of histopathological examination when it is added to an *in vivo* test method. Thus, there is no need for an additional validation study based solely on the addition of this endpoint. A standardized histopathological scoring system is suggested, but this should be arrived at by the experts in the field and will not require validation. NICEATM/ICCVAM should facilitate the development of a histopathological scoring system for corneal damage (with visual aids).

Changes in the calculation method for the BCOP test score, or the use of the individual endpoint data instead of a calculated score also do not need to be validated.

When validation studies are conducted, the studies proposed in the BCOP BRD are appropriate but should be limited to the classes of test substances in question. Validation studies should be carefully planned. Tests should first be done to confirm that any modifications of the protocol do not decrease reliability. Once the inter- and intra-laboratory variability is defined, it will not be necessary to have a large number of laboratories test every chemical in the validation study. Validation should focus on the class of chemicals in question. The study should involve a very small number of experienced laboratories with only a limited number of duplicate samples at each laboratory.

Any validation or optimization 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 information obtained (e.g., wound healing) and to minimize the number of animals used.

Minority Opinion
According to Dr. Martin Stephens, **Section III – 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 BCOP BRD REFERENCES

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

The papers of J.V. Jester and J.K. Maurer should be added as they support the use of short-term endpoints to predict longer-term results.

Also add to the BCOP BRD any other publications cited in **Section III** of this report and listed below that were not included in the BRD.

14.0 PANEL REPORT REFERENCES

Bailey PT, Freeman JJ, Phillips RD, Merrill JC. 2004. Validation of the BCOP assay as a predictor of ocular irritation of various petrochemical products [Abstract]. Toxicologist 78(S-1):266.

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, Evans MG, McPherson JP, Southee JA, Williamson PS. 1998. Investigation of ingredient interactions in cosmetic formulations using isolated bovine corneas. Toxicol In Vitro 12:669-690.

Casterton PL, Potts LF, Klein BD. 1996. A novel approach to assessing eye irritation potential using the bovine corneal opacity and permeability assay. J Toxicol Cutaneous Ocul Toxicol 15:147-163.

Cater K, Patrick E, Harbell J, Merrill J, Schilcher S. 2004. Comparison of *in vitro* eye irritation potential by BCOP assay to erythema scores in human eye sting test of surfactant-based formulations [Abstract]. Toxicologist 78(S-1):268.

EEC. 1984. Acute toxicity – eye irritation. In Directive 67/548 (6th adaption); Annex V, Part B: Methods for the Determination of Toxicity. Official Journal of the European Community 27 L251:109.

EEC. 1991. Classification of irritant substances and preparations. Official Journal of the European Community. L180:52.

EPA. 1996. Label Review Manual. 2nd 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.

Gautheron P, Giroux J, Cottin M, Audegond L, Morilla A, Mayordomo-Blanco L, Tortajada A, Haynes G, Vericat JA, Pirovano R, Tos EG, Hagemann C, Vanparys P, Deknudt G, Jacobs G, Prinsen M, Kalweit S, Spielmann H. 1994. Interlaboratory assessment of the bovine corneal opacity and permeability (BCOP) assay. Toxicol In Vitro 8:381-392.

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.

Hughes Jr. WF. 1946. Alkali burns of the eye. II. Clinical and pathologic course. Arch Ophthalmol 36:189-214.

ICCVAM. 1997. Validation and regulatory acceptance of toxicological test methods: A Report of the ad hoc Interagency Coordinating Committee on the Validation of Alternative Methods. NIH Publication No.: 97-3981. Research Triangle Park, NC:National Institute of Environmental Health Sciences.

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.

Maurer JK, Parker RD, Jester JV. 2002. Extent of initial corneal injury as the mechanistic basis for ocular irritation: key findings and recommendations for the development of alternative assays. Regul Toxicol Pharmacol 36:106-117.

McCulley JP. 1987. Chemical Injuries. In: The Cornea: Scientific foundation and clinical practice. (Smolin G, Thoft RA, eds). Boston: Little, Brown and Co, 527-542.

Nussenblatt RB, Bron R, Chambers W, McCulley JP, Pericoi M, Ubels JL, Edelhauser HF. 1998. Ophthalmologic perspectives on eye irritation testing. J Toxicol Cutaneous Ocul Toxicol 17:103-109.

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.

Schermer A, Galvin S, Sun T-T. 1986. Differentiation-related expression of a major 64K corneal keratin in vivo and in culture suggests limbal location of corneal epithelial stem cells. J Cell Biol 103:49-62.

Southee JA. 1998. Evaluation of the Prevalidation Process. Part 2, final report. Volume 2. The Bovine Corneal Opacity and Permeability (BCOP) Assay. European Community contract no. 11279-95-10F 1ED ISP GB.

Swanson JE, Harbell JW. 2000. Evaluating the eye irritancy potential of ethanolic test materials with the bovine corneal opacity and permeability assay. The Toxicologist 54:188-189.

Swanson JE, Lake LK, Donnelly TA, Harbell JW, Huggins J. 1995. Prediction of ocular irritancy of full-strength cleaners and strippers by tissue equivalent and bovine corneal assays. J Toxicol Cutaneous Ocul Toxicol 14:179-195.

Tseng SCG, Sun T-T. 1999. Stem cells: ocular surface maintenance. In:Corneal Surgery: Theory, Technique, and Tissue. (Brightbill FS, ed]. St. Louis: Mosby, 9-18.

Ubels JL, Paauw JD, Casterton PL, Kool DJ. 2002. A redesigned corneal holder for the bovine cornea opacity and permeability assay that maintains normal corneal morphology. Toxicol In Vitro 16:621-628.

Ubels JL, Ditlev JA, Clusing DP, Casterton PL. 2004. Corneal permeability in a redesigned corneal holder for the bovine cornea opacity and permeability assay. Toxicol In Vitro 18:853-857.

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.

Hen's Egg Test – Chorioallantoic Membrane Test Method

[This Page Intentionally Left Blank]

IV. HEN'S EGG TEST-CHORIOALLANTOIC MEMBRANE TEST METHOD

1.0 HET-CAM TEST METHOD RATIONALE

1.1 Scientific Basis for the HET-CAM Test Method

1.1.1 Mechanistic Basis of the HET-CAM Test Method

The rabbit eye is the current reference standard in predicting what will happen when the human eye is directly exposed to a chemical, even though the rabbit eye is somewhat structurally different from the human eye. It should always be noted, however, that suitable human data would be vastly preferred as a comparative standard. The chorioallantoic membrane (CAM) contains vascular membrane structures. The Hen's Egg Test – Chorioallantoic Membrane (HET-CAM) test system is used as a model of the cornea, conjunctiva, and iris to detect ocular corrosives and severe irritants. However, the CAM tissue structure is not similar to the cornea as the latter is not vascularized epithelium. Exposure of the rabbit eye to a chemical results in a pathophysiological reaction whereas the HET-CAM assay detects vascular injury. The differences in the structure of the CAM and the mammalian eye must be considered when using the HET-CAM assay as a predictor of potential for human eye irritation.

It is recognized that HET-CAM is an *in ovo* assay but for purposes of consistency, the term *in vitro* will be used when referring to this test method.

It is recommended that the draft HET-CAM BRD include discussions on:

- cellular mechanisms of corrosion and severe irritation (e.g., necrosis, apoptosis) and relevance to *in vitro* testing, and
- the role of responsive inflammatory cells in isolated rabbit eyes and how this compares to the responsive inflammatory cells in the CAM.

Furthermore, additional literature and laboratory research to review the following questions are recommended:

- How much and what kind of data are available for using eggs at incubation day 7?
- What is known about the development of the chorioallantoic membrane, its sensitivity and its reactivity on incubation day 7 compared to incubation day 9?
- What kinds of data about pain receptors are present on the CAM on either incubation day 7 or day 9?
- How does the incubation day affect the reliability and variability of the data?

1.1.2 Advantages and Limitations of Mechanisms/Modes of Action of the HET-CAM Test Method

The HET-CAM test method appears to be suitable as a limited screen for a broad array of different types of chemicals. A deficiency of the CAM is that it has no structures comparable to the iris and cornea. Chemical exposure in the rabbit eye can be relatively long (usually never washed) as compared to the HET-CAM assay, which is relatively short (5 minutes).

The actual endpoints assessed in the two test systems are different. The rabbit eye test assesses each specific major eye structure endpoints up to 21 days post exposure while the HET-CAM test method uses a scoring system and formula to evaluate the degree of blood vessel hemorrhage, lysis, and coagulation.

1.1.3 Similarities and Differences of Mechanisms/Modes of Action and Target Tissues Between the HET-CAM Test Method and Humans and Rabbits

Much is known about differences in mechanisms/mode of action between the HET-CAM test method and humans and rabbits. All of these differences have to be considered and kept in mind as comparisons are made. Exposure of the rabbit conjunctiva to a chemical results in an immunological reaction whereas in the HET-CAM assay, the result is a measure of vessel necrosis. The differences in response of adult tissues (with a developed immune system) verses embryonic tissues (with a much undeveloped immune system) also need to be kept in mind when reviewing the results from the HET-CAM test method. Due to these differences, it cannot be assumed that adverse changes that occur in the HET-CAM test method are going to be similar to what may occur in the rabbit or human eye.

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

Due to the differences in the mechanisms of the response between the tests, the *in vivo* rabbit eye test will more closely predict what changes will occur in the human eye over a period of days. The *in vivo* rabbit test follows the eye over a period of up to 21 days and any long-term effects can be noted in endpoints very relevant to human exposure (iris, cornea, conjunctiva). Comparatively, the HET-CAM test method is a short-term test (5 minutes) with few endpoints (hemorrhage, lysis, coagulation) and no responses related to the cornea or iris.

Any relationship between the short-term effects observed in the HET-CAM test method to the long-term effects seen in rabbits or humans should be explored in the HET-CAM BRD. Such an evaluation may provide additional support for the use of the HET-CAM method to assess the delayed and long-term effects of corrosives and severe irritants.

1.2 Regulatory Rationale and Applicability

1.2.1 Similarities and Differences Between Endpoints Measured in the HET-CAM Test Method and the *In Vivo* Rabbit Eye Test Method

The endpoints are very different between the *in vivo* rabbit eye and the HET-CAM test methods. The *in vivo* rabbit eye endpoints are very similar, if not identical, to what may happen to a human eye after exposure to a substance. The HET-CAM endpoints are a representation of what may happen by inferring from the onset of blood vessel necrosis in the CAM.

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

The BRD has summed up these issues as five criteria that must be achieved. Four of the five criteria seem to be achievable. One criterion, which may be difficult to achieve, is criterion number 4: "Provide improved prediction of adverse health effects in the human". This criterion would be difficult to achieve unless comparative data are generated using substances from a standardized repository that are already known to cause specific effects in humans. The HET-CAM assay and other identified assays must all be tested using the same standard substances to determine if the assay can improve the prediction of adverse eye effects for humans.

It is hard to visualize that the HET-CAM test method, in its current state of performance, would do more than add another level of testing which would rarely supplant the existing tests. Rather, the HET-CAM test method may have the potential to complement other tests in a tiered-testing approach.

2.0 HET-CAM TEST METHOD PROTOCOL COMPONENTS

2.1 Description and Rationale of the Components for the Recommended HET-CAM Test Method Protocol

The recommendations from the draft HET-CAM BRD appear to appropriately integrate protocol components and specific procedures from the various published literature. These BRD recommendations also include developing consistent scoring and calculation of irritation indices.

Reference substances that are part of the performance standards developed for the HET-CAM test method should be identified in the BRD. These reference substances would be used to evaluate test methods similar to HET-CAM. The HET-CAM BRD also should clarify the decision criteria for identifying ocular corrosives and severe irritants.

2.2 Basis for Selection of the HET-CAM Test Method System

Historically, the chick embryo has been extensively utilized. The specific strain, stock and age of White Leghorn eggs, which has been recommended in the BRD, is common and fairly easy to obtain; use of these eggs would provide consistency for the HET-CAM assay results.

2.3 Identification of Proprietary Components

The Panel agrees with the BRD, there are no proprietary components of the test system.

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

The BRD recommendations on the numbers of replicates and/or repeat experiments would provide uniformity and consistency to the HET-CAM assay in interpreting the results. Many alternative assays that are submitted to regulatory agencies have, as part of the protocol, a

standardized number of replicates that must be used in order for the test system to be considered valid.

2.5 Study Acceptance Criteria for the HET-CAM Test Method

Since the study acceptance criteria varied between the various test method protocols, a definition of what constitutes a positive result is needed. Also, since there are times when the concurrent control can show quite a bit of variation, tabulation and use of historical control data need to be considered. More objective criteria for assessment would enhance the repeatability and reliability of the HET-CAM test method. Objective criteria also would enhance the validity of interlaboratory comparisons.

2.6 Basis for Any Modifications made to the Original HET-CAM Test Method Protocol

The Panel agrees with the BRD recommendations on the bases for any modifications made to the original HET-CAM test method protocol.

2.7 Adequacy of the Recommended Standardized Protocol Components for the HET-CAM Test Method

The Panel agrees with the BRD recommendations for the development and use of a standardized HET-CAM test method protocol. A critical recommendation is the inclusion of BOTH concurrent negative and positive controls each time the assay is conducted. In addition, investigators need to accumulate historical data for their positive and negative controls in order to better define the range of positive and negative responses as different materials are tested in the HET-CAM assay.

3.0 SUBSTANCES USED FOR PREVIOUS VALIDATION STUDIES OF THE HET-CAM TEST METHOD

3.1 Substances/Products Used for Prior Validation Studies of the HET-CAM Test Method

The types and numbers of substances/products used in prior validation studies appear adequate.

3.2 Coding Procedures Used in the Validation Studies

It was difficult to determine if the coding procedures used in the validation studies were appropriate. There was not enough information to determine the appropriateness of the coding used. As long as the quality and multiplicity of sources of the data were sufficient to draw meaningful conclusions, it does not matter if coding was not used.

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

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 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* 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

If there are a few test substances that lack original study records, then they should not be given the same weight as those test substances with original study records. However, if there are many test substances that lack original study records and it appears that obtaining the original study records may be difficult, then such studies should be given equal weight with those that have original study records. In the case of the HET-CAM test method, original study data (e.g., laboratory notebooks) 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 is adequate (as is the case for the studies evaluated in the HET-CAM BRD), the study should be used.

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

The criteria used in selecting agents in some of the studies for the HET-CAM test method cited in the BRD were not specified. The Balls et al. (1995) project included the criterion that the *in vivo* data were from GLP-compliant post-1981 studies conducted in accordance with OECD TG 405 (OECD 1987). The Spielmann et al. (1996) project was conducted under blind conditions according to GLP standards in laboratories of the chemical and drug industry in Germany. The Panel recommends that the status or availability of additional information on GLP compliance needs to be pursued more diligently.

However, as the GLP regulations do not deal with the actual performance of the tests as much as with documentation, no distinction needs to be made in the weight given to GLP-compliant versus non-GLP-compliant studies in the BRD as long as the work was performed in well-established laboratories (e.g., stable workforce, significant throughput in that section of the laboratory, long term experience with the test method, historical data, adequate supervisory staff). It is recognized that these are some characteristics of a well-established

laboratory and are not meant to be criteria for determining such laboratories. Furthermore, 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 criteria 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 have been accessed 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, it would seem worthwhile to determine if the current ocular hazard classification schemes are working correctly to protect workers and the public from severe eye injury. While it is difficult to obtain specific data from the various databases, they can be useful to give reassurance that current schemes appear to be protecting the public. According to the European Cosmetics, Toiletries and Perfumery Association (COLIPA) Task Force on Eye Irritation workshop report (Bruner et al. 1998), the International Life Sciences Institute (ILSI) has published a human eye irritation classification scheme (see Table II in Bruner et al. 1998) and planned to search databases on human eye irritation. Therefore, it is recommended that COLIPA and ILSI be consulted for human data.

The Panel also recommends that a greater effort be made to obtain, consider, and use information on human topical ocular chemical injury. The USEIR may be one source of such information. Literature sources of human topical ocular chemical injury include, but are not limited to, Grant (1974), Fox and Boyes (2001), and Fraunfelder (1982).

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

There should be more discussion in the HET-CAM 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 test (e.g., Weil and Scala 1971; Spielmann 1996), 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, the differences in reproducibility/variability of the *in vivo* Draize eye test data have to be taken into account. Therefore, before data analysis is performed, it has to be defined how this special feature of the Draize eye test will be taken into account when comparing it to results from *in vitro* tests and when attempting to determine the predictive value of the *in vitro* alternatives.

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 it is well documented in the scientific literature (e.g., Figure 1 in Balls et al. [1995]) and in a review by Spielmann (1997), additional discussion in the HET-CAM BRD is warranted.

The Draize eye irritation test has never gone through a validation process. However, data on the 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). Using a fixed protocol and a single supply of chemical agents tested in 25 laboratories, these investigators identified "good" laboratories as those, which 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.

Using data from the Weil and Scala (1971) study, another evaluation showed the difference in MAS values that can be obtained between different laboratories. For three of the ten substances tested, the *in vivo* Draize eye irritation test indicated that the substances were classified as nonirritant (MAS < 20) to irritant (MAS > 60) when tested in 24 laboratories (Spielmann 1996).

It is documented that the Draize eye test has 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) (Spielmann 1996). However, in the middle range, the variability is very high for such substances (as indicated by the high CV and SD values in Balls et al. [1995]). While any repeat performance of *in vivo* rabbit eye irritancy testings or testing of known corrosives or severe irritants should be strongly discouraged, it is important to have available multiple *in vivo* rabbit eye test data that demonstrate reproducible results.

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* method. These questions typically center on two major concepts. The first is the availability of a reference standard for measuring the intended effect. The second is the reproducibility and reliability 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.

Not all substances evaluated in the HET-CAM BRD were tested concurrently in both the *in vivo* rabbit eye and the HET-CAM test methods. 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".

An effort should be made to determine if the *in vivo* results are consistent with the known toxicity of these materials (e.g., as indicated in the RTECS or IUCLID databases) would be useful. It is imperative that a greater effort be made to access suitable human data from other sources such as Hazardous Substances Data Bank (HSDB), the Physician's Desk Reference (PDR) and the Poison Control Center network.

The Panel recommends that any future optimization and validation studies should use existing animal data if they are available. If important data gaps are identified, additional animal studies should only be conducted with the minimum number of animals. Such studies should be carefully designed to maximize the amount of pathophysiological information obtained and conducted under GLP conditions.

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 IV - 12.3**).

5.0 HET-CAM TEST METHOD DATA AND RESULTS

5.1 HET-CAM Test Method Protocols Used to Generate Data Considered in the BRD

The test method protocols used to generate each set of data considered in the BRD were adequately described. It is recommended that the type of irritation score (IS) (A or B) analysis method used by each study be detailed in Section 5.4 of the HET-CAM BRD.

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

For the validation of the BCOP test method (Gautheron et al. 1994), an *in vivo* study was performed by one laboratory. Draize data from this *in vivo* study may be a source of data that could be used in the BRD evaluation for available HET-CAM data.

5.3 Statistical and Nonstatistical Approaches Used to Evaluate HET-CAM Data in the BRD

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

The accuracy analysis was complicated by a lack of consistent test and evaluation methods in the literature. Analysis methods in the HET-CAM BRD include the IS(A), IS(B), Q-Score, S-Score, and IS and Irritation Threshold Concentration scores, or in some cases, just classifications based on any of these analysis methods. Results were reformulated in the BRD to be consistent with regulatory agency classifications. The procedure was as good as possible given the lack of consistency among studies. This certainly is not optimal and more internally consistent data are needed.

The classification criteria using these analysis methods should be optimized, including considering the formula for combining information and the irritancy categorization of that result.

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

Whether coded chemicals were tested, or the identity of the chemicals is unknown is adequately documented (HET-CAM BRD Section 3.4). Whether GLP guidelines were followed is detailed in Appendix B of the BRD. How well the studies followed GLP guidelines cannot be determined from the studies. In most of the studies, quality assurance was likely not involved. If studies were conducted following GLP principles, which is likely the case for most of the studies, they should be accepted. GLP-criteria should not overrule all the other criteria for final acceptance of studies for retrospective validation of the HET-CAM test.

Ideally minimal criteria or requirements, such as (1) a well described materials and methods section and (2) criteria for a corrosive or severe irritant call, should be provided and be used to determine an adequate study. However, it is recognized that not all studies would provide such information. Consequently, as long as the data from the study can be interpreted and does not have any serious deficiencies, such as inadequate number of animals, it should be acceptable.

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

There is not enough information on "lot-to-lot" consistency. It is expected that different batches of substances may give some quantitative differences in irritation classification results but a major qualitative difference in irritation classification would not be expected (i.e., classification of a highly severe substance should remain severe between batches of substances). When the irritancy classification of a substance is on the borderline between nonsevere irritant and severe irritant, "lot-to-lot" variations may have an effect on the results. In other words, one batch of a borderline substance may produce a severe irritant response while another batch may produce a nonsevere irritant response.

6.0 HET-CAM TEST METHOD ACCURACY

6.1 Accuracy Evaluation of the HET-CAM Test Method for Identifying Ocular Corrosives and Severe Irritants

The accuracy of the *in vitro* test using the different evaluation criteria has been adequately evaluated. Accuracy evaluations were limited to the substances evaluated in nine *in vitro-in vivo* comparative studies.

1. Accuracy was assessed separately for each *in vitro-in vivo* comparative study.
2. Accuracy was assessed after pooling data across comparative studies that used the same method of data collection and analysis.

Overall, false positive rates ranged from 20% (8/40) to 27% (12/45) and false negative rates from 0% (0/12) to 7% (1/14) compared with *in vivo* rabbit eye test method data classified according to the GHS (UN 2003), the EPA (1996), or the EU (2001) ocular irritancy classification systems. To what degree false results can be reduced with more replicates, more understanding of the various sources of variability, and further optimization of the categorization decision rule is unclear. It will be essential to identify which structural classes of chemicals this test system works for and which ones it performs poorly for.

Tables 6-1 to 6-3 and Table 6-7 of the HET-CAM BRD are quite helpful in summarizing results on all the required accuracy measurements and give a good overview of the performance of the HET-CAM test method. HET-CAM BRD Table 6-9 provides clear information on discordant results, which also are well described in the text.

In addition to the analyses conducted in the BRD, the Panel recommends an assessment based on ranking of experimental data for severity for both the reference method and the *in vitro* test be conducted.

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 HET-CAM Test Method

Concordance assessments are severely limited by the lack of reported data and the differences between methods and analysis methods used in the different studies. False positives and false negatives are identified where possible. Categorization methods used by the authors in the original studies were not designed to meet regulatory agencies requirements. These limitations are clearly spelled out.

It is known that there is much variability among Draize data (Weil and Scala 1971; Spielmann 1996). In the case where an *in vitro* classification is different from the *in vivo* classification, the variability of the *in vivo* response should be reviewed.

6.3 HET-CAM Test Data Interpretation

Because of the limited nature of the reported data, considerable effort was necessary to interpret the data. Data interpretation and specific endpoints applied are sufficiently detailed, to the level possible. The description makes the reader quickly aware that the IS(B) analysis method is the best one to identify most ocular corrosives and severe irritants. A standardized test method is needed to produce more interpretable and consistent data.

It is recommended that IS(B) data of non-accepted studies (HET-CAM BRD Section 9.0) be compiled into a table to see what the outcomes are in these studies.

7.0 HET-CAM TEST METHOD RELIABILITY (REPEATABILITY/ REPRODUCIBILITY)

7.1 Selection Rationale for the Substances Used in the HET-CAM Test Method Reliability Assessment

The rationale for compound selection is based primarily on the easy availability of *in vivo* rabbit eye data. The quality of such data is a weakness for all *in vitro* validation studies. A rationale based on the quality of *in vivo* data (after a thorough investigation and independent checks) would have been better. Selection of substances of which *in vivo* irritancy grades are confirmed by at least two studies would have given more power to the validation of HET-CAM and other test methods. The Panel notes that the above limitations are limitations of the studies used in the analysis and thus limitations of the analysis in the HET-CAM BRD.

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

Analysis on intralaboratory repeatability and intralaboratory reproducibility could not be done due to lack of available data at the time of BRD development. This is a weakness in the validation of the HET-CAM, but should not be a roadblock for its use in identifying ocular corrosives and severe irritants.

Qualitative and quantitative analysis on the interlaboratory variability was well addressed in the HET-CAM BRD. Interlaboratory data were available from four to five laboratories. Ocular irritancy classifications from HET-CAM studies are compared to *in vivo* rabbit eye classifications for each agency classification system. Comparisons are given in HET-CAM BRD Tables 7-1 to 7-3. The participating laboratories agreed on at least half the calls and total agreement occurred frequently. This analysis shows that less agreement among laboratories is obtained with nonsevere irritants/nonirritants. The interlaboratory correlations given in BRD Table 7-7 (for Balls et al. [1995]) vary considerably; S-Scores for chemicals insoluble in water range from -0.9 to 0.852. Clearly, additional work is needed to improve interlaboratory consistency, when using the S-Score analysis method.

Use of %CV values has limitations when evaluating a narrow range of scores (i.e., 0-21 for the HET-CAM test method). Alternative approaches for measuring reproducibility (intra- and inter-laboratory) could be used and are recommended. One approach to assess variability could be the use of a non-parametric analysis, which is useful for small sample sizes and when the data may well not be normally distributed. The Kruskal-Wallis and Mann-Whitney tests evaluate for differences between groups, K groups (where $K > 2$ groups) and 2 groups, respectively. These tests are appropriate for comparing data with continuous outcomes, such as the IS score, to answer the question "do scores differ between laboratories" when comparing replicate scores from the same substance. An assumption for both tests is that observations are independent and identically distributed, and this would not be the case for different substances. So these tests would be useful for a substance-by-substance evaluation if the raw data are, or can be made, available.

A chi-square test for homogeneity of substances across laboratories may be used. But there are so many test substances that this test will not perform well. One could test whether the proportions of substances called severe significantly differ between the laboratories. For HET-CAM, there are enough substances to assume normality of the proportions, so one could do a global test for differences and then use one of a variety of methods for assessing multiple comparisons if the global test for no difference is rejected. This would be a straightforward measure of laboratory differences.

The Spearman rank correlation also is a good non-parametric measure of correlation. It would apply to the IS scores, but not to severe versus not severe outcomes.

The following items are noted for revision in the HET-CAM BRD:

- In BRD Tables 7-4 and 7-5, it would be helpful to have the sample size noted in the table to verify understanding of the text (this is true for some other tables as well). There is nothing in the Table heading or footnotes that say measurements are taken across laboratories.
- Motivation for inclusion of Balls et al (1995) was given. This should also be done for Hagino et al. (1999) on BRD page 7-2 (line 36).
- BRD Page 7-16, line 339: reference is made to Ohno et al. (1999) but no information on this publication can be found in Appendix B.

7.3 Availability of Historical Control Data

The absence of historical negative and positive control data is a weakness in the validation of the HET-CAM test method but this should not be a roadblock for the acceptance of this model as alternative test to detect ocular corrosives and severe irritants.

The Panel notes that some non-accepted publications (HET-CAM BRD Section 9) included positive controls. These publications may give some more information on the reproducibility of HET-CAM. Gilleron et al. (1996, 1997) included a positive control in all HET-CAM studies. Historical control data (90 tests with 0.9% NaCl as negative control, 80 tests with *N,N*-dimethylformamide as a negative control, and 15 tests with imidazole as a positive control) were obtained from Johnson & Johnson Pharmaceutical Research and Development laboratories (Beerse, Belgium) to assess intralaboratory reproducibility. The fact that a test substance applicator was used (which is different from all the other studies discussed in the BRD) should not influence the outcome of the study.

It also is noted that some studies used positive controls that are typically considered nonirritants. Appropriate recommendations are made for the use of concurrent positive and negative controls in the HET-CAM BRD.

7.4 Effect of Minor Protocol Changes on Transferability of the HET-CAM Test Method

The sensitivity of the method to minor protocol changes is impossible to evaluate without having more standardized studies with measures of variability.

Optimization and validation studies are needed for routine regulatory use for hazard classification. Accuracy and reliability may be improved by tailoring the *in vitro* classification scheme to the classification systems of the regulatory agencies and further optimizing the criteria for these systems.

8.0 TEST METHOD DATA QUALITY

8.1 Impact of GLP Noncompliance and Lack of Coded Chemical Use

As scoring of the effects is still somewhat subjective, knowledge of the substances might have influenced scoring of the endpoints during the conduct of the *in vitro* test. Failure to use GLP guidelines may have had a qualitative impact on borderline classifications of nonsevere/severe irritants. The use of GLP guidelines assures that there was good control of the test system, acceptance criteria were defined, evaluation criteria were defined, and there were data audits. Lack of GLP compliance may be overcome by use of coded substances and appropriate data handling.

The Panel recommends that information on coding provided in Section 3.4 of the HET-CAM BRD also be included in Appendix B2.

8.2 Results of Data Quality Audits

The Panel agrees that no data quality checks could be done. This is a weakness not only for the HET-CAM validation but probably also for all other tests as a data quality check is included in the GLP guidelines where an independent group (Quality Assurance Unit; QAU) performs this task. Involvement of QAU is rarely included in validation studies.

8.3 Impact of GLP Deviations in the Data Quality Audits

As this cannot be deduced from the available information, the Panel agrees with the BRD conclusion that the impact of the deviations from GLP guidelines cannot be evaluated.

8.4 Availability of Original Records for an Independent Audit

The Panel agrees that the availability of laboratory notebooks or other records is adequately discussed in the BRD. Evaluation presented in the BRD has been done with the available data and information. 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). Availability of notebooks or other records would increase the "trust index" of the conclusions presented in the HET-CAM BRD.

9.0 OTHER SCIENTIFIC REPORTS AND REVIEWS

9.1 Other Published or Unpublished Studies Conducted Using the HET-CAM Test Method

The Panel agrees that a comprehensive review is made on available publications. The Panel wonders if the criteria for acceptance of literature for evaluation were too strict and relaxing the criteria would have allowed more studies to be included in the final evaluation discussed in the BRD. Additionally, by requesting some additional information on publications closely satisfying the inclusion criteria might have resulted in more studies considered for final evaluation of the performance of the HET-CAM test.

It is recommended that an evaluation on the impact of relaxing the data inclusion criteria be conducted, and additional resources should be placed on contacting authors of relevant papers and individuals that may have *in vitro* and/or *in vivo* data that may be used in the evaluation of the performance of HET-CAM. Additionally, it is recommended that information be placed into the HET-CAM BRD that indicates from which publications additional information was obtained and from which publications additional information was not obtained.

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

The conclusions published in independent peer-reviewed reports and other independent scientific reviews were adequate and complete. It was useful to have the motivation for exclusion of the studies for the final evaluation on the performance of the HET-CAM test. But, once again, the criteria may have been too strict for inclusion of some studies.

Recommendations made by the Panel in **Section IV - 9.1** of this report are applicable to this section.

9.3 Approaches to Expedite the Acquisition of Additional Data

An approach to expedite the process for obtaining additional in-house data could be to make a review on *in vivo* data of a preferred list of compounds and ask companies if they can deliver additional data supporting or contradicting the conclusions made by the Panel.

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

10.1 Extent to Which the HET-CAM Test Method Refines, Reduces, or Replaces Animal Use

This section of the HET-CAM BRD addresses many of the considerations relevant to the 3Rs of refinement, reduction, and replacement. However, the discussion of some issues seems incomplete. In addition, other animal welfare considerations (perhaps not explicitly related to the 3Rs) still need to be discussed, or at least mentioned.

- It is recognized that HET-CAM is an *in ovo* assay but for purposes of consistency the term *in vitro* will be used when referring to this test method.
- Section 10.0 of the HET-CAM BRD mentions that pain perception is unlikely to occur prior to incubation day 9. It is recommended that discussion on pain perception (as is discussed in Section 2 of the BRD) in this section should be expanded.
- It is recommended that Section 10.0 of the HET-CAM BRD also mention the tiered-testing strategy that is being envisioned, namely, the use of HET-CAM test as a first tier test and *in vivo* testing as the second tier, triggered only by a negative finding in the first tier. Thus animals would be needed only to confirm the absence of a severe or corrosive response in the initial tier.
- Given HET-CAM's place in this potential two-tiered battery, the test method would probably best be considered a "partial replacement" in 3Rs parlance, albeit one that also results in refinement and reduction.
- In this section of the HET-CAM BRD or elsewhere, it should be stated that:
  - additional optimization and validation studies should rely on existing *in vivo* data

- the low rate of false negatives (underpredictions) for HET-CAM has the animal welfare advantage of reducing the exposure of rabbits in the follow-on testing to severe irritants or corrosives
- any test method optimization should seek to further decrease the false negative rate

11.0 PRACTICAL CONSIDERATIONS

11.1 HET-CAM Test Method Transferability

The proposed test method, as detailed in Appendix A of the HET-CAM BRD, should be readily transferable to properly equipped and staffed laboratories. A video on the method and on scoring would make implementation easier and ensure correct conduct of the test method.

11.1.1 Facilities and Major Fixed Equipment Needed to Conduct the HET-CAM Test Method

The Panel agrees with the BRD on the facilities and major fixed equipment needed to conduct the HET-CAM test method. All the equipment and supplies seem to be readily available. In addition, technicians who are trained in the assay do not need to be trained in proper animal handling techniques, husbandry and all the other regulatory issues that arise when intact animals need to be housed and used.

11.1.2 General Availability of Other Necessary Equipment and Supplies

The Panel agrees with the BRD on the general availability of other necessary equipment and supplies.

11.2 HET-CAM Test Method Training

11.2.1 Required Training to Conduct the HET-CAM Test Method

The Panel agrees with the BRD on the required level of training and expertise needed for personnel to conduct the HET-CAM test method. In addition, training on the HET-CAM assay should involve both positive and negative controls, identifying the critical endpoints, and calculating the irritation indices.

11.2.2 Training Requirements Needed to Demonstrate Proficiency

The Panel agrees with the BRD on the training requirements needed for personnel to demonstrate proficiency. In addition, some kind of limited refresher training should be conducted periodical (e.g., every 2 years). A training video and other visual media on the technical aspects of the assay is recommended. Training approaches in the application of this test method should be developed and implemented for use in training.

11.3 Relative Cost of the HET-CAM Test Method

The Panel agrees with the BRD on the costs involved in conducting the *in vivo* test. Rabbit test costs are consistent with past experience.

11.4 Relative Time Needed to Conduct a Study Using the HET-CAM Test Method

The Panel agrees with the BRD on the amount of time needed to conduct a study. The duration of the *in vivo* rabbit eye test is consistent with past experience. 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 HET-CAM Test Method

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

The Panel agrees that the most appropriate version of the HET-CAM test method for use in a tiered-testing strategy and/or optimization and validation studies is the test method protocol recommended in the HET-CAM BRD. It is recommended that for the purpose of detecting severe eye irritants in the tiered-testing scheme outlined in the BRD, the HET-CAM test is useful for identification of severe or corrosive ocular irritants with the caveat that the HET-CAM has a high false positive rate. Positive results could be re-tested in a modified HET-CAM test method (e.g. using a lower concentration of test substance) to confirm the results. Alternatively, the positive substance could be tested in a different *in vitro* test method (e.g., ICE, IRE, BCOP). It is noted that data and information on the use of lower concentrations of test substances in the HET-CAM test method exist. Such information should be included in the BRD.

The proposed HET-CAM standardized test method protocol is adapted from the one by Spielmann and Liebsch (INVITTOX 1992). The method contains a negative control, a solvent control (if appropriate), a positive control and benchmark controls (if appropriate). Overall, the method is similar to those used by most investigators, but recommends using the time required for an endpoint to develop as the criteria for assessing irritation potential (Kalweit et al. 1987, 1990). The IS(B) method exhibited the highest accuracy rate (78%) and the lowest false negative rate (0%) in identifying ocular corrosives and severe irritants.

More specifically, the use of a standardized protocol in future studies will allow for new data to be generated, which will allow further evaluation of the usefulness and limitations of the recommended test method protocol. The proposed standardized HET-CAM test method protocol includes the use of concurrent positive and negative control test substances, whereas the published protocols are inconsistent on the use of such control test substances. Including concurrent control substances in the HET-CAM test method protocol allows for an assessment of experimental variability across time, establishment of a historical control database, and development of acceptance criteria for each test based on the positive control substance inducing an appropriate response. The test method protocol also recommends the inclusion of appropriate benchmark substances, where possible, to aid in evaluating the

ocular irritancy potential of test substances of a specific chemical class, or for evaluating the relative irritancy potential of a test substance within a specific range of irritant responses.

When using this method for substance classification, substances producing positive results (e.g., HET-CAM score defined as corrosive or severe irritant) obtained from this test method can be used to classify a substance as an ocular corrosive or severe irritant. Substances producing negative results (e.g., HET-CAM score defined as nonirritant, mild irritant, or moderate irritant) obtained from this test method would follow the tiered testing strategy.

12.2 Recommended Standardized HET-CAM Test Method Protocol

12.2.1 Appropriateness of the Recommended Standardized HET-CAM Test Method Protocol and Suggested Modifications to Improve Performance

The Panel recommends that procedures for applying and removing solids from the CAM be included in the standardized test method protocol. Solid substances may adhere to the CAM and demolish the CAM upon removal. Therefore, procedures for evaluating solids in this test method should be included in the test method protocol provided in Appendix A of the HET-CAM BRD.

Further optimization of the recommended standardized test method protocol should be possible. Optimization should increase the accuracy of the HET-CAM test method by reducing the moderate false positive rate while maintaining the low false negative rate. Optimization also should increase the reliability of the HET-CAM test method. Therefore, a retrospective analysis should be conducted to determine if different decision criteria might enhance the accuracy and/or reliability of the test method for the detection of ocular corrosives and severe irritants, as defined by the EU (2001), GHS (UN 2003), and EPA (1996) classification systems. Since it appears that the appropriate data are not available, a subset of substances in the recommended list of reference substances (HET-CAM BRD Section 12.4) should be tested to provide the necessary data.

12.2.2 Other Endpoints that Should be Incorporated into the HET-CAM Test Method

Other endpoints may be considered for use with the HET-CAM test method, but inclusion of these endpoints should not block retrospective validation of the HET-CAM test method with the parameters previously used to evaluate eye irritation potential.

The endpoints evaluated in HET-CAM are quite different from those evaluated in ICE, IRE, and BCOP, the organotypic test methods. For example, all three organotypic test methods include an evaluation of corneal opacity. Comparatively, the endpoints used in HET-CAM (development of lysis, hemorrhages, and coagulation) are unique to this test method; their use is based on proposed physiological similarities between the CAM and various structures of the eye (i.e., conjunctiva, cornea). Further optimization of the HET-CAM test method for the detection of ocular corrosives and severe irritants may be possible by considering different endpoints (e.g., trypan blue absorption, antibody staining, membrane changes) for evaluation and inclusion in the calculation of irritancy potential. Some of these may be comparable to those of the IRE, ICE and BCOP methods: membrane swelling, dye retention, visual evaluation, and microscopic evaluation. These additional tests may help reduce the

number of false positives with the HET-CAM test.

12.3 Recommended Optimization and Validation Studies

It is recommended that an evaluation to determine the relationship or predictability between the short-term effects observed in the HET-CAM and long-term effects observed in rabbits or humans be conducted. Such an evaluation may provide additional support for the use of the HET-CAM method to assess the delayed and long-term effects of corrosives and severe irritants.

12.3.1 Recommended Optimization Studies to Improve Performance of the Recommended HET-CAM Test Method Protocol

No optimization studies are needed to lower the false negative rate of the HET-CAM test method. However, studies to lower the false positive rate are needed. Optimization studies should make maximum use of retrospective analyses to preclude the need for further, time-consuming studies. Any further optimization and/or validation work should take full advantage of the modular approach to validation that the ECVAM is developing. The work could identify needed modules (e.g., interlaboratory reliability) and focus on gathering data for those needed modules. This would avoid the time and expense of a full-blown validation study.

It is recommended that any future optimization and validation studies should use existing animal data, if they are available. If important data gaps are identified, additional animal studies should only be conducted with the minimum number of animals. Such studies should be carefully designed to maximize the amount of pathophysiological (e.g., depth of injury) information obtained and conducted under GLP conditions. Any optimization and/or validation studies also should aim to minimize the number of animals used.

Optimization studies could increase the accuracy of the HET-CAM test method by reducing the moderate false positive rate while maintaining the low false negative rate. Therefore, a retrospective analysis should be conducted to determine if different decision criteria might enhance the accuracy of the test method for the detection of ocular corrosives and severe irritants, as defined by the EU (2001), GHS (UN 2003), and EPA (1996) classification systems. Optimization studies also may involve the development of a protocol that includes re-testing of positive substances using a modified HET-CAM test method protocol, as described above.

It is noted that optimizing a method involves validation of the method only if the modifications do not have a major impact on the conduct of the study. The recommendation to optimize and to use an optimized method should not minimize the value of data already obtained with the method of Spielmann and Liebsch (INVITTOX 1992). As some laboratories already apply this method, the data generated in these laboratories should still be valid and be used for labeling of corrosives and severe irritants.

An optimized test method may be used when a positive finding is obtained in the HET-CAM test method of Spielmann and Liebsch (INVITTOX 1992); the optimized protocol should be applied as a second step. This optimized protocol should then be validated.

The high variability of the Draize test does not allow for 100% accuracy with any of the recommended optimized methods or any other proposal for change. Because not enough human data are available, reference is made to the Draize test. However, this test cannot be seen as a "gold standard" (see **Section IV - 4.6** of this report) and should be defined as a "reference standard".

The Panel also recommends that this BRD section should discuss the pros and cons of the immediate implementation of the HET-CAM test for ocular corrosion and severe irritation. For example, the discussion should answer the question: What, if anything, is the downside of foregoing the proposed optimization and validation work?

12.3.2 Recommended Validation Studies to Evaluate Performance of the Optimized HET-CAM Test Method Protocol

If optimization of the method is done to reduce the false positive rate and modifications have a major impact on the conduct of the study, a validation study should be done with the optimized method. As the false negative rate is 0%, it is recommended that validation of the optimized method to reduce the false positive rate while maintaining the low false negative rate.[3]

The Panel also recommends identification of reference substances that would be included as part of the performance standards developed for the HET-CAM test method. These reference substances would be used to evaluate optimized test methods that are similar to the HET-CAM test method.

Minority Opinion

According to Dr. Martin Stephens, **Section IV – 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] Practical use of the IS(B) method in pharmaceutical industry for other purposes: In the pharmaceutical industry, the IS(B) analysis method is used to assess irritating potential of nasal or intravenous formulations. In this respect the IS(B) analysis method was found to be very powerful to select the right formulations. Formulations that were identified as nonirritants by the IS(B) analysis method did not induce irritation in animals. Intravenous formulations, which came out as severe irritating in the IS(B) analysis method induced severe irritation in the blood veins of animals even with necrosis of blood vessel cells (Vanparys P, personal communications).

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 HET-CAM BRD REFERENCES

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

It is recommended that the references in the public comments provided by Dr. med. Horst Spielmann, which lists relevant publications, should be included in the 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, de Silva O, Earl LK, Easty DL, Pape W, Spielmann H. 1998. Report on the COLIPA workshop on mechanisms of eye irritation. ATLA 26:811-820.

EPA. 1996. Label Review Manual. 2nd 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.

Fox DA, Boyes WK. 2001. Toxic responses of the ocular and visual system. In:Casarett & Doull's Toxicology: The Basic Science of Poisons, 6th Edition. (Klaassen CD ed). New York:McGraw-Hill Press, 565-596.

Fraunfelder FT, ed. 1982. Drug-induced ocular side effects and drug interactions. Philadephia, PA:Lea & Febiger.

Gautheron P, Giroux J, Cottin M, Audegond L, Morilla A, Mayordomo-Blanco L, Tortajada A, Haynes G, Vericat JA, Pirovano R, Tos EG, Hagemann C, Vanparys P, Deknudt G, Jacobs G, Prinsen M, Kalweit S, Spielmann H. 1994. Interlaboratory assessment of the bovine corneal opacity and permeability (BCOP) assay. Toxicol In Vitro 8:381-392.

Gilleron L, Coecke S, Sysmans M, Hansen E, van Oproy S, Marzin D, van Cauteren H, Vanparys P. 1996. Evaluation of a modified HET-CAM assay as a screening test for eye irritancy. Toxicol In Vitro 10:431-446.

Gilleron L, Coecke S, Sysmans M, Hansen E, van Oproy S, Marzin D, van Cauteren H, Vanparys P. 1997. Evaluation of the HET-CAM-TSA method as an alternative to the Draize eye irritation test. Toxicol In Vitro 11:641-644.

Grant WM. 1974. Toxicology of the eye; drugs, chemicals, plants, venoms. Springfield, IL:Thomas.

Hagino S, Kinoshita S, Tani N, Nakamura T, Ono N, Konishi K, Iimura H, Kojima H, Ohno Y. 1999. Interlaboratory validation of in vitro eye irritation tests for cosmetic ingredients. (2) Chorioallantoic membrane (CAM) test. Toxicol In Vitro 13:99-113.

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.

INVITTOX 1992. HET-CAM Test. ECVAM. Available: https://ecvam-sis.jrc.it/invittox/published/indexed_47.html [accessed 18 February 2004].

Kalweit S, Gerner I, Spielmann H. 1987. Validation project of alternatives for the Draize eye test. Mol Toxicol 1:597-603.

Kalweit S, Besoke R, Gerner I, Spielmann H. 1990. A national validation project of alternative methods to the Draize rabbit eye test. Toxicol In Vitro 4:702-706.

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.

Spielmann H, Liebsch M, Kalweit S, Moldenhauer F, Wirnsberger T, Holzhutter H, Schneider B, Glaser S, Gerner I, Pape WJW, Kreiling R, Krauser K, Miltenburger HG, Steiling W, Luepke NP Müller N, Kreuzer H, Mürmann P, Spengler J, Bertram-Neis, E, Siegemund B, Wiebel F. 1996. Results of a validation study in Germany on two *in vitro* alternatives to the Draize eye irritation test, HET-CAM test and the 3T3 NRU cytotoxicity test. ATLA 24:741-858.

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]

Proposed Reference Substances for Validation Studies

[This Page Intentionally Left Blank]

V. PROPOSED REFERENCE SUBSTANCES FOR VALIDATION STUDIES

1.0 ADEQUACY AND COMPLETENESS OF THE RECOMMENDED LIST OF REFERENCE SUBSTANCES

The list of proposed substances is fairly comprehensive in that the three major groups of products to which the eye is exposed (i.e., industrial chemicals, pharmaceuticals, cosmetics) are represented. Individual substances have been chosen based on: the availability of high quality *in vivo* data; commercial availability; lack of excessive hazard or prohibitive disposal costs. The substances appear to be readily available and in acceptably pure form. The range of possible ocular toxicity responses in terms of severity and types of lesions appears to be represented. Appropriately, there are presently no substances with color that will interfere with the observation of the endpoints. However, while the list covers a broad range of organic chemical classes, only two inorganic substances (sodium hydroxide and ammonium nitrate) were included. If possible, additional inorganic chemicals (including more alkali substances) that are used in consumer products should be included. Surfactants are over-represented and correspond to an area where the panel can make selective recommendations. The use of substances at different concentrations (which are included in the reference list) is important as it allows for determination of test sensitivity. However, different substance concentrations should not be included in early studies that evaluate reproducibility. The source of the *in vivo* data should be provided in the list of reference substances in each BRD. For clarity, the identity of the individuals charged with selecting the list of reference chemicals should be specified in each BRD and any potential biases among these individuals identified. Conversely, classification data for each *in vitro* test should not be included in a list of test substances that are proposed for validating *in vitro* tests, and therefore this information should be removed from the list.

Where applicable, within a chemical class, substances of lower, medium and higher molecular weight should be included (although as noted above, it is recognized that selection of substances may have been limited by the availability of high quality *in vivo* rabbit eye test data). Finally, the recommended substances should represent the entire spectrum of injury as defined by each *in vivo* test.

To declare this list adequate and complete is difficult. The current list has entirely too many substances and, thus, is unwieldy. Perhaps, a worthy effort would be to select from the list an appropriate number of specific substances that the Panel believes optimal for validation and optimization studies.

With that in mind, one possible approach for determining the adequate and most efficient number of substances could be to employ a two-stage study design for validation studies. In this two-stage approach, the first stage would be for a subset of substances to be tested in multiple laboratories to yield an estimate of test method reliability. The substances to be included in each stage would be selected from the list of recommended reference substances included in Section 12.4 of each test method BRD. In the first stage, a subset of substances (e.g., n = 10) could be tested in multiple laboratories to yield an estimate of test method reliability. Because negative substances provide little information with regard to test method

reliability, severe ocular irritants/corrosives should be the focus of this stage. Also, the nonsevere irritants or nonirritants that would be included (e.g., n = 2) should be moderate irritants (i.e., GHS Category 2A). This initial set of substances would cover a broad range of chemical classes, as well as encompassing the range of GHS Category 1 responses (i.e., GHS Category 1 subcategories as detailed in Section 12.4 of each test method BRD; one per chemical class and including at least one per Category 1 subcategory). Product class does not seem to be as important a factor in selecting test substances. In constructing this initial list of reference substances, the focus might be on substances to which individuals are most likely to come into contact (e.g., the 50 highest production volume non-polymeric substances in commerce). In most instances, volume of production (apart from pharmaceuticals) is a good surrogate for risk of exposure. However, it is recognized that inclusion of substances in this list is limited in part by the availability of high-quality *in vivo* rabbit eye test reference data. Therefore, representatives from the following classes would seem most appropriate for inclusion in this list: acids (organic and mineral); alkalis; amines, imines, and amides; alcohols (including polyols); ethers; esters; thiols; halides; quaternary ammonium compounds; N- and S- heterocyclics; and hydrocarbons. The list should also include a reasonable range of molecular weights, but no formulations, prototypes, or products should be included, and testing should be in several laboratories. Limiting this initial list to liquid substances (as they represent the majority of substances for which "real world" testing would be performed) would also minimize the complexity of the resulting analysis that would result from the inclusion of too many variables in this early stage.

If results from this initial stage indicate that the test method is suitably reliable, a second stage that includes a larger number of substances could be conducted to evaluate test method accuracy. During this stage, the list of substances to be tested would be expanded to include multiple representatives from each chemical class and GHS Category 1 subcategory. In addition, within each chemical class, testing substances of different physical properties (solubility, molecular weight, pH) would seem appropriate, where feasible. At issue during this stage would be the appropriate number of chemical classes necessary to assess accuracy, and the extent of generalization of results that would be anticipated across classes. A possible design might include a set of five substances per class (covering the range of irritancy responses).

Presently in each test method BRD, the criteria for selection include "substances which represent the range of known or anticipated mechanisms or modes of action for severe/irreversible ocular irritation or corrosion." Section 1.2.2 of each test method BRD purports to discuss similarities and differences of modes and mechanisms of action between the *in vitro* test method and ocular irritancy in humans and/or rabbits. Despite a very illuminating discussion of the anatomy of the human, rabbit, bovine, and/or chicken eye, there is no discussion of mechanism of action of irritants, only a description of the effects. That criterion for agent selection should be deleted or appropriate justification provided.

Regarding health and safety concerns, laboratory personnel doing the testing should be well trained in general safety associated with handling of potentially toxic chemicals. Information regarding the test substances with respect to handling and inadvertent exposure should be readily available, if needed. Therefore, for all validation studies, Material Safety Data Sheets

(MSDS) for the recommended substances should be provided (i.e., as a coded MSDS) and prestudy safety briefings should be conducted.

2.0 OTHER CRITERIA THAT SHOULD BE ADDRESSED IN THE SELECTION OF REFERENCE SUBSTANCES

Substances known to induce severe lesions, *in vivo*, in the eyes of humans should be included, even in the absence of rabbit data.