APPENDIX A
EXPERT PANEL REPORTS

[This Page Intentionally Left Blank]

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.

[This Page Intentionally Left Blank]

APPENDIX A2
EXPERT PANEL REPORT: EVALUATION OF THE CURRENT VALIDATION STATUS OF *IN VITRO* TEST METHODS FOR IDENTIFYING OCULAR CORROSIVES AND SEVERE IRRITANTS - ADDENDUM

[This Page Intentionally Left Blank]

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

November 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/ EPrptAddend.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 A. 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

PREFACE

On November 1, 2004, The National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) made available draft Background Review Documents (BRDs) that provided information and data about the current validation status of four *in vitro* test methods for detecting ocular corrosives and severe irritants. The four test methods were the Bovine Corneal Opacity and Permeability (BCOP) assay, the Hen's Egg Test - Chorioallantoic Membrane (HET-CAM) assay, the Isolated Chicken Eye (ICE) assay, and the Isolated Rabbit Eye (IRE) assay. These draft BRDs were based on published studies using the identified test methods, and other data and information submitted in response to a 2004 *Federal Register* (*FR*) request.

The Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) convened an Expert Panel meeting on January 11-12, 2005, to independently assess the validation status of these four *in vitro* test methods for identifying ocular corrosives or severe irritants, as determined by the rabbit response. Public comments at the meeting revealed that additional relevant data were available that had not yet been provided in response to earlier requests for data. The Expert Panel recommended that the additional data be requested and that a reanalysis of the accuracy and reliability of each test method be conducted where appropriate.

In response to this recommendation, a *FR* notice was published on February 28, 2005. The notice requested all available *in vitro* data on these four *in vitro* ocular irritancy test methods and corresponding *in vivo* rabbit eye test method data, as well as any human exposure data (either via ethical human studies or accidental exposure). A request for relevant data was re-sent directly to the primary developers or users of each test method. In response to these requests, additional *in vitro* test method data and corresponding *in vivo* rabbit eye test results were submitted for the BCOP, HET-CAM, and ICE test methods, which were used for the reanalyses presented in this BRD Addendum.

In addition to the additional test method data, clarification of European Union (EU 2001) and United Nations (UN) Globally Harmonized System (GHS) (UN 2003) ocular hazard classification rules for severe irritants was obtained subsequent to the release of the four draft BRDs. This change resulted in 10 to 15 substances being reclassified based on their *in vivo* data from nonsevere to severe irritants, depending on which *in vitro* ocular irritancy test method and ocular hazard classification system was used.

The original draft BRDs also provided an evaluation of the accuracy of each test method by chemical class. The chemical classes assigned to each test substance were revised based on a chemical classification system consistent with the U.S. National Library of Medicine's Medical Subject Headings (MeSH), an internationally recognized standardized classification scheme. This scheme was used to ensure consistency in classifying substances by chemical class among all the *in vitro* ocular test methods under consideration, and resulted in some chemicals being re-classified into different chemical classes. As a result, the accuracy of each test method by chemical class was reanalyzed; the results of each reanalysis are also provided in this BRD Addendum.

Finally, an additional accuracy analysis was conducted. In this analysis, the accuracy of each *in vitro* ocular irritancy test method for detecting ocular corrosives or severe irritants, depending on whether the *in vivo* rabbit classification was based on the severity of the response and/or its persistence to day 21 post-treatment, was determined.

A list of proposed reference substances for validation of *in vitro* tests to detect ocular corrosives and severe irritants was included in the draft BRDs released on November 1, 2004. The BRD Addendum provides a revised list of proposed reference chemicals, which was prepared after consideration of the following:

- recommendations of the Expert Panel that resulted from their deliberations on January 11-12, 2005
- submission of additional Draize rabbit eye test results for approximately 300 substances
- clarification regarding the GHS rules for classification of severe irritants (UN 2003) that resulted in the reclassification of two proposed reference substances from nonsevere to severe irritants
- reassignment of the candidate reference substances to chemical classes using MeSH (NLM 2005)

The BRD Addendum was released on July 26, 2005, with notification of its release via an *FR* notice and notification through the ICCVAM electronic mailing list. The Panel was subsequently reconvened via teleconference on September 19, 2005 to discuss the BRD Addendum. Prior to this meeting, public comments on the Addendum were received from three organizations and provided to the Panel for their consideration. The Panel provided formal comment on each of the four *in vitro* test methods, as well as the proposed list of reference substances. In addition, the public were provided time at the public meeting to comment (although no public comments were provided). The Panel then provided final endorsement regarding the effects, if any, of the information in the BRD Addendum on their original evaluation from the January 11-12, 2005 meeting.

EXECUTIVE SUMMARY

This report describes the conclusions and recommendations of the Expert Panel ("Panel") made during the September 19, 2005 teleconference on the utility of four *in vitro* ocular toxicity test methods for identifying ocular corrosives and severe irritants (i.e., the Bovine Corneal Opacity and Permeability [BCOP] assay, the Hen's Egg Test - Chorioallantoic Membrane [HET-CAM] assay, the Isolated Chicken Eye [ICE] assay, and the Isolated Rabbit Eye [IRE] assay). This second Panel report is a supplement to the March 2005 report entitled, "Expert Panel Report: Evaluation of the Current Validation Status of *In Vitro* Test Methods for Identifying Ocular Corrosives and Severe Irritants." Unless indicated, all conclusions and recommendations made by the Panel in their March 2005 report remain unchanged.

For each test method, the Panel was asked to determine if the information provided in the Addendum to the November 2004 Background Review Documents (BRD) were appropriate for inclusion in the accuracy and reliability re-analyses, and then if any changes to the original recommendations established at the January 11-12, 2005 meeting (http://iccvam.niehs.nih.gov/methods/ocudocs/EPreport/ocureport.htm) were warranted based on the updated information detailed in the BRD Addendum. The Panel agreed that, for all four test methods, the information in the BRD Addendum was appropriate for inclusion, and that no errors or omissions were present. For three of the four methods (i.e., BCOP, ICE and IRE), the Panel agreed that there was no basis for changing the original conclusions and recommendations established at the January 11-12, 2005 meeting. However, the Panel concluded that, given the increases in both false positive and false negative rates based on the reanalysis, the HET-CAM IS(B) analysis method, using the decision criteria of Leupke, 1985, may have limited utility for the identification of severe ocular irritants and/or corrosives. In contrast, during the January 11-12, 2005 meeting, the Panel concluded that, for the purpose of detecting severe eye irritants in the tiered-testing strategy, the HET-CAM test has been shown to be useful for the identification of severe or corrosive ocular irritants.

The Panel was also asked to consider the adequacy of the proposed list of reference substances, which was updated (in part) based on comments received from the Panel at the January 11-12, 2005 meeting. The Panel reaffirmed the comments stated in the original Panel report and still considered the list too large if the list is intended to be the minimum number of substances required for validation of a new test method. The Panel also recommended that substances of the highest purity available from major suppliers be used.

During the deliberations of the Panel, the question was raised as to how closely the performance of an *in vitro* test must match the performance of an *in vivo* test before the *in vitro* test is considered a sufficiently accurate measure of the risk to humans. It was acknowledged that this was an appropriate and important question to bring to ICCVAM, but one that was beyond the scope of the charge to this expert panel.

[This Page Intentionally Left Blank]

I. The Isolated Rabbit Eye (IRE) Test Method

On January 11-12, 2005, 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.

During the September 19, 2005 Panel meeting, three questions were addressed with regard to the IRE BRD Addendum as follows:

Is the information provided in the Addendum to the November 2004 Background Review Document (BRD) appropriate for inclusion in the accuracy and reliability analyses?

The Panel concluded that the information was appropriate.

Are there any errors or omissions that should be corrected?

The Panel agreed that there were no errors or omissions. The Panel recognized and supported the rationale for excluding some substances from the evaluation based on lack of adequate *in vivo* rabbit eye test data (i.e., severe ocular irritancy/corrosivity classification based solely on skin corrosivity, pH extremes, etc., or no classification feasible based on eye test data provided to NICEATM). While the pH and/or dermal corrosive effects of a test substance are utilized as substitutes for animal eye irritation data for the purposes of ocular hazard classification, the goal of this evaluation was to determine whether the four *in vitro* test methods can be used to predict the outcome of the *in vivo* rabbit eye test for the same test substance. Therefore, including data on pH extremes and/or dermal corrosivity (in the absence of *in vivo* rabbit eye test data) was judged to be inappropriate due to the uncertainty of its performance in predicting the *in vivo* rabbit eye test outcome. In addition, the Panel recommended that text be included in the final BRD to underscore the fact that, where such information was available, data derived from scientifically acceptable *in vivo* rabbit eye tests terminated based on humane endpoints (e.g., severe discomfort) were included in the accuracy and reliability analysis.

Based on the revised accuracy and reliability analyses for the IRE test method for identifying ocular corrosives and severe irritants, does this new information provide the basis for any changes in the Panel's conclusions and recommendations made at the January 11-12, 2005 meeting?

The Panel agreed that there was no basis for changes to the original conclusions and recommendations.

II. The Isolated Chicken Eye (ICE) Test Method

At the January 11-12, 2005 Expert Panel meeting, the Panel concluded that the ICCVAM criteria for validation (ICCVAM 2003) have not been fully met for the ICE test method. Cited deficiencies included:

- 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
- 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 previously recommended that, 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. Subsequent to the January 11-12, 2005 meeting, additional data were received that allowed for such analyses to be conducted.

During the September 19, 2005 Panel meeting, three questions were addressed with regard to the ICE BRD Addendum as follows:

Is the information provided in the Addendum to the November 2004 BRD appropriate for inclusion in the accuracy and reliability analyses?

The Panel concluded that the information was appropriate.

Are there any errors or omissions that should be corrected?

The Panel agreed that there were no errors or omissions. The Panel recognized and supported the rationale for excluding some substances from the evaluation based on lack of adequate *in vivo* rabbit eye test data (i.e., severe ocular irritancy/corrosivity classification based solely on skin corrosivity, pH extremes, etc., or no classification feasible based on eye test data provided to NICEATM). While the pH and/or dermal corrosive effects of a test substance are utilized as substitutes for animal eye irritation data for the purposes of ocular

hazard classification, the goal of this evaluation was to determine whether the four *in vitro* test methods can be used to predict the outcome of the *in vivo* rabbit eye test for the same test substance. Therefore, including data on pH extremes and/or dermal corrosivity (in the absence of *in vivo* rabbit eye test data) was judged to be inappropriate due to the uncertainty of its performance in predicting the *in vivo* rabbit eye test outcome. In addition, the Panel recommended that text be included in the final BRD to underscore the fact that, where such information was available, data derived from scientifically acceptable *in vivo* rabbit eye tests terminated based on humane endpoints (e.g., severe discomfort) were included in the accuracy and reliability analysis.

Based on the revised accuracy and reliability analyses for the ICE test method for identifying ocular corrosives and severe irritants, does this new information provide the basis for any changes in the Panel's conclusions and recommendations made at the January 11-12, 2005 meeting?

The Panel agreed that there was no basis for changes to the original conclusions and recommendations. The Panel added that the reanalysis using the new information and the GHS classification scheme showed that the test performance was essentially unchanged (1-2% difference) or directionally poorer (Table ES-1 in the Addendum).

III. The Bovine Corneal Opacity and Permeability (BCOP) Test Method

At the January 11-12, 2005 Expert Panel meeting, 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.

While the Panel believed these modifications were 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.

During the September 19, 2005 Panel meeting, three questions were addressed with regard to the BCOP BRD Addendum as follows:

Is the information provided in the Addendum to the November 2004 BRD appropriate for inclusion in the accuracy and reliability analyses?

The Panel concluded that the information was appropriate.

Are there any errors or omissions that should be corrected?

The Panel agreed that there were no errors or omissions. The Panel recognized and supported the rationale for excluding some substances from the evaluation based on lack of adequate *in vivo* rabbit eye test data (i.e., severe ocular irritancy/corrosivity classification based solely on skin corrosivity, pH extremes, etc., or no classification feasible based on eye test data provided to NICEATM). While the pH and/or dermal corrosive effects of a test substance are utilized as substitutes for animal eye irritation data for the purposes of ocular hazard classification, the goal of this evaluation was to determine whether the four *in vitro* test methods can be used to predict the outcome of the *in vivo* rabbit eye test for the same test substance. Therefore, including data on pH extremes and/or dermal corrosivity (in the absence of *in vivo* rabbit eye test data) was judged to be inappropriate due to the uncertainty of its performance in predicting the *in vivo* rabbit eye test outcome. In addition, the Panel recommended that text be included in the final BRD to underscore the fact that, where such information was available, data derived from scientifically acceptable *in vivo* rabbit eye tests

terminated based on humane endpoints (e.g., severe discomfort) were included in the accuracy and reliability analysis.

Based on the revised accuracy and reliability analyses for the BCOP test method for identifying ocular corrosives and severe irritants, does this new information provide the basis for any changes in the Panel's conclusions and recommendations made at the January 11-12, 2005 meeting?

The Panel agreed that there was no basis for changes to the original conclusions and recommendations.

IV. The Hen's Egg Test - Chorioallantoic Membrane (HET-CAM) Test Method

At the January 11-12, 2005 Expert Panel meeting, 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 further 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.

Subsequent to the January 11-12, 2005 meeting, additional data were received and the full data set was reanalyzed. During the September 19, 2005 Panel meeting, three questions were addressed with regard to the HET-CAM BRD Addendum as follows:

Is the information provided in the Addendum to the November 2004 BRD appropriate for inclusion in the accuracy and reliability analyses?

The Panel concluded that the information was appropriate.

Are there any errors or omissions that should be corrected?

The Panel agreed that there were no errors or omissions. The Panel recognized and supported the rationale for excluding some substances from the evaluation based on lack of adequate *in vivo* rabbit eye test data (i.e., severe ocular irritancy/corrosivity classification based solely on skin corrosivity, pH extremes, etc., or no classification feasible based on eye test data provided to NICEATM). While the pH and/or dermal corrosive effects of a test substance are utilized as substitutes for animal eye irritation data for the purposes of ocular hazard classification, the goal of this evaluation was to determine whether the four *in vitro* test methods can be used to predict the outcome of the *in vivo* rabbit eye test for the same test substance. Therefore, including data on pH extremes and/or dermal corrosivity (in the absence of *in vivo* rabbit eye test data) was judged to be inappropriate due to the uncertainty of its performance in predicting the *in vivo* rabbit eye test outcome. In addition, the Panel recommended that text be included in the final BRD to underscore the fact that, where such information was available, data derived from scientifically acceptable *in vivo* rabbit eye tests terminated based on humane endpoints (e.g., severe discomfort) were included in the accuracy and reliability analysis.

Based on the revised accuracy and reliability analyses for the HET-CAM test method for identifying ocular corrosives and severe irritants, does this new information provide the basis for any changes in the Panel's conclusions and recommendations made at the January 11-12, 2005 meeting?

As indicated above, at the January 11-12, 2005 meeting, 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. However, at the September 19, 2005 meeting, the Panel concluded that, given the increases in both false positive and false negative rates, the HET-CAM IS(B) analysis method, using the decision criteria of Leupke, 1985, may have limited utility for the identification of severe ocular irritants and/or corrosives, although it may be useful for the identification of mild to moderate irritants. As stated in the Panel's March 2005 Report, 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.

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

At the January 11-12, 2005 Expert Panel meeting, 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. However, the Panel concluded that: 1) the current list has too many substances; 2) surfactants are over-represented; 3) more inorganic substances should be added; and 4) substances known to induce severe ocular lesions in humans should be included in the list, even in the absence of rabbit data.

The Panel noted that the number of inorganic substances in the revised list of proposed reference substances was increased from 2 to 11; that the current list includes 10 substances that are known human ocular corrosives or severe irritants, even in the absence of *in vivo* rabbit data; that all formulations were removed; and that the number of surfactants were decreased from 12 to 7. However, the total number of proposed reference substances was increased from 89 to 122. ICCVAM justifies this increase because, for the detection of ocular corrosives and severe irritants, the list of substances needs to include substances that:

- Induce very severe responses within a relatively short period, as well as those where the response is delayed
- Adversely affect the cornea, iris, and/or conjunctiva
- Induce persistent and non-persistent lesions
- Represent a diverse population of chemical classes and physicochemical properties

During the September 19, 2005 Panel meeting, one question was addressed with regard to the recommended list of reference substances included in the BRD Addendum as follows:

Is the revised list of proposed reference substances, selected from the list of available candidate substances, sufficiently adequate and complete for validation studies to evaluate the usefulness and limitations of in vitro test methods proposed for identifying ocular severe irritants and corrosives?

The Panel reaffirmed the comments stated in the original Panel report (e.g., providing the list as a reference from which to generate a subset of substances to be used for evaluating *in vitro* ocular toxicity test methods on a scientifically sound, case-by-case basis.) The Panel still considered the list too large if the list is intended, as stated in the BRD Addendum, to be the minimum number of substances that should be used for validation of a new test method. A focus on mechanism of action may reduce the number of substances that need to be used to evaluate the relevance and reliability of a proposed test method. The Panel recommended that the highest purity level available from major suppliers for each substance be used and ideally, information on major impurities provided.

APPENDIX A3
SUMMARY MINUTES FROM EXPERT PANEL MEETING ON JANUARY 11-12, 2005

[This Page Intentionally Left Blank]

Department of Health and Human Services
National Institutes of Health
National Institute of Environmental Health Sciences
Interagency Coordinating Committee on the Validation of Alterative Methods (ICCVAM)
Expert Panel Meeting

Summary Minutes of the Expert Panel Meeting to Assess the Current Validation Status of *In Vitro* Test Methods for Identifying Ocular Corrosives and Severe Irritants.

Introduction

A public meeting of an independent Expert Panel was convened on January 11-12, 2005, at the National Institutes of Health (NIH), Natcher Center, Bethesda, Maryland, to evaluate several i*n vitro* ocular irritation test methods. The purpose of this meeting was to assess the current validation status of the Bovine Corneal Opacity and Permeability (BCOP), Isolated Chicken Eye (ICE), Isolated Rabbit Eye (IRE), and Hen's Egg Test-Chorioallantoic Membrane (HET-CAM) test methods for identifying ocular corrosives and severe irritants. The meeting was organized by ICCVAM and the National Toxicology Program (NTP) Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM), and sponsored by the National Institute of Environmental Health Sciences (NIEHS) and the NTP. A comprehensive report of the Expert Panel is provided as an attachment to these minutes.

The following scientists served on the Expert Panel:

- Robert Scala, Ph.D., (Panel Chair), Tucson, Arizona, United States
- Sally S. Atherton, Ph.D., Professor, Medical College of Georgia, Augusta, Georgia, United States
- Roger Beuerman, Ph.D., Professor, Louisiana State University, New Orleans, Louisiana, United States
- June Bradlaw, Ph.D., International Foundation for Ethical Research, Rockville, Maryland, United States
- Ih Chu, Ph.D., Health Canada, Ottawa, Canada
- Henry Edelhauser, Ph.D., Professor, Emory University, Atlanta, Georgia, United States
- Donald Fox, Ph.D., Professor, University of Houston, Houston, Texas, United States
- Jim Freeman, Ph.D., Lab Director, ExxonMobil Biomedical Sciences, Inc., Annandale, New Jersey, United States
- Sidney Green, Ph.D., A.T.S., Graduate Professor, Howard University, Washington, DC, United States
- Frederick Guerriero, M.S., Senior Occupational Toxicologist, GlaxoSmithKline, King of Prussia, Pennsylvania, United States

- 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, United States
- Hiroshi Itagaki, Ph.D., Deputy Director of JSAAE, Manager of Alternative Section, 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 JSAAE, National Institute of Health, Japan
- Robert Peiffer, D.V.M., D.A.C.V.O., Senior Investigator, Merck Research Laboratories, West Point, Ohio, United States
- Lionel Rubin, V.M.D., D.A.C.V.O., Emeritus Professor of Ophthalmology, University of Pennsylvania, Philadelphia, Pennsylvania, United States
- 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, United States
- Katherine Stitzel, D.V.M., Consultant, West Chester, Ohio, United States
- 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, United States
- Scheffer Tseng, M.D., Ph.D., Director, Ocular Surface Research and Education Foundation, Miami, Florida, United States
- Philippe Vanparys, Ph.D., Senior Research Fellow, Johnson and Johnson, Belgium

The following scientists were invited guests:

- Dr. Chantra Eskes, European Centre for the Validation of Alternative Methods, Ispra, Italy
- Mr. Robert Guest, SafePharm Industries, Derby, United Kingdom
- Dr. John Harbell, Institute for *In Vitro* Sciences, Gaithersburg, Maryland, United States
- Dr. Klaus Krauser, Abbott Laboratories, Abbott Park, Illinois, United States
- Mr. Menk Prinsen, TNO Nutrition & Food Institute, The Netherlands

The following ICCVAM agency representatives were present:

- Dr. Robert Bronaugh, (Ocular Toxicity Working Group), U.S. Food and Drug Administration
- Dr. Kailash Gupta, (OTWG), U.S. Consumer Product Safety Commission
- Dr. Karen Hamernik, (OTWG), U.S. Environmental Protection Agency
- Dr. Abigail Jacobs, (OTWG), U.S. Food and Drug Administration
- Ms. Deborah McCall (OTWG), U.S. Environmental Protection Agency
- Dr. Amy Rispin (OTWG), U.S. Environmental Protection Agency
- Dr. Leonard Schechtman, U.S. Food and Drug Administration

- Dr. Margaret Snyder, National Institutes of Health
- Dr. William Stokes, (OTWG), National Institute of Environmental Health Sciences
- Dr. Marilyn Wind, U.S. Consumer Product Safety Commission

The following additional members of the ICCVAM OTWG were present:

- Dr. Meta Bonner, U.S. Environmental Protection Agency
- Dr. Wiley Chambers, U.S. Food and Drug Administration
- Ms. Donnie Lowther, U.S. Food and Drug Administration
- Dr. Jill Merrill, U.S. Food and Drug Administration

The following members of the NICEATM Staff were present:

- Dr. David Allen, Integrated Laboratory Systems, Inc.
- Mr. Bradley Blackard, Integrated Laboratory Systems, Inc.
- Dr. Neepa Choksi, Integrated Laboratory Systems, Inc.
- Ms. Christina Inhof, Integrated Laboratory Systems, Inc.
- Ms. Linda Litchfield, Integrated Laboratory Systems, Inc.
- Ms. Debbie McCarley, National Institute of Environmental Health Sciences
- Dr. Raymond Tice, Integrated Laboratory Systems, Inc.
- Mr. James Truax, Integrated Laboratory Systems, Inc.

The following members of the public were present:

- Ms. Sara Amundson, Doris Day Animal League
- Dr. Daniel Bagley, Colgate-Palmolive
- Ms. Kathleen C. Cater, The Dial Corporation
- Ms. Nicole Cuellar, S.C. Johnson & Son, Inc.
- Dr. Rodger D. Curren, Institute for *In Vitro* Sciences, Inc.
- Ms. Sadhana Dhruvakumar, People for the Ethical Treatment of Animals
- Dr. Carol Eisenmann, Cosmetic, Toiletry and Fragrance Association
- Ms. Megha S. Even, Physicians Committee for Responsible Medicine
- Ms. Myra Karstadt, U.S. Environmental Protection Agency
- Mr. Ray Kemper, DuPont Haskell Lab
- Ms. Sue A. Leary, Alternatives Research and Development Foundation
- Dr. Dan Marsman, Procter and Gamble
- Mr. David J. McCanna, Bausch & Lomb
- Mr. Claude McGowan, Johnson and Johnson CPDW
- Dr. Pauline M. McNamee, The European Cosmetic, Toiletry, and Perfumary Association (COLIPA)
- Mr. Hidenori Meiseki, Dojindo
- Mr. Hans A. Raabe, Institute for *In Vitro* Sciences
- Dr. Michael W. Rohovsky, Johnson and Johnson
- Mr. Chad Sandusky, Physicians Committee for Responsible Medicine
- Mr. Dean Scott, Bureau of National Affairs News

- Ms. Judith E. Swanson, S.C. Johnson & Son, Inc.
- Dr. Kristina Thayer, National Institute for Environmental Health Sciences
- Dr. Kevin J. Trouba, Institute for *In Vitro* Sciences, Inc.
- Ms. Amanda Ulrey, Institute for *In Vitro* Sciences, Inc.
- Ms. Sarah B. Vieh, The Rose Sheet
- Dr. Sherry L. Ward, Physicians Committee for Responsible Medicine
- Mr. Keith Wyatt, National Eye Institute
- Mr. Gary Wnorowski, Product Safety Labs

The purpose of this meeting was to evaluate the validation status of *in vitro* test methods for identifying ocular corrosives and severe irritants. The Expert Panel was asked to evaluate four draft background review documents (BRDs) prepared by NICEATM.

The four BRDs reviewed and discussed were:

- Current Status of *In Vitro* Test Methods for Identifying Ocular Corrosives and Severe Irritants: The Bovine Corneal Opacity and Permeability (BCOP) Test Method
- Current Status of *In Vitro* Test Methods for Identifying Ocular Corrosives and Severe Irritants: The Hen's Egg Test-Chorioallantoic Membrane (HET-CAM) Test Method
- Current Status of *In Vitro* Test Methods for Identifying Ocular Corrosives and Severe Irritants: The Isolated Chicken Eye (ICE) Test Method
- Current Status of *In Vitro* Test Methods for Identifying Ocular Corrosives and Severe Irritants: The Isolated Rabbit Eye (IRE) Test Method.

Call to Order and Introductions

Dr. Robert Scala, Panel Chair, called the meeting of the Expert Panel (Panel) to order at 8:30 a.m. and asked each attendee to state their name and affiliation. Dr. Scala stated that the public would be given the opportunity to speak at various times during the meeting. Each speaker from the public would be limited to seven (7) minutes, and anyone addressing the group should state their name for the benefit of the transcriptionist.

Dr. William Stokes, Executive Secretary for the Expert Panel and the designated government official, read the Statement of Conflict of Interest and explained policies and procedures regarding confidentiality and avoidance of conflict of interest, as follows:

The members of this expert panel serve as individual scientists and not as representatives of any organization. Each member is to exercise judgment as to whether a potential conflict of interest might exist relative to one or more of the topics being discussed due to his or her occupational affiliation, professional activity or financial interest. Should there be a potential conflict of interest, the member is to recuse him or herself from participating in the discussion of panel recommendations and/or decisions on the topic. You will be signing a conflict of interest certification which declares that during this panel meeting you did not participate in discussion of panel recommendations and/or decisions that involve a particular matter that could have a direct and predictable effect on: 1) Any organization,

institution or university system in which a financial interest exists for yourself, spouse, parent, minor child or partner. 2) Any organization in which you, your spouse, parent, minor child or partner serves as an officer, director, trustee or employee or is otherwise similarly associated. 3) Any organization with which you, your spouse, parent, minor child or parent [sic] is negotiating or have any arrangements concerning prospective employment or other such associations. Panel members are asked to identify at the beginning of this meeting the nature of any such conflicts.

None of the Panel members declared a conflict of interest.

Overview of the ICCVAM Test Method and Evaluation Process

Dr. Stokes (Director, NICEATM, NIEHS) provided a brief overview of ICCVAM and NICEATM. ICCVAM was established as an *ad hoc* committee in 1994 in response to revisions in the 1993 NIH Revitalization Act (Public Law [P.L.] 103-43) that mandates that the NIEHS develop criteria for validation and regulatory acceptance of test methods, and develop a process to achieve regulatory acceptance of scientifically valid methods. The *ad hoc* committee issued its report in 1997, and the ICCVAM committee was formally established that year to implement P.L. 103-43 directives. In 2000, the ICCVAM Authorization Act (P.L. 106-545) established ICCVAM as a permanent committee.

The 15 member agencies of ICCVAM include those involved in regulatory and research activities (CPSC; Department of Agriculture [DOA]; Department of the Interior [DOI]; Department of Transportation [DOT]; U.S. Environmental Protection Agency [EPA]; U.S. Food and Drug Administration [FDA]; Occupational Safety and Health Administration [OSHA]) and those involved in non-regulatory research (Agency for Toxic Substances and Disease Registry [ATSDR]; Department of Defense [DOD]; Department of Energy [DOE]; National Cancer Institute [NCI]; NIEHS; National Institute for Occupational Safety and Health [NIOSH]; National Library of Medicine [NLM]; NIH).

The purposes of ICCVAM, as set forth in P.L. 106-545, are to:

- Increase efficiency and effectiveness of U.S. Federal agency test method review
- Eliminate unnecessary duplicative efforts and share experiences between U.S. Federal regulatory agencies
- Optimize utilization of scientific expertise outside the U.S. Federal Government
- Ensure that new and revised test methods are validated to meet the needs of U.S. Federal agencies
- Reduce, refine, or replace the use of animals in testing (i.e., 3Rs), where feasible

The duties of ICCVAM are to:

- Facilitate and provide guidance on test method development, validation criteria, and validation processes

- Consider petitions from the public for review and evaluation of validated test methods
- Facilitate acceptance of scientifically valid test methods
- Review and evaluate new or revised or alternative test methods applicable to regulatory testing
- Submit test method recommendations to U.S. Federal agencies
- Facilitate interagency and international harmonization of test methods

NICEATM is located at NIEHS in Research Triangle Park, North Carolina, and has the following responsibilities:

- Administers ICCVAM
- Provides operational and technical support for ICCVAM activities
- Communicates and partners with stakeholders
- Organizes test method peer reviews, expert panel meetings, and workshops
- Conducts independent validation studies, as resources permit

The definition of validation used by ICCVAM is that it is the process by which the reliability and relevance of a procedure are established for a specific purpose. Validation characterizes the usefulness and limitations of a test method for a specific purpose. Adequate validation is a prerequisite for regulatory acceptance.

The criteria for test method validation are:

1. Clear statement of proposed use and regulatory rationale
2. Biological basis/mechanistic relationship to effect of interest
3. Formal detailed protocol
4. Reliability adequately assessed
5. Relevance adequately assessed
6. Limitations described
7. All data (raw) available for review
8. Data quality: Ideally generated according to Good Laboratory Practices (GLPs)
9. Independent scientific peer review

The criteria for test method acceptance are:

1. Fits into the regulatory testing structure
2. Adequately predicts the toxic endpoint of interest
3. Generates data useful for risk assessment
4. Adequate data available for specified uses
5. Robust and transferable
6. Time and cost-effective
7. Adequate animal welfare consideration (3Rs)

The ICCVAM test method evaluation process also was described.

Dr. Stokes then described the history of *in vitro* alternatives for ocular irritation. Numerous methods were developed in the 1980s-90s, and numerous validation studies were conducted

in the 1990s. In 1993, a workshop was held by the Interagency Regulatory Alternatives Group (IRAG) that evaluated several *in vitro* test methods as replacements for *in vivo* tests. None of the test methods evaluated were considered a valid replacement. However, test guidelines (e.g., EPA [1998] and the Globally Harmonized System [GHS; UN 2003] tiered testing strategy) were modified to allow for the use of *in vitro* test methods following future validation and acceptance. Some countries in the European Union (EU) will accept positive results for classification of R41 (risk of serious damage to the eye).

A recent European Commission (EC) Directive (EU 2004) regarding IRE, BCOP, ICE, and HET-CAM was described. The directive states that, "These tests are not yet validated, and therefore not included in Annex V." However, positive results can be used to consider a substance a severe irritant and R41 applied with no further testing. But, "where a negative result is obtained, an *in vivo* test should subsequently be required, as the *in vitro* tests have not been shown to adequately discriminate between eye irritants and non-irritants."

Dr. Stokes also described the background and history of the ICCVAM evaluation of *in vitro* ocular irritation assays. In August 2003, EPA announced plans to nominate *in vitro* ocular toxicity test methods for review by ICCVAM. Emphasis was placed on those test methods that may be able to identify severe irritants without animal testing. ICCVAM's Scientific Advisory Committee on Alternative Toxicological Methods (SACATM) unanimously recommended the methods as high priority for ICCVAM evaluation. In October 2003, EPA submitted a formal nomination of four ocular evaluation activities to NICEATM and ICCVAM. ICCVAM endorsed the four EPA nominations as high priority in January 2004. The highest priority was evaluation of *in vitro* screening methods for ocular corrosives/severe irritants. An OTWG was then established to coordinate the evaluation with NICEATM.

A *Federal Register* (*FR*) notice was published in March 2004 requesting public comment on the nominations, and data on chemicals evaluated by *in vitro* or *in vivo* ocular irritancy test methods. A second *FR* notice was published in April 2004 requesting nominations of scientific experts for an independent expert panel. Between April and October of 2004, the four BRDs on the BCOP, HET-CAM, ICE, and IRE test methods were prepared by NICEATM. In November 2004, a *FR* notice announced the dates of this meeting, the availability of the BRDs, and a request for public comments. A *FR* notice announcing the availability of additional data and analyses was published in December 2004.

Charge to the Expert Panel and Organization of the Review

Dr. Stokes explained the charge to the Expert Panel. The Panel was requested to evaluate, for each of the four test methods, the extent and adequacy that each of the applicable ICCVAM validation and acceptance criteria have been addressed, based on available information and data, or will be addressed in the proposed studies for the purpose of identifying ocular corrosives and severe irritants in a tiered testing strategy. The Panel was also asked to develop conclusions and recommendations on:

- Current usefulness and limitations of each of the four test methods for identifying ocular corrosives and severe irritants

- The test method protocol that should be used for future testing and validation studies
- Adequacy of proposed optimization and/or validation studies
- Adequacy of reference substances proposed for future validation studies

A tentative post-meeting timeline also was presented.

Acknowledgments

Dr. Stokes acknowledged the many individuals and organizations who helped with this review. These include the following invited test method experts: Menk Prinsen (TNO-CIVO Institutes, The Netherlands); Dr. Klaus Krauser (Abbott Laboratories, Abbott Park, Illinois, United States); Robert Guest (SafePharm Laboratories Ltd., Derby, United Kingdom); and Dr. John Harbell (Institute for *In Vitro* Sciences [IIVS]; Gaithersburg, Maryland, United States). An ICCVAM Working Group (OTWG) comprised of government scientists that is co-chaired by Drs. Karen Hamernik and Jill Merrill, worked with NICEATM to develop the questions that were addressed to the Panel. The OTWG also recommended experts to serve on the Panel and reviewed the BRDs for completeness. The OTWG will review the recommendations proposed by the Panel and develop draft ICCVAM recommendations. ICCVAM recommendations and the Panel's final report will be forwarded to the EPA and other Federal Agencies for consideration.

Welcome and Introduction to the Meeting by the ICCVAM Chair

Dr. Schechtman (Chair, ICCVAM; FDA) added his welcome to the Panel and the meeting attendees. He then briefly described the composition of the Panel, which was composed of scientists from Europe, Japan, Canada, and the United States, with expertise in toxicology, human and veterinary ophthalmology, biostatistics, pharmacology, anatomy and physiology, laboratory animal medicine, and pathology. Dr. Schechtman also discussed the importance of science-based expert panel recommendations. He noted that the advice of the Panel regarding the validation status of the *in vitro* ocular test methods, including their usefulness and limitations, will help guide:

- The formulation of ICCVAM/NICEATM recommendations regarding the validation status of the four *in vitro* ocular test methods of interest
- The conduct of any future studies that might be warranted that could help them satisfy ICCVAM's criteria for validation and acceptance of toxicological test methods and render any of these methods more acceptable for regulatory purposes
- Regulatory agencies on the use of data generated by these test methods that could help in their regulatory decision-making processes

Dr. Schechtman also briefly described regulation of food, drugs and cosmetics by the FDA. FDA is charged with protecting American consumers through the Federal, Food, Drug, and Cosmetic Act and its amendments. He recognized that today's regulations grew out of a series of health-related tragedies that caught the public's attention beginning in the early 1900's. The beginning of ocular irritancy testing in the United States also was described.

Overview of Current Ocular Toxicity Regulatory Testing Procedures

Ms. Debbie McCall (EPA) provided an overview of the current U.S. and European statutes and regulations that require ocular irritation testing. In the United States, the EPA, CPSC, FDA, and OSHA have authority to require testing of particular chemicals and products to determine their ocular irritation potential, as a result of various statutes and regulations (e.g., Toxic Substances Control Act [TSCA] and Federal Hazardous Substances Act [FHSA]). Testing guidelines are in place to aid the regulated community in meeting these testing requirements.

Ms. McCall then discussed the basic procedures for conducting the *in vivo* rabbit eye test. Relevant Testing Guidelines were also described and compared, to include those of the EPA, the EU, the FHSA, and the Organisation for Economic Co-operation and Development (OECD). All four test guidelines are based on the original method of Draize et al. (1944). FHSA requires the greatest number of animals in an initial test (six). EPA, EU, and OECD recommend up to three animals in an initial test, with the possibility of using only one animal for classifying a corrosive substance. The four test guidelines permit use of anesthetics, generally when pain is anticipated. EPA, EU, and OECD require studies to be carried out to 21 days to evaluate reversible/irreversible effects, while FHSA only requires observations out to three days. Irrigation of the eyes is allowed in all four test guidelines after 24 hours; EU and OECD allow for irrigation at one hour for solid substances.

All four ocular damage regulatory guidelines use the same rabbit eye scoring system. The eye of a treated rabbit is subjectively evaluated using the Draize method for three endpoints: corneal opacity, iris effects, and conjunctival effects. The scores for each of these endpoints were described in detail.

Ocular Hazard Regulatory Testing Requirements and Classification Schemes

Ms. Debbie McCall also provided an overview of ocular toxicity classification definitions and criteria among regulatory hazard classification systems (EPA, EU, GHS, FHSA). All current ocular toxicity classification systems are based on the Draize rabbit eye test method (Draize et al. 1944) and scoring system; however, the classification definitions and criteria vary among the systems.

The EPA classification system (1996) was described first. At least three animals per test are usually required for classification, with a one-animal screen permitted. The maximum score in any animal is used for classification of a substance (i.e., the most severe response is used). The EPA classifies substances into four ocular irritant categories, ranging from I to IV. Category I substances are defined as corrosive or severe irritants, while classification from II to IV is based on decreasing irritation severity, as well as the time required for irritation to clear. EPA labeling signal words, statements and protective equipment/actions for each of the four categories were described. For example, the signal word for Category I is Danger, and the statements required for labeling are “Corrosive. Causes irreversible eye damage. Do not get in eyes or on clothing.”

In the EU classification system (2001), at least three animals are usually required for classification, with a one-animal screen permitted for corrosive effects. There are two possibilities for classification:

- If a study includes > 3 animals, mean study values (each endpoint averaged over days 1-3 of the study for all animals) are used
- If a study includes 3 animals, individual animal mean values (each endpoint averaged over days 1-3) are used

Hazard classification of ocular irritation in the EU system corresponds to two risk phrases: 1) R36 denotes "irritating to eyes"; 2) R41 denotes "risk of serious damage to the eyes."

Ms. McCall proceeded to describe the GHS classification system (UN 2003). Classification is based on severity of effect and reversibility of the effect. The GHS includes two harmonized categories, one for irreversible effects on the eye/serious damage to the eye (Category 1), and one for reversible effects on the eye (Category 2). Reversible effects are further subclassified, based on the duration of persistence, as Category 2A (irritating to eyes; reverses in 21 days), and Category 2B (mildly irritating to eyes; reverses within seven days). GHS labeling symbols, signal words and caution statements also were described.

The FHSA classification system (CPSC 1995) uses at least six animals per test. In this system, there are three categories: corrosive, irritant, or nonirritant. A classification of corrosive is used if one or more animals exhibit destruction or irreversible alterations at the site of contact. For irritants, the maximum score in any animal on any day is used for classification. The irritant classification depends on the incidence of test animals exhibiting a positive ocular response. Depending on the number of animals with positive scores in a study, additional testing may be carried out.

Ms. McCall compared and contrasted the four systems. The EPA, EU, and GHS systems allow for classification of corrosive based on a one-animal screen. If the initial animal indicates corrosivity, no additional testing is required. Classification according to EPA and FHSA is based on the most severe lesion in any animal on any day. However, the EU and GHS systems take into account the most severe mean scores over days 1-3, in addition to persistent lesions. All four systems have only one classification for ocular corrosives/severe irritants. However, there are different numbers of classifications for nonsevere irritants:

- EPA (Category II, III, or IV)
- EU (R36)
- FHSA (Irritant)
- GHS (Category 2A or 2B)

Organization of the Panel Review

During the course of the two-day meeting, the Panel addressed a detailed list of questions concerning the completeness of each BRD and the performance of each test method evaluated. The Expert Panel was subdivided into four groups (one group per test method).

Each subgruop was responsible for addressing the questions for the relevant BRD, and drafting responses for consideration by the entire Panel.

Prior to the presentations and discussions by each of the four groups, an invited test method expert presented information on the test method protocol for which he had expertise. A NICEATM staff member then provided a brief summary of the information contained in the test method specific BRD, including accuracy and reliability analyses for the test method, and proposed optimization and validation studies for the test method.

Each Panel group presented its draft responses for each of the questions assigned for the particular test method BRD. After each presentation, the entire Panel discussed the draft positions and offered additional comments and suggestions. The Chairman summarized the discussion for each question and sought consensus from the Panel on the topic. Public comments were accepted following the Panel's discussion of each BRD.

I. ISOLATED CHICKEN EYE (ICE) TEST METHOD EVALUATION

Primary Reviewers: Drs. Robert Scala, Roger Beuerman, June Bradlaw, Wallace Hayes, Robert Peiffer, Nancy Flournoy

Note: Due to a family emergency, Dr. Flournoy was unable to attend the Panel meeting. However, her comments and suggestions were included in the ICE Panel Report.

Overview of the ICE Test Method Procedure

Mr. Menk Prinsen (invited expert from TNO) provided an overview of the ICE test method. Included in his presentation was a description how the ICE test is conducted, and how the data are used to predict the ocular irritancy classification of a test substance. Mr. Prinsen indicated that the ICE test has been used at TNO since 1992 for eye irritation testing.

Mr. Prinsen indicated that the ICE test was adapted from the Isolated Rabbit Eye (IRE) test. He stated that he looked at the possibilities to use slaughterhouse animals as a replacement as an eye donor and we looked at several species, the bovine, the pig and the chicken. He described that the bovine and the pig for us were less suitable because the cornea was too thick, while the structure of the chicken cornea appeared comparable to that of the rabbit. The process of obtaining the chicken eyes was then described, along with the experimental setup, including the 11-chamber superfusion apparatus and the saline drip system. Mr. Prinsen estimated that the total cost of the experimental setup would be approximately $15,000 U.S. He then described how the chicken heads were transported from the slaughterhouse to the laboratory in humidified boxes at ambient temperature, and that the eyelids close spontaneously after death, providing a protective barrier for the corneal surface. He stated that the eyes could be dissected very quickly (approximately 10 seconds each), and placed in the superfusion apparatus under a saline drip at 32˚C for equilibration for 45 minutes.

Mr. Prinsen then detailed the experimental procedure, beginning with the pre-test measurements recorded to ensure the adequate integrity of each test eye. Mr. Prinsen next detailed the method of dosing, in which 30 µL of liquid or 30 mg of solid is applied to each of three eyes per test substance, with one eye remaining untreated as a control eye. He explained that the rationale for this quantity was based on the relative size of the chicken and rabbit, where the chicken eye is roughly one-third the size of the rabbit eye (which is dosed with 100 µL of liquid or 100 mg of solid). He then described how corneal opacity, corneal thickness, and fluorescein retention are measured with a slit-lamp microscope. Representative photographs of each endpoint were provided, along with video images of actual dosing and resulting opacity formation.

Mr. Prinsen then outlined the decision criteria used to assign an ocular irritancy classification using a categorization scheme for each endpoint. He described that a logical subdivision of the combined categories was used to derive an overall irritancy classification, and how these categories could be applied to yield an EU classification. He also noted that in addition to the combination of the three categories, an immediate corneal opacity score of three or

higher, a corneal opacity score of four throughout the test period, or if there is severe loosening of epithelium, would be criteria for assigning a severe irritant classification. Mr. Prinsen also indicated that histopathological effects may also be used to assign an irritancy classification, but the precise decision criteria for this endpoint were not provided.

After Mr. Prinsen completed his presentation, Dr. Scala invited the Panel to ask him questions on the procedural elements of the ICE test, a full record of which is available in the meeting transcript.

Summary of the ICE Test Method BRD

Dr. David Allen (NICEATM) presented a summary of the BRD for the ICE test method. Dr. Allen detailed that the primary data sources that were used in evaluating the performance of the ICE test were extracted from three publications (Prinsen and Koëter 1993, Balls et al. 1995, and Prinsen 1996). The number of substances evaluated for the EU (n=121) was the largest because the EU classification was given for certain substances for which individual animal data were not provided. Therefore, classification based on the GHS (n=92) and EPA (n=90) classification system was not feasible for all 121 substances. Fifteen chemical classes were tested; the most frequently being alcohols, acids, and surfactants. Also, fourteen product classes were tested, the most frequent being chemical/pharmaceutical intermediates, herbicides/pesticides, industrial chemicals, and soaps/surfactants/detergents.

Dr. Allen described the fact that the major ICE test method protocol variation among the three studies was the number of eyes tested per substance. Originally, five eyes per test substance were included, but that number was later reduced to three, reportedly with no effect on test method performance.

Dr. Allen also described that accuracy statistics were calculated for each test method protocol by report and where appropriate, classifications were pooled into one classification per substance as well as using individual studies where a balanced design existed. Overall accuracy was reported as 82% to 85% (depending on the classification scheme used). Likewise, the false positive rate was between 8% to 10%, while the false negative rate was between 30% to 40%.

Dr. Allen also presented the substances that were used to evaluate test method reliability. There were no substances tested in intralaboratory studies, but the Balls et al. (1995) study provided an interlaboratory reproducibility analysis (n=59 substances tested in four laboratories). This analysis was performed both quantitatively (coefficient of variation analysis) and qualitatively (the extent of agreement among laboratories). For the qualitative analysis, roughly 75 percent of the time, all four laboratories got the same outcome, while up to at least three out of four laboratories got the same answer approximately 90% of the time. For the quantitative analysis, median coefficient of variation (CV) values of approximately 35% was noted for all endpoints except corneal swelling, which was approximately 75%. Dr. Allen speculated that this discrepancy may be due to the use of different thickness measurements among the four laboratories, which could result in variability.

Dr. Allen closed by summarizing the draft test method proposals that were presented in the BRD. A proposed test method version was identified which evaluates corneal opacity, corneal swelling, fluorescein retention and morphological effects, along with a proposed standardized protocol based on a method of TNO. The only significant difference in the NICEATM-proposed protocol and the protocol used by TNO is the inclusion of additional eyes for negative controls where three eyes per negative control are proposed, as opposed to just one, in addition to concurrent positive controls and when appropriate, a benchmark control. Potential optimization studies that might enhance the performance of the test method were also identified. These included: 1) a retrospective analysis of the decision criteria that are used to identify corrosives; 2) an evaluation of the potential causes of the lower level of interlaboratory reproducibility for the corneal swelling endpoint; 3) additional evaluation of possible increased interlaboratory variability that was identified for particular chemical classes, albeit with relatively small numbers such as alcohols, acetates and esters and cationic surfactants; 4) determining the feasibility of introducing a quantitative measurement of corneal opacity; and 5) determining the utility of histopathology and when exactly it should be included. Once optimized, the protocol should undergo additional validation studies to further characterize the accuracy and reliability of the optimized method and that is a summary of the ICE test method BRD.

A discussion ensued regarding the use of CVs in the context of interlaboratory variability. Dr. Lovell stated that in this evaluation, caution should be observed in how these CV measures are interpreted. He noted that there will likely be significant variability in the CV values based on the range of endpoint scores. For example, you could easily get a zero value for the fluorescein retention value since that endpoint score ranges from 0 to 3. Conversely, the corneal swelling value has a much larger potential range, and thus would be expected to have greater variability.

Proposed Panel Recommendations for the ICE Test method

1.0 ICE Test Method Rationale

Dr. Scala presented the draft recommendations for the ICE test method for consideration and concurrence by the Panel. He noted that although the mechanistic basis of the ICE test is not known, this may not be of concern given the fact that correlation with irritancy classification was the predominant goal in this context for the assay. He also recognized that the anatomy and structure of the eyes of chickens, rabbits, and humans are different. He mentioned the differences between the ICE test and the *in vivo* rabbit test, and the endpoints that the ICE does not evaluate (i.e., conjunctival, iris effects, no assessment of reversibility, does not account for systemic effects). In a discussion among the Panel in a subsequent test method, they recommended that the ICE BRD should add discussion of cellular mechanisms of corrosion and severe irritation (e.g., necrosis, apoptosis) and relevance to *in vitro* testing, along with the role of responsive inflammatory cells in isolated rabbit eyes.

2.0 ICE Test Method Protocol Components

Dr. Scala continued by highlighting areas of concern in the protocol that were identified by the Group. These included: the potential for variability due to use of different depth measuring devices; temperature not being well controlled; the drip system, which appears difficult to control and results in removal of the tear film; the vertical position of the superfusion apparatus; randomization of the test eyes; length of exposure time; lack of divalent cations in the superfusion medium, and the number of test eyes per substance (n=3). Following a discussion, the Panel recommended that reference substances that are part of the performance standards developed for the validated test method should be identified.

3.0 Substances Used for Previous Validation Studies of the ICE Test Method

Dr. Scala then discussed the adequacy of the types and number of substances evaluated. He indicated that although only one of the studies evaluated used a coding system for the test substances, a lack of coding was not justification for excluding the remaining data.

4.0 In Vivo Reference Data Used for an Assessment of Test Method Accuracy

Dr. Scala then continued by noting that the interpretation of the *in vivo* results appeared correct. However, he pointed out that the regulatory classification methods may be less than adequate for use in evaluating or making distinctions between *in vitro* methods and their suitability for chemical product class evaluations. He then noted that original study records were not available for any of the reports evaluated, but an evaluation of the results could be made and the quality of the studies otherwise appears to be adequate. Dr. Scala then summarized the extent of GLP compliance of the *in vivo* studies, along with the need for future studies to be conducting according to GLPs. He did note that the Primary ICE Reviewers believed that lack of GLP compliance was not an adequate basis for excluding data from the evaluation and that future validation studies should be conducted under GLP compliance and original study records should be readily available. He then recognized that most human eye data was from accidental exposures. During Panel discussion, Dr. Kathy Stitzel raised the point that animal testing did not necessarily precede human testing and Dr. Wiley Chambers (FDA) confirmed that point. Dr. Chambers also pointed out that irritating compounds are tested in controlled clinical trials. Dr. Martin Stephens then raised the point that there needs to be greater effort to obtain and consider information on human topical ocular chemical injury. Dr. Roger Beuerman recommended the Alabama Ocular Injury Registry, and Dr. Donald Fox recommended Fraunfelder's registry of ocular injury. Dr. Stitzel also noted that since this evaluation only deals with severe ocular irritants/corrosives, qualitative data would likely suffice as reference data.

Dr. Scala then acknowledged the ongoing debate over the variability of the *in vivo* rabbit eye test. He indicated that the potential variability of the *in vivo* rabbit data has not been adequately discussed in the BRD, and that the evaluation of an alternative method is unavoidably biased by the selection of the *in vivo* data used in that evaluation. Subsequent to a discussion, the Panel recommended 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. Dr. Martin Stephens expressed a minority opinion that no animal testing was needed for this purpose, as the current database should be considered adequate.

5.0 ICE Test Method Data and Results

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

6.0 ICE Test Method Accuracy

Dr. Scala then discussed the ICE test method accuracy evaluation. The overall false positive rate (8-10%) was considered adequate, but the acceptability of the false negative rate (30-40%) was less evident since this would result in corrosives/severe irritants to be tested *in vivo* (according to the tiered testing strategy). Dr. Scala stated that a comprehensive accuracy assessment in the absence of suitable human data should take account of the variability in the Draize test itself, such as the analysis by Dr. Joe Haseman that was distributed prior to the meeting. Subsequent to discussion, the Panel recommended an assessment based on ranking of experimental data for severity for both the reference method and the *in vitro* test. A minority opinion was expressed by Drs. Stephens and Peter Theran that the use of the term "accuracy" in this context is not appropriate because the *in vitro* test may in fact be a more accurate estimate of the human response. For this reason, the term "concordance" should be used in favor of "accuracy" when comparing the *in vitro* test to the *in vivo* rabbit eye test.

7.0 ICE Test Method Reliability (Repeatability/Reproducibility)

Dr. Scala then stated that the selection rationale for the substances used to evaluate test method reliability was considered adequate, and only one study was used for this analysis. Test method reliability analyses and conclusions were considered sound and appropriate, and both qualitative and quantitative evaluations of interlaboratory variability were conducted appropriately. Dr. Scala noted that no intralaboratory repeatability and reproducibility were conducted because of a lack of appropriate information.

8.0 ICE Test Method Data Quality

Next, Dr. Scala indicated that, given the lack of original records, caution should be exercised when evaluating these data, but that the lack of original records should not be used as a rationale for excluding these data. However, any future validation studies should include coded test substances of known purity, from a common source and centrally distributed; appropriate controls; and be conducted in compliance with GLP guidelines. He then discussed data quality audits, which were not feasible given the absence of original data, and that a more complete retrospective evaluation would be possible of such data were made available.

9.0 Other Scientific Reports and Reviews

No concerns were raised regarding this section. However, Dr. Scala did suggest that personal contacts by agencies to which ICE data have been submitted may be the best method to expedite acquiring more data. Furthermore, if such data are not received, additional *in vivo* studies may be necessary to compile an adequate reference database

10.0 Animal Welfare Considerations

Dr. Scala noted that although there is no additional inflicting of pain or distress to the animal as a result of the testing procedures, because chickens do not come under the Animal Protection Act, there is still a need to ensure their humane treatment.

11.0 Practical Considerations

Dr. Scala then discussed the transferability of the test method, which does not appear to be a significant obstacle to its use. However, he did indicate that specifications for the custom-built superfusion apparatus must be readily available. Following discussions, the Panel recommended that a training video and other visual media on the technical aspects of the assay be produced and that training approaches in the application of this test method should be developed/implemented. The relative cost of the ICE test and the *in vivo* rabbit eye test were considered comparable, and the ICE test can be completed in much less time than the full *in vivo* test (extending out to 21 days). However, during discussions Drs. Stitzel and Itagaki raised the point that a corrosive or severe irritant may be detected within a few hours using a single rabbit, and thus the reduction in time afforded by the ICE would not always be applicable.

12.0 Proposed ICE Test Method Recommendations

Finally, Dr. Scala summarized the draft recommendations for the ICE test method. He stated that the ICE test method appears to be useful in the identification of ocular corrosives from severe irritants in a tiered testing strategy with the following limitations: alcohols tend to be over predicted; surfactants tend to be under predicted; and solids and insoluble substances may be problematic because they may not come in adequate contact with the corneal surface. He also highlighted that the reliability of the ICE test has not been adequately assessed.

A discussion ensued regarding the context of these statements and how they relate to the optimization/validation studies that have also been recommended. Dr. Stokes noted the difference between declaring a test method as useful versus saying that it has been fully validated. He stated that a test method may be used if it is considered to be useful, and once it has been through adequate validation, it will be mandated for routine use before an *in vivo* test. He continued that alternatives have to be considered by animal care and use committees (at least in the United States). Therefore, by stating that a test method can be used, then institutional animal care and use committees are going to have to ensure that it has been considered before a rabbit test is done. He concluded that this doesn't imply that a mandate

for their use will occur, but they need to be considered and some rationale will have to be provided why they are not being proposed for use now.

Notwithstanding the above conclusions, the Panel agreed on following recommendations: a formal evaluation of the optimum number of eyes per test substance should be conducted; a standardized scoring scheme for histopathology, along with identification of the appropriate circumstances to include in such an evaluation, should be developed; and reference photographs for all subjective endpoints should be provided. During discussions, the recommendation from the Dr. Edelhauser to install centering lights on the optical pachymeter to make it easier to take precise corneal thickness measurements was agreed to. Dr. Fox also recommended an evaluation of the impact of delayed use of chicken eyes on performance. Dr. Scala continued by identifying a number of optimization studies considered useful to enhance the performance of the ICE test. These included:

- optimizing the decision criteria to reduce the false negative rate while maintaining a low false positive rate
- determining the utility of rotating the superfusion apparatus to a horizontal position
- determining the utility of including divalent cations in the assay medium
- determining the optimum mechanism for handling differences in corneal swelling values for test substances from different laboratories

The Panel also recommended that reference substances should be identified that can be used as part of performance standards, and that NICEATM/ICCVAM should facilitate the development of a histopathology scoring system for corneal damage (with visual aids). Finally, the Panel recommended that any optimization and validation studies should use existing animal data, if available; that additional animal studies should only be conducted if important data gaps are identified; and that such studies should be carefully designed to maximize the amount of pathophysiological information obtained (e.g., wound healing). Dr. Stephens again expressed a minority opinion that there is sufficient data so that additional animal testing for this purpose is not warranted.

Panel Vote on the ICE Report

Dr. Scala concluded this discussion with a vote among the Panel members. He noted that everyone on the Panel, with the exception of Dr. Stephens, was in agreement with the conclusions and recommendations for the ICE test method. Dr. Stephens's dissenting opinions are noted in the relevant sections above.

Public Comment Session

1. Dr. Rodger Curren (IIVS)

Dr. Curren presented a public comment that dealt with the variability of the Draize eye test. He indicated that the question at hand is whether the *in vitro* tests are good enough to replace the *in vivo* test. He stated that the first step would be knowing as much as possible about what the performance of the test that is going to be replaced. He was therefore critical of the

BRD as having, "virtually a complete absence of discussion about the performance of the animal test that we were looking at." He questioned as to how judgments about how well the *in vitro* test performs can be made in the absence of information about the animal test. He continued by stating that given the variability of the animal test, a severe irritant outcome doesn't mean that that chemical is severe by some degree of a physical-chemical property that never changes, because if it is tested again, a different outcome may result.

Dr. Curren referenced what he considered to be the best animal data that he could find to bias it towards the best animal data (the Cosmetics, Toiletry, and Fragrance Association [CTFA] evaluation study of surfactants and surfactant-containing materials conducted in 1993-94). Citing an evaluation done by Dr. John Harbell, he noted that there is underprediction of the severe irritants among these substances of varying degrees. During a subsequent discussion, it is revealed that this analysis is included in the BCOP BRD as Appendix H so that it could remain intact, as it was not part of the performance analysis conducted in the BRD. Dr. Tice also noted during the discussion that Dr. Harbell's overall figure for the underprediction rate of the *in vivo* test (19%) is comparable to that of Dr. Haseman's estimate (8% to 18%), which also is near the value obtained using the Weil and Scala (1971) data (15%).

2. Sara Amundson (Doris Day Animal League)

Ms. Amundson provided general comments on the BRDs and their subsequent review by the Panel. She expressed concern that there seemed to be a tendency for the BRDs to lead the Panel in very specific directions with regard to recommendations and she hoped that was carefully considered as the Panel moves forward with their recommendations. She also noted that accuracy is not relevance and that relevance is validation, a concept that she stated has reached international agreement. She continued by indicating that the fact that this was left out of the direct information communicated to the Panel in the Executive Summary of each BRD was unconscionable.

Ms. Amundson then referenced S.C. Johnson's written comments that indicated that they have utilized the BCOP, one of the test methods under consideration during the meeting, for the past 15 years or so. She continued that they have made internal decisions and have submitted data to EPA for regulatory decisions. She noted that the potential outcome of the meeting was that the ability of companies like S.C. Johnson to continue to use a test method that they have been able to utilize both internally and externally for regulatory decision making purposes for a great number of years would be hindered.

She closed by advising the Panel to keep in mind that if there is confirmatory testing that is required under any sort of strategy or paradigm that is put forward, the concerns about regulatory and internal corporate decision-making may actually stunt the submission of test methods to the ICCVAM for consideration.

II. BOVINE CORNEAL OPACITY AND PERMEABILITY (BCOP) TEST METHOD EVALUATION

Primary reviewers: Drs. Kathy Stitzel (Group Chair), Ih Chu, Henry Edelhauser, Hiroshi Itagaki, Lionel Rubin, Scheffer Tseng, David Lovell

Overview of the BCOP Test Method Procedure

Dr. Harbell (Invited Expert, IIVS) began his presentation by acknowledging Dr. Pierre Gautheron and Dr. Joseph Sina of Merck Research for much of the initial work on the BCOP assay. He then described the corneal lesions associated with eye irritation *in vivo*. There are four major types of corneal lesions:

- Epithelial cell loss -- partial or full thickness loss over some fraction of the cornea
- Stromal swelling -- results from either a change in hydration state, which is reversible, or protein denaturation, which is poorly reversible
- Death of keratocytes -- depth of injury or loss of these cells is associated with increased severity of the lesion and initiation of inflammation
- Endothelial cell loss -- these cells do not regenerate in humans

Dr. Harbell also described the common modes of chemical action in eye irritation. Membrane lysis can occur when surface-active agents solubilize membrane lipids or when organic solvents extract lipids. A second mode of chemical action involves protein coagulation or denaturation, which can result from exposure to acids and certain solvents. Saponification can result from alkalis and is often progressive. Lastly, alkylation or oxidative damage to macromolecules can result from reactive materials such as bleaches and peroxides.

Dr. Harbell then explained the "depth of injury" model first introduced by Drs. Maurer and Jester, who proposed that depth of injury in the cornea is predictive of the degree and duration of injury. The conclusion of these authors is that the changes in the early post-exposure period (i.e., the first 3 hours, or so) are predictive of the degree and duration of injury *in vivo*. This model provides a link between the *in vivo* situation, whether human or rabbit, and the *in vitro* testing situation.

Using the *ex vivo* cornea, as done in the BCOP assay, provides appropriate cellular and structural targets for eye irritation testing, allows determination of depth of injury, and responds to the various modes of action of irritants on the cornea. Factors that impact exposure in the eye also were described. Binding or trapping of substances in the eye will increase irritation, while vaporization, dilution, and flushing will decrease irritation. Rapid binding and/or penetration will favor increased irritation potential.

Dr. Harbell then described the procedures involved in conducting the BCOP assay. A local slaughterhouse supplies the bovine eyes. The eyes are transported in a bucket in Hanks' balanced salt solution (HBSS), and the bucket is maintained over ice. The time between first

slaughter and the time of arrival at the laboratory is between four and five hours. After arrival at the laboratory, the cornea is inspected for defects, scratches and opacity; the quality of the eyes is very important. The cornea is then excised. Corneas are held in HBSS, then are mounted very carefully on a holder. The holder has an O-ring side and a non-O-ring side. The endothelium is placed on the O-ring; it is critical that once placed on the O-ring, the cornea is not moved. The other half of the corneal holder is placed on top of the cornea, and the two chambers are screwed together. The chambers are filled with medium, and the system is allowed to equilibrate. Opacity is then measured quantitatively with an opacitometer. Sodium fluorescein is used to measure damage to the epithelium. If histology is conducted on the corneas, the corneas are fixed after the sodium fluorescein evaluation is completed.

One of the key features of the BCOP assay is complete control over exposure conditions. One can control the exposure concentration as well as the exposure time at the specified concentration, and ensure that exposure is over the whole corneal surface. Control over the post-exposure (expression) period is another important feature.

The *in vitro* score is an algorithm developed by Merck for pharmaceutical intermediates. It uses the opacity score, which ranges from 0 to 700, plus the net optical density (OD) of fluorescein multiplied by 15 (i.e., *in vitro* score = opacity + 15 x OD).

Certain chemicals do not induce direct opacity (e.g., anionic and nonionic surfactants) and so only the permeability score is used. Positive controls are always used at IIVS. Benchmark materials are used whenever possible.

The protocols for liquid and solid test substances were also described. Liquids are normally tested at 100%, while solids are tested at 200 mg/mL. The exposure period for liquids is normally 10 minutes, while that for solids is 4 hours. After the liquid is rinsed off the cornea there is a 2-hour "post-exposure" period, but there is no post-exposure period for solids.

Dr. Harbell then explained that exposure and post-exposure times can be modified to address certain chemical classes or expected consumer exposure scenarios. Reactive chemistries (e.g., peroxides) require extended post-exposure incubation because *in vivo* there is a delayed response. Something with a delayed onset, where apoptosis or delayed necrosis of keratocytes is a factor, requires a longer post-exposure period and histology.

Experience with the European Commission/Home Office (EC/HO) chemicals has shown that when the mode of action is not known, the most conservative approach is to address the reactive chemistry mode by using extended post-exposure incubation and histology. The more rapid changes will also be detected with this approach.

After Dr. Harbell completed his presentation, Dr. Scala invited the Panel to ask him questions on the procedural elements of the BCOP assay.

Summary of the BCOP Test Method BRD

Dr. Neepa Choksi (NICEATM) first discussed the current U.S. regulatory status of BCOP. ICCVAM agencies were surveyed by NICEATM, and the EPA and FDA indicated that BCOP data have been submitted to these agencies for consideration. She proceeded to describe the primary data sources for the BCOP BRD. Nine studies were used for the accuracy and/or reliability analyses in the BRD. In addition to these nine reports, there were 31 studies (discussed in Section 9) that were not used in the BRD analyses for a variety of reasons, including the lack of appropriate *in vivo* data or quantitative *in vitro* data.

The database used for the accuracy analyses included 166 different substances or formulations, representing 15 chemical classes and 20 product classes. The test method protocols used to generate BCOP data were similar to each other, but not identical. Some of the protocol variations include different storage conditions of the bovine eyes during transport, use of positive controls, and analysis of the resulting data.

Dr. Choksi then described the different BCOP data analysis methods. The most commonly used analysis method is the *In Vitro* Irritancy Score (IVIS), which is equal to the opacity value plus 15 times the optical density value. An IVIS >55.1 is considered a severe irritant. A few laboratories use the endpoint (opacity or permeability) with the highest score. One of the studies in the BRD analyzed permeability data only for substances that produce significant permeability without appreciable opacity. The distribution of BCOP tests among analysis methods was briefly described.

Accuracy analyses were performed to determine the ability of BCOP to correctly identify ocular corrosives and severe irritants as defined by GHS (Category 1), EPA (Category I), and EU (R41). Accuracy statistics were calculated for:

- each BCOP study that had acceptable *in vitro* and *in vivo* data
  - by test substance
  - by test
- pooled data from studies with similar protocols

In addition, false negative and false positive rates were calculated for different chemical classes and available physicochemical properties (liquid/solid).

The accuracy using pooled studies was relatively good for all three classification systems, ranging from 77% to 80%. False positive rates for the pooled data ranged from 17% to 23%, while false negative rates ranged from 22.5% to 27% for the three classification systems. The analyses by chemical class showed that the false negative (14%) and false positive (5%) rates were rather good for surfactant-containing formulations (n=34). Liquids had lower false negative (18%) and false positive (21%) rates, than the corresponding rates (33% and 29%) for solids. A major limitation of the BCOP accuracy analysis is that for a majority of the chemical classes (63%; 20/32), only a small (two or less) number of substances were tested. Another limitation is the limited information on the physicochemical properties for some test substances.

For the reliability analyses, BCOP data were available to evaluate intralaboratory repeatability and reproducibility, as well as interlaboratory reproducibility. For the intralaboratory analyses, quantitative CV analyses were conducted. For the interlaboratory analyses, a CV analysis and a qualitative analysis that evaluated extent of agreement between testing laboratories were performed. Intralaboratory repeatability was evaluated for three studies (Dr. Sina's submission, Swanson et al. 1995, Southee 1998) by calculating the CV of the IVIS obtained for replicate corneas exposed to the same test substance. For substances predicted as severe irritants, the mean %CV ranged from 8.2 to 11.1 for the three studies. Intralaboratory reproducibility was evaluated for two studies (Gettings et al. 1996 and Southee 1988), which tested substances in two or more independent experiments (trials). For the Gettings study, which evaluated permeability only for 25 substances in three different trials, the mean and median %CV for the permeability value was 33.4 and 29, respectively. In the Southee study, which evaluated 16 substances in two or more trials in three laboratories, the mean %CV for the IVIS ranged from 12.6 to 14.8 for the three laboratories. As far as the interlaboratory reproducibility, the classification agreement among laboratories was very good. The interlaboratory %CV values for IVIS were also good for substances predicted as severe, but were much higher when the datasets included moderate to mild irritants, which have lower scores and, thus, greater "noise" in the data. No limitations were identified for the BCOP reliability analyses.

The draft BRD on BCOP included a number of proposals. A recommended BCOP version was identified which evaluates opacity and permeability, as well as histology on a case-by-case basis. A standardized protocol was proposed for the recommended version of the BCOP test method; this protocol is based on the method used by the Institute for *In Vitro* Sciences. The only significant difference between the two protocols is that the recommended protocol in the BRD lacks the detailed histology procedures provided in a separate IIVS protocol on histology for the BCOP assay. The decision criteria recommended are those previously described by Gautheron et al. (1994). Proposed optimization studies include a retrospective analysis of decision criteria used to identify corrosives and severe irritants, an evaluation of possible increased interlaboratory variability for specific chemical classes appearing more variable (e.g., alcohols), an evaluation of reduced exposure times for alcohols and possibly other volatile solvents, and determination of the utility of histopathology and when it should be used in the BCOP assay. Once optimized, additional studies are recommended to further assess the accuracy and reliability of BCOP, so that the applicability domain can be better defined and data gaps are filled.

Proposed Panel Recommendations for the BCOP Test Method

1.0 BCOP Test Method Rationale

The Panel agreed that use of living corneal tissue is a good model. Opacity is an important endpoint, but the proposed protocol doesn't differentiate the different mechanisms of opacity. Permeability measures integrity of the cornea and adds important information on degree of injury. Limitations of the BCOP test method are that the method evaluates only corneal effects, it may not identify materials that cause serious corneal injury without changes to opacity or permeability, it doesn't evaluate damage to limbal stem cells, and it doesn't model

protective mechanisms (e.g., blinking, tearing) that affect the outcome of *in vivo* studies. The Panel discussed the importance of the limbus for evaluating damage to the human eye; it is part of the conjunctiva and the most important issue when classifying human eye damage.

The endpoint of corneal opacity is measured in both BCOP and the accepted *in vivo* method. However, BCOP does not measure changes in iris and conjunctiva, and systemic toxicity is not identified by BCOP. Although reversibility is not evaluated in BCOP *per se*, initial depth of injury *in vitro* may be useful to evaluate potential long-term effects. The Panel, therefore, recommended that the BCOP BRD include a discussion of the work of Maurer and Jester, which provides evidence that initial changes can predict long-term effects. The BRD should also discuss the human clinical experience with injury scales that are used to predict long-term effects from immediate injury.

2.0 BCOP Test Method Protocol Components

The Panel agreed with the protocol components in the BRD, with the exception of the following items:

Eyes

- Discourage use of antibiotics since they are not effective at 4º C, could potentially increase permeability of the epithelial cells, and possibly cause drug-compound interactions; however, to inhibit bacterial growth the eyes must be kept cold
- The recommended storage time of 4-5 hours may be too restrictive
- Bovine Spongiform Encephalopathy (BSE) is a risk, thus, appropriate precautions should be taken

Solvent for preparing solutions

- Use 0.9% NaCl, *not* sterile water
- Osmolarity and pH of solutions should be known

Corneal culture medium

- Minimum Essential Medium with Fetal Bovine Serum is not necessary
- Balanced salt solutions should be acceptable

Optimize corneal holder

- It should clamp on the sclera and not the cornea
- Holder should maintain curvature of cornea

Exposure

- Optimize exposure duration for 'volatile solvents'
- Exposure method for solids is problematic

Rinsing procedures

- Recommend optimizing these procedures especially for viscous materials and solids

Histopathology

- Must be added unless the substance is from a class of materials known to be accurately predicted using only opacity and permeability in the BCOP assay
- A grading system for histopathology is needed

Reference substances

- Identification of reference substances that are part of the performance standards developed for the validated test method

Controls Needed

- Positive, negative and benchmark controls are needed
- Each laboratory must establish acceptable ranges

Reexamine Prediction Model

- Is a calculated score advisable/necessary?
- Optimize to identify severe irritants
- The BRD should identify the decision criteria for identifying ocular corrosives and severe irritants and discuss rationale for development

Additionally, the Panel discussed some issues surrounding BSE. It takes a full day after slaughter to determine whether cattle have BSE, but the eyes must be used before then. Another point is that calves up to 16 months do not have BSE, so some companies are trying to use calves eyes instead of eyes from mature animals. It was recommended that NICEATM obtain the BCOP data obtained using calf eyes for further evaluation.

3.0 Substances Used for Previous Validation Studies of the BCOP Test Method

The Panel agreed that the number and classes of substances tested in previous validation studies were acceptable. However, materials known to be severe eye irritants in humans should be confirmed to be severe in BCOP. Since available data indicate alcohols, ketones, and solids are problematic in BCOP, better chemical characterization and physicochemical data on all the test substances are needed.

The Panel considers coding procedures to be important, since data quality could be affected if they are not used. The coding procedures used in the previous validation studies were considered adequate.

4.0 In Vivo Reference Data Used for an Assessment of Test Method Accuracy

The Panel agreed that the *in vivo* rabbit eye test method protocols used to generate reference data in the studies cited in the BRD were appropriate. However, the use of the three regulatory classification systems to evaluate *in vitro* methods was questioned by the Panel. Regarding the data quality of the *in vivo* studies, the lack of original study records was a concern of the Panel, but was not considered serious enough to prevent use of the data. Also,

the BRD should include more information on the extent of GLP compliance of the *in vivo* studies.

With respect to the availability of relevant human ocular toxicity information, the Panel recommended confirming that current ocular hazard classification schemes by examining Poison Control Center databases, Department of Labor data, and by reviewing published case reports. Also, the Panel stated greater effort should be put forth to obtain and consider information on human ocular injury from chemical exposures.

Regarding the accuracy and reliability of the *in vivo* rabbit eye test, the Panel agreed that the potential variability of the rabbit eye data was not adequately discussed in the BRD. Therefore, it was recommended that the BRD be modified to include discussion of several publications that address this issue, such as Weil and Scala (1971). An effort should be made to confirm *in vivo* classifications using other data sources such as the Registry of Toxic Effects of Chemical Substances or the International Uniform Chemical Information Database. Any optimization and validation studies should use existing animal data, if available. A majority of the Panel agreed that 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). However, Drs. Stephens and Theran had a minority opinion that no additional animal testing should be conducted for this purpose.

5.0 BCOP Test Method Data and Results

The Panel agreed with the BRD assessment of the information presented in Section 5.0 of the document.

6.0 BCOP Test Method Accuracy

With respect to the accuracy evaluation of the test method for identifying ocular corrosives and severe irritants as defined by the EPA (1996), the EU (2001), and the GHS (UN 2003), the Panel agreed that the currently used BCOP assay, with the addition of histology, is acceptable to assess the ability of materials to cause corrosive or severe injury to the eye as part of the screening strategy described in the BRD. However, based on the data presented, the assessment of alcohols, ketones, and solids with the protocol as written is problematic.

The Panel further discussed that alcohols were overclassified in three (BCOP, ICE, IRE) of the four test methods under review. In the BCOP test method, the 10-minute exposure protocol for liquids does not appear suitable for alcohols, but it was mentioned that a 3-minute exposure protocol might be more suitable. The Panel recommended that it would be useful to do a retrospective evaluation of any data obtained from the 3-minute exposure protocol.

The Panel agreed that the accuracy parameters must indicate that the values are a concordance comparison with the results of a single rabbit eye test. Because there is potentially a 10-20% misclassification rate for severe eye irritants in the rabbit eye test, the

Panel discussed the need to correct the performance statistics of BCOP for the Draize test variability. The misclassification in the *in vivo* test method would affect the false negative and false positive rates of the BCOP assay, and must somehow be accounted for in the performance statistics.

Regarding the strengths and limitations of the test method, the Panel stated that the effect of colored substances was not discussed in the BRD. Also, to better determine if certain physicochemical properties are problematic in the test method, consideration should be given to exploring physicochemical effects by using a structure activity or structure property relationship program.

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

7.0 BCOP Test Method Reliability (Repeatability/Reproducibility)

The Panel agreed with the BRD assessment of the selection rationale for the substances used to evaluate test method reliability. Regarding the analyses and conclusions regarding intralaboratory repeatability and intra- and inter-laboratory reproducibility, the Panel agreed that the data from existing studies was extensively reviewed and considered in the BRD. The data indicated acceptable levels of intra- and inter-laboratory variability. However, the Panel stated that the use of CVs should be used with care with this data. Optimization of the protocol may decrease variability.

Positive control data were presented in the BRD; however, negative control data were not included.

With respect to the effect minor protocol changes might have on the recommended test method protocol and the transferability of the test method, the Panel stated that the data indicate the test method is transferable. At what point minor protocol changes will be sufficiently significant to require further validation could not be determined with the information provided.

8.0 BCOP Test Method Data Quality

The Panel agreed that coding of chemicals should be used for all subsequent validation studies. While spot checks of data not part of multilaboratory validation studies could be conducted, the Panel believes this is not necessary. The lack of original notebook data was of some concern to the Panel but not sufficient to remove the data from consideration. The Panel noted that recent information indicates that raw data may be available for many, if not all, of the studies in accuracy evaluation of the BRD.

9.0 Other Scientific Reports and Reviews

The Panel agreed that the BRD adequately identified relevant data from other published or unpublished studies. The BRD also adequately summarized the conclusions published in

independent peer reviewed reports or other independent scientific reviews. It is possible that more data could be obtained by working with trade associations, but much of the data in the BRD is from these types of efforts, so whether more data could be obtained is unclear.

10.0 Animal Welfare Considerations (Refinement, Reduction, Replacement)

The Panel agreed that the BCOP will reduce the numbers of animals exposed to severe irritants. Also, the BCOP will aid in classifying some substances as severe without further animals tests.

11.0 Practical Considerations

The Panel agreed that the BRD adequately addresses the facilities, major fixed equipment, and availability of other supplies needed to conduct the BCOP test method. The required level of training and expertise to conduct BCOP were also adequately considered. However, the Panel thought the description of training of technicians for the *in vivo* test may be incorrect – proficiency in the *in vivo* test is demonstrated the same way as for BCOP.

The Panel recommended development of a training video and other visual media on the technical aspects of the assay. Also, training approaches in the application of this test method should be developed and implemented.

The discussion of the test method cost in the BRD should be modified to reflect the written public comments submitted by S.C. Johnson in December 2004.

The Panel noted that for very corrosive substances and some severe irritants, an *in vivo* evaluation may be completed within 4 hours in the rabbit eye test. Thus, it is not always true that *in vivo* evaluations would be extended to 21 days.

12.0 Proposed BCOP Test Method Recommendations

Regarding the recommended version of the BCOP test method, the Panel suggested confirming with several active laboratories that the proposed changes are workable.

Regarding the Panel's conclusions on the recommended standardized BCOP test method protocol, the Panel discussed at length whether the BCOP assay can be considered "valid", "validated", "acceptable", "useful" or whether the "validation criteria have been met" to identify ocular corrosives and severe irritants. However, even after the lengthy discussion, the Panel was uncertain what terminology to use to describe the conclusions reached about the recommended standardized BCOP test method protocol. The Panel Chair ultimately decided that this section of the BCOP report (Section 12.2) required a Panel vote.

A majority of the Panel agreed with the following conclusions regarding the recommended standardized BCOP test method protocol:

- For the purpose of detecting severe eye irritants in the testing scheme outlined in the BRD, the BCOP test method presented is useful in identifying ocular

corrosives and severe irritants, as described in the BRD, with the exception of:
  - Alcohols, ketones, and solids
  - Histopathological examination must be added, unless the substance is from a class of materials known to be accurately predicted using only opacity and permeability in the BCOP assay
- There is a need to confirm that the BCOP test identifies substances known to cause serious eye injury in humans
- Negative, positive, and benchmark (where possible) controls should be included
- Eyes from young adult cattle should be used
- Users should be aware of the risk of BSE and other zoonoses and use proper precautions
- 0.9% sodium chloride should be used as the standard diluent and rinse
- the osmolarity and pH of test solutions should be determined

Although Dr. Freeman abstained from voting on Section 12.2 since he believed the discussion had not been satisfactorily resolved due to time constraints, he agreed to provide a written abstaining opinion. Drs. Stephens and Theran were opposed to the language presented for Section 12.2 because they felt there was undue pressure on the group to "back off" the issues; they agreed to prepare a written dissenting opinion.

With respect to the recommended BCOP optimization studies in the BRD, the Panel made the following suggestions as future improvements to the test method:

- Using the larger holder designed by Ubels
- Reexamining the calculated total score
- Optimize media used to bathe the eyes
- Optimize rinsing procedures
- Consider use of younger animals
- Discourage the use of antibiotics

Optimization studies will be necessary to ensure any changes to the protocol will decrease the variability of the test method.

With respect to the recommended BCOP validation studies in the BRD, the Panel believes that validation studies are not necessary except, possibly, for solids, alcohols, and ketones. It is possible that submission of additional historical data for these types of substances may be sufficient. The Panel also stated that validation is not required for the addition of histopathology or changes in scoring system.

The Panel made one additional comment with respect to the BCOP test method. It was suggested that consideration be given to use of porcine eyes, since the porcine eye is a better model for human eyes than the bovine eye.

Panel Vote on the BCOP Report

The Panel Chair asked the Panel members to indicate whether they accepted the revised document that had been presented, with all the revisions that had been placed in it. A show of hands indicated that all Panel members, except for Drs. Freeman, Theran, and Stephens, accepted the BCOP report. Drs. Theran and Stephens stated that they voted no because of their opinion that animals should not be used for future validation studies, and because of their minority opinion on Section 12.2. Dr. Freeman stated that he voted no because he thought "the question at hand [in Section 12.2] was the most important question that we were asked to address over the two days that we have been here … if we are not taking a position to state clearly that the criteria have been met, this was so close that we should have specified what criteria were not met." Drs. Freeman, Theran, and Stephens indicated they agreed with the Panel on all other parts of the BCOP report.

Public Comment Session

1. Dr. Sherry Ward (Physicians Committee for Responsible Medicine)

Dr. Sherry Ward prefaced her comments by stating that she is a member of an animal advocacy organization, but is also a scientist with more than seven years of experience working in an industrial *in vitro* toxicology laboratory, where she spent significant time working on the validation and development of *in vitro* ocular toxicology test methods.

One important point she thought was missing from the BRDs is the importance of these methods to industry. If the methods are approved as screening assays, this will give companies the flexibility they need to choose the proper method or methods that are compatible with the testing of their products.

Another issue that she thought could have been better clarified in the BRDs is the potential to optimize *in vitro* decision criteria to look at how to reduce over- or under-prediction of the methods. It would have been helpful to have some of this analysis in the BRDs because this may have helped the Ocular Expert Panel to make a better decision on whether any of the particular methods should have been recommended for validation at this time with the ICCVAM. Without this information, it is really hard to make that decision.

Additionally, in the United States, a method has to be validated before it can be considered for acceptance by regulatory agencies. That is one of the ICCVAM regulatory-acceptance criteria. Inclusion in the Institutional Animal Care and Use Committee (IACUC) reviews due to this designation would be helpful, but this seems to be a minor application for saying the methods are acceptable, and it is not sufficient to ensure their use.

Dr. Ward stated that her organization hoped to see progress on one or more of the methods being recommended for validation by the ICCVAM and would like ICCVAM to act to validate the methods as soon as possible. Her organization also strongly opposes any additional animal testing for conducting new optimization or validation studies and requests that the isolated rabbit eye protocol contain stronger wording prohibiting the use of

laboratory rabbits. The word "should" in the recommendation is not a very strong statement to keep from using laboratory rabbits.

2. Ms. Nicole Cuellar (S.C. Johnson & Son)

Ms. Nicole Cuellar provided a corporate perspective on use of the BCOP assay. For the past 10 years, S.C. Johnson has put significant effort into trying to reduce the use of animals in hazard assessment and frequently has used alternative assays for product development and labeling decisions. The company currently uses the BCOP in a weight-of-evidence approach for hazard classification and labeling purposes for its non-registered products. A benchmark-related approach, using the BCOP assay, has been used with a variety of different product types. For non-registered products, such as air fresheners and cleaning products, the company has used alternative assays for labeling and hazard. For registered products, the company has used alternative assays more in product-development situations and worker safety assessments. By combining in-house historical data, toxicology information on raw materials, and postmarket surveillance, the company is comfortable using the BCOP alone, without *in vivo* testing, on non-registered products. This assay is an indispensable tool that the company has used for addressing the potential irritation of S.C. Johnson products.

Ms. Cuellar then described how S.C. Johnson conducts the BCOP assay; a standard protocol is used with concurrent benchmarks, controls, and histology if needed. The exposure and post-exposure times are chosen to be appropriate for the formula or chemical class or the test material. Each formula is carefully matched with a specific benchmark material for which the irritancy potential is well understood. Histology is conducted on both the test sample and the benchmark, for a complete assessment of degree and depth of injury. Histology is conducted under the following situations: to understand new chemistries and formulas; to investigate known chemistries with delayed effects; for chemicals where the mode of action is not easily predicted or the complete picture is needed. It is also used to further characterize damage not obvious from the standard BCOP endpoints and to resolve borderline cases.

In conclusion, Ms. Cuellar stated that she appreciated the enormous effort that has gone into the production of the BRD and review of the data for support of this assay. S.C. Johnson submitted five datasets for this evaluation, and is very supportive of this effort. The company respectfully requests that its comments be considered due to the wealth of investigation and its application by S.C. Johnson.

3. Dr. Rodger Curren (IIVS)

Dr. Roger Curren first addressed a few of the points addressed during the Panel recommendations for the BCOP test method. He stated that IIVS has negative control data available. Also, regarding the issue of whether the protocol is acceptable to laboratories other than IIVS that are running the BCOP test method, one of the things IIVS has done over the years is to have workshops just for users of the BCOP assay, where we have discussed, in general, among our laboratories, what protocols are appropriate.

Dr. Curren then addressed three issues related to the BCOP test method evaluation. He first described how histology was found to be useful in identifying the severe irritancy of a few substances that did not produce significant opacity or permeability in the EC/HO validation study of BCOP. IVIS, histology slides, *in vivo* Maximum Average Score (MAS) values, and EU and EPA classifications were presented for quinacrine and sodium oxalate. The substances are classified as Category I and R41, respectively, in the EPA and EU systems, but in the EC/HO study, they were predicted as mild in the BCOP assay (IVIS <25). However, histologically, these substances produce severe effects. For sodium oxalate, there is severe destruction of the epithelium and penetration of the material into the stroma. For quinacrine, there are severe changes in the stroma, and the endothelial cells are damaged. It is reasonable to call these lesions severe. Adding histology to this study improved the sensitivity from 81% to 90% for the EPA classification system.

Dr. Curren also discussed the prevailing definitions of accuracy versus concordance. He reviewed the American Society for Testing and Materials (ASTM) standard practice report, which has years of experience with validation-type exercises. The ASTM report defines accuracy as "expressing the closeness of a test result to a true value or an accepted reference value." An accepted reference value must be of fairly high quality. For the purposes of this Panel evaluation, many scientists think of the true value as a human result, with the animal as only an imprecise surrogate. Dr. Curren believes it is more accurate to refer to the BRD accuracy analyses as performance of the *in vitro* methods relative to the rabbit for eye irritation. The statistics in the BRDs form a set of performance statistics, not accuracy measurements. Concordance has often been the historical way to do it, from the original Cooper statistics. Performance statistics are measures of the concordance with the rabbit-test results.

Finally, Dr. Curren suggested considering the work of the biostatistician, Feinstein, from Yale. Dr. Feinstein talks about how sensitivity and specificity are not user statistics. The relevant statistics may be slightly different than looking at sensitivity and specificity.

4. Ms. Sara Amundson (Doris Day Animal League)

Ms. Sara Amundson provided some clarification on the recommendation to use eyes from 18- to 48-month old cattle. Based on her experience as an animal advocate and her knowledge of the slaughtering industry, the preponderance of beef cattle are going to be in that age range naturally, without having to optimize this as a goal set within the test-method protocol. It is not economically feasible to have beef cattle that are going to be more aged than that, from a rancher's perspective. She asked the Panel to keep in mind that this information is generally available directly from the slaughterhouse.

5. Ms. Sadhana Dhruvakumar (People for the Ethical Treatment of Animals [PETA])

Ms. Sadhana Dhruvakumar was pleased with the presentation of the Panel's recommendations and the very positive conclusions. She understood from the Panel presentation that the BCOP is acceptable for use, with certain caveats, and that the caveats

can be resolved retrospectively, with more existing data, which is something that can be done pretty quickly. She also requested that the Panel consider using the term "considered scientifically valid," or better yet, "validated" instead of using the term "acceptable for use." She believed that would have greater clarity than just saying "acceptable for use," if validity is indeed what the panel means by that term.

III. ISOLATED RABBIT EYE (IRE) TEST METHOD EVALUATION

Primary reviewers: Drs. James Freeman (Group Chair), Sally Atherton, David Lovell, Yasuo Ohno, Horst Spielmann, Peter Theran

Overview of the IRE Test Method Procedure

Mr. Robert Guest (SafePharm Laboratories; Derby, United Kingdom), an invited expert with many years of experience using the IRE test method, provided an overview with background information, technical aspects, a description of the ocular scoring system and a discussion of the decision criteria for identification of an ocular corrosives or severe irritants.

Dr. Guest indicated that the IRE test method was similar to the ICE test method previously discussed by the Panel. He indicated that SafePharm Laboratories has been performing toxicology studies for over 30 years and that 300 to 400 eye irritation tests are performed each year on a variety of test substances. Data from these tests are submitted to regulatory bodies worldwide and are used for occupational safety assessment. A tiered-testing strategy was used in accordance with OECD Test Guideline (TG) 405, EPA, and other regulatory guidelines. Initially, physicochemical properties, reactivity, corrosivity and other factors are considered before testing. Once tested, a positive outcome results in labeling as a severe irritant/corrosive. A negative outcome results in testing in a single animal, then a decision is made to follow-up with a full study, if necessary. The original IRE protocol was prevalidated using 14 chemicals, most from the European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC) reference bank, followed by a study with GlaxoSmithKline (GSK) using a total of 30 test substances. The principal advantage of the IRE is the use of the main target tissue of the rabbit eye (cornea) without pain or distress. He pointed out that the weighted Draize ocular scoring system places most of the weight (73%) on corneal scores. In addition to a qualitative endpoint used *in vivo* such as corneal opacity, quantification of corneal swelling is possible in the IRE. Corneal opacity (area and intensity), corneal swelling (percent increase in corneal thickness), and uptake of sodium fluorescein (area and intensity) are measured routinely. Histopathology is not performed routinely, but is available if necessary.

Dr. Guest discussed the technical aspects of the IRE test which are outlined as follows:

- The perfusion apparatus consisted of 11 custom-built temperature-controlled chambers with four temperature probes to monitor temperature (1200 British pounds [1740 Euros, $2268 U.S. dollars at the current exchange rates]).
- A portable slit-lamp and an ultrasonic pachymeter are used.
- Other standard laboratory equipment/instruments are needed such as temperature-controlled baths, peristaltic pumps, etc.
- New Zealand White Rabbit weighing 2.5-4 Kg are used.
- Control eyes are occasionally obtained from animals that have undergone skin testing, but would have had either no reaction or a mild reaction.
- Corneal pachymetry is performed at the optical center and at four other points, namely the 3, 6, 9, and 12 o'clock positions. During this procedure, the

animals do not even blink and there is no distress or damage to the cornea as a result of this measurement.

- Eyes are carefully dissected, placed in holders held in place by jaws, and then equilibrated for 30 minutes after placing the holders in the chambers with the bath temperature maintained at 32°C.
- Eyes are examined prior to testing and any with corneal effects or fluorescein penetration are rejected.
- Following equilibration, the isolated eyes are perfused with saline for hydration.
- Before application of test substances, the eyes are re-evaluated by slit-lamp for corneal effects, corneal thickness is measured again and a fluorescein examination is conducted. Eyes with an increase in corneal swelling greater than 10% are rejected.
- For application of the test substance, the holders are removed from the chamber and then placed to keep the eyes in a horizontal position. A volume of 0.1 mL or weight of 100 mg is applied to the surface of the cornea. For solids, the material is either gently compacted into a 0.1 mL volume in a syringe with the tip cut off or weighed in a gelatin capsule and sprinkled over the cornea.
- The corneas are evaluated macroscopically and by slit-lamp and scored for opacity and area. The mean corneal thickness from five measurements is calculated. The advantage of the optical pachymeter is that measurements are not precluded by corneal opacity or dye uptake. Corneal evaluation and thickness measurements with epithelial changes (i.e., mottling, stippling, sloughing, or ulceration) are evaluated under diffuse illumination (1, 2, and 3 hours) and a more detailed evaluation using a slit-lamp biomicroscopic examination is then conducted (1, 2, 3, and 4 hours) to look at the stroma and endothelium.

The ocular scoring system of Hackett and McDonald (1991) was used; a more detailed scoring system than the Draize. At 4 hours, fluorescein uptake was scored for area and intensity.

The decision criteria used to identify a test substance as a severe ocular irritant or corrosive were developed retrospectively, by looking at the data generated in-house. A severe irritant/corrosive is identified as any test substance that produces a maximum corneal score (in three tested eyes) of 3 or greater (opacity x area), maximum fluorescein penetration of 4 or greater (area x intensity), mean corneal swelling (n=3) equal to or greater than 25%, or any single incidence of disruption of the epithelium (stippling, mottling, etc.) in which the control eyes did not respond. If any of these criteria are met, the substance is labeled a severe irritant/corrosive and it is not tested *in vivo*. This testing strategy has been in use since 1999 and has resulted in a reduction in the number of animals exposed to severe irritants. Use of the IRE in conjunction with other *in vitro* tests (e.g., human reconstituted tissue models), would, hopefully, result in replacement of the rabbit eye test.

After Mr. Guest completed his presentation, Dr. Scala invited the Panel to ask him questions on the procedural elements of the IRE assay.

Summary of the IRE Test Method BRD

Dr. Allen (NICEATM) thanked Mr. Truax (NICEATM) for compilation of the BRD. Dr. Allen pointed out that although the IRE test is currently accepted in the EU as a screen for the identification and labeling of ocular corrosives and severe irritants, IRE test data have not been submitted to U.S. regulatory agencies. Dr. Allen indicated that there were four primary sources of data for the IRE test method (CEC 1991; Balls et al. 1995; Gettings et al. 1996; Guerriero et al. 2004). Dr. Allen showed the number of test substances used for each of the regulatory classification systems (GHS, EPA and EU) and he noted that the numbers with an EU classification were higher, because some *in vivo* data were provided with assigned EU classifications, but the *in vivo* raw data were unavailable to allow conversion of the data to the other classification systems. No intralaboratory data were available for analysis in any of the reports, although interlaboratory data (three to four laboratories) were provided in the CEC (1991) and Balls et al. (1995) reports. The EU classifications assigned to eight compounds on the basis of skin corrosivity test results or pH extremes were excluded, since the accuracy analysis is base don the ability of the ICE test method to correctly predict ocular responses in the intact rabbit.

The database consisted of 149 test substances with 124 chemicals and 25 products or formulations. Fifteen chemical classes (heterocyclics/aromatics, acetates/esters, and formulations were the most common) and 14 product classes (industrial chemicals, soaps, and surfactants were the most common) were identified. The endpoints from the recommended IRE protocol in the BRD were corneal opacity, corneal swelling, fluorescein penetration, and evaluation of epithelial integrity. However, the Gettings et al. (1996) study measured only corneal swelling, while Balls et al. (1995) evaluated the first two parameters, the CEC (1991) study used the first three endpoints, and Guerriero et al. (2004) used all four endpoints. Accuracy of each test method was evaluated based on use of all three regulatory classification systems, when possible. The Guerriero et al. (2004) data demonstrated accuracy in the 77-78% percent range with false positive rates of 33 to 34% and false negative rates of 0, albeit with a small number of compounds tested (n=36-44). The small numbers of test substances within any particular chemical class (n=2 to 3) made it difficult to evaluate the performance of any specific chemical classes.

Reproducibility analyses indicated that in the Balls et al. (1995) study using 59 substances and four laboratories, 59% of the time all four laboratories were in agreement with respect to the outcome. Three of four laboratories were in agreement 85% of the time. In the CEC (1991) study with 21 substances among three laboratories, all three laboratories were in agreement 81% of the time. Three of four laboratories were in agreement 95% of the time. If limited to severe irritants, the four laboratories in Balls et al. (1995) were in agreement 100% of the time and the three laboratories in the CEC (1991) study were in agreement 83% of the time. In the Balls et al. (1995) study, a wide range of %CV values (0-200%) using corneal opacity and corneal swelling as endpoints were obtained; the median %CV values was 43.4% and 49.7%, respectively. If only GHS Category 1 substances (i.e., severe

irritants) were considered, median %CV's were 35% to 40%. For the CEC (1991) data, median %CV's were in the range of 24 to 43% (endpoint times were slightly different for some laboratories in this study). For GHS Category 1 irritants, %CV's ranged from 15 to 30%.

A recommended standardized protocol using corneal opacity, corneal swelling, fluorescein penetration, and evaluation of epithelial integrity was proposed in the BRD. This is essentially the SafePharm Laboratories protocol described by Guerriero et al. (2004) with additional positive, negative, and reference controls included.

Proposed Panel Recommendation for the IRE Test Method

1.0 IRE Test Method Rationale

Dr. Freeman pointed out that the Panel concurred with the description in the BRD, but recommended inclusion of a discussion of potential cellular mechanisms of ocular corrosion and severe irritation. This discussion should include the relevance of necrosis and apoptosis to *in vitro* testing, the role of resident and migrating inflammatory cells and their products in ocular irritation *in vivo*, and the consequence of having an incomplete response *in vitro*. The BRD needs to be updated to reflect the basis of the test method as a correlation of descriptive observations or toxicity, rather than mechanistic. In addition, the IRE Panel suggested that additional studies such as microscopy and immunohistochemistry might add to the accuracy of the test method.

The regulatory use and rationale was thoroughly covered. However, it should be noted that the IRE test method does not account for effects on the iris and conjunctiva, nor does it account for reversibility of corneal effects or systemic effects or identification of slow-acting irritants.

The IRE Panel recommends consideration of the use of microscopy or histopathology to improve sensitivity and scope. These efforts may provide insights into early markers of effect or identify transient *versus* progressive changes.

2.0 IRE Test Method Protocol Components

There is a limited dataset using the recommended IRE protocol. The recommended protocol enhancements do appear to improve the accuracy of the test. The recommended protocol, however, has not been directly assessed across other laboratories. In addition, the decision criteria (prediction model) may need to be modified by use of a statistical paradigm (e.g., discriminant analysis) to enhance performance (i.e., to reduce the false positive rate without appreciably increasing the false negative rate). In addition, positive and negative controls and reference substances should be identified from a validated reference substance list such as that being prepared by the Expert Panel Reference Substances Subcommittee.

The IRE Panel recommends that appropriate sources of rabbit eyes be defined. Acceptable rabbit strains should be identified. Acceptable storage and transport conditions (e.g.,

temperature limits, time limits, required buffers, salt or other solutions, containers, etc.) of the isolated eyes for shipment should be defined.

The IRE Panel agrees with the BRD that corneal opacity and swelling, fluorescein penetration, and epithelial integrity should be used as endpoints in the IRE test method. In addition, identification of the reference substances that are part of the performance standards developed for the validated test method is recommended. Data should be collected according to GLP-compliant procedures. Finally, the BRD should clarify the orientation of the eye during application of the test material. Other considerations include application of confocal microscopy or histopathology to detect changes at the cellular level, quantification of the observation (e.g., counting pixels) where possible, and use of descriptive statistics based on individual scores. Finally, the statistical algorithm or rationale used to establish the decision criteria should be more clearly defined and stated in the BRD.

Any further additions to the test method should be backed by a specific rationale. The recommended protocol was adequately covered in the BRD. Consideration of the use of histopathology and defined, validated reference substances should be included as previously described by the IRE Panel. The types and numbers of substances used for prior validation have been adequately described in the BRD. However, further optimization or validation would require use of the reference substances list being developed by the Expert Panel subgroup.

3.0 Substances Used for Previous Validation Studies of the IRE Test Method

The Panel agreed that the number and classes of substances tested in previous validation studies were acceptable.

4.0 In Vivo Reference Data Used for Assessment of Test Method Accuracy

The IRE Panel noted that the Draize test has been used basically unchanged for decades. However, it was suggested that the Draize test could be improved *vis à vis* by use of some of the technology being considered for use in the *in vitro* studies.

The IRE Panel considered the interpretation of the results to be correct as described in the BRD. However, a question was raised by the Expert Panel regarding the adequacy of using regulatory classification systems for evaluating *in vitro* methods and the suitability for chemical or product class evaluations.

Issues regarding data quality were adequately written in the BRD. The IRE Panel felt that if evaluation of the results can be made and the quality of the study appears to be adequate, then lack of original study records does not raise undo concern about a study. Reference to data quality and use of GLPs is covered in the BRD. If the work is performed in a well-established laboratory, then no distinction between GLP-compliant or non-compliant studies is required, and a lack of GLP-compliance *per se* is not a sufficient criterion for exclusion of data of evaluation of performance.

Previously, the Expert Panel had expressed the need for greater effort to find and consider human topical ocular chemical injury data. However, it was recognized that limited data are available and dose and exposure rates would be difficult to quantitate. In addition, no scoring or time course data would likely be available for comparison to an *in vitro* test method.

The IRE Panel indicated that more discussion of the variability of the *in vivo* data is needed in the BRD. The question is how concordant is rabbit data with human data. There is a need to develop an acceptable reference standard, since we do not know if inaccuracy results from inconsistencies in the *in vitro* test method or from misclassification based on a single *in vivo* result.

Further optimization or validation studies should use existing animal data, if possible. Additional animal studies would be used only if data gaps are identified. Such studies should be carefully designed to garner as much information as possible and maximize the amount of pathophysiological information obtained.

Dr. Stevens expressed and provided a written minority opinion that no additional animal tests should be performed for this purpose.

5.0 IRE Test Method Data and Results

The Panel agreed with the BRD assessment of the information presented in Section 5.0 of the document.

6.0 IRE Test Method Accuracy

The recommended protocol includes the additional parameters that enhance accuracy (e.g., Guerriero et al. 2004). No additional datasets were produced with that method. The statistical methods were limited, but appear to be appropriate for descriptive toxicology data, and the conclusions on reliability in the BRD appear to be sound.

Documentation of the data quality was adequate. The studies using the recommended protocol were conducted according to the spirit of GLP.

The reference studies analyzed in the BRD were independent efforts, so lot-to-lot consistency really did not apply here. The consistency was controlled and described within three of the four studies, but not described in the fourth. The stability of chemicals over each study's time frame was not discussed in the BRD.

This section was adequately described in the BRD. Accuracy results summarized in Section 6.1 (Tables 6.1 to 6.3) of the BRD provide a correct overview of performance as reported in the literature, as well as discordant results. Accuracy appears to be improved, based on a small n, with the recommended method, resulting in a false negative rate of 0 and a false positive rate of 33 to 34%.

Draize variability must be included in the discussion in the BRD. There is a weakness in this evaluation. That is, there is a lack of a common protocol for all of the studies analyzed. However, the IRE Panel found it encouraging that the accuracy appeared to improve in the protocol that has become the recommended protocol.

This section was adequately described in the BRD. However, the BRD needs to be revised to assure that the temporal sequence of the studies described is consistent with the publication dates. In Section 6.3 (Tables 6.4 and 6.5), the source of the *in vivo* and *in vitro* data and appropriate author information should be included, and the datasets used to calculate the irritancy classifications identified.

Differences in reproducibility of the Draize test must be taken into account when comparing the predictive value of *in vitro* alternatives. Other relevant information (e.g., weight-of-evidence approach) may clarify the performance of the IRE. It was noted that variability of the Draize test for corrosives or severe irritants might not occur to the same extent as it does for milder irritants.

7.0 IRE Test Method Reliability

The BRD should incorporate information from Bland and Altman (1986) that discusses statistical comparison of methods with poor reproducibility. Information from the European Centre for the Validation of Alternative Methods (ECVAM) skin-irritation prevalidation study on repeatability and reproducibility should be obtained and incorporated where relevant to the ocular test systems. Information from Dr. Sebastian Hoffman's detailed variability analysis comparing SD's and CV's for two skin-irritation models, where relevant, should be incorporated. A strategy to evaluate reliability in any future optimization and validation testing should be developed and implemented.

The IRE Panel concurred with the BRD. No data were provided for multiple studies from a single laboratory. Neither intralaboratory repeatability nor reproducibility could be assessed. Quantitative interlaboratory reproducibility was assessed in two of the four studies, which used slightly different protocols. The recommendation is that reproducibility analyses should be conducted from studies using the recommended protocol and the approved list of reference substances.

The availability of historical negative and positive control data was appropriately covered. However, positive controls have not been consistently used. In future studies, this information should be tracked.

There appears to be no impediments to minor protocol changes or to transferability of the IRE test method. It may be useful to contrast results developed using the SafePharm recommended protocol *versus* earlier renditions. Good agreement across the board with *in vivo* data would suggest that existing data from all of the protocols could be used as validation data. Any differences in protocols used for future studies should be specifically justified.

8.0 IRE Test Method Data Quality

The IRE Panel concurred with the BRD. A lack of GLP compliance *per se* is not an exclusion criterion. Although not all studies were considered GLP-compliant, the reviewed data appear to be of satisfactory quality.

This was covered in the BRD adequately. Verification of accuracy of data against original data records is beyond the scope of the IRE assessment. The impact of deviations from GLP guidelines in Section 8.3, are appropriately covered. Noncompliance with GLP was not considered a mandatory exclusion criterion. All laboratories that performed the studies were considered reputable.

This was well covered in the BRD. The original raw *in vitro* data for all studies was not available for review. This data cannot be audited retrospectively. The quality of the institution, the reputation of the individual researcher, and evidence of reproducibility of results must guide our confidence in the accuracy of the data.

9.0 Other Scientific Reports and Reviews

This was adequately covered in the BRD. The submitted Procter and Gamble Enucleated Rabbit Eye Test (ExRET) and Low Volume Eye Test (LVET) data were not readily transferable to other studies for regulatory classification, and thus were excluded from the overall analysis. Reviews of all relevant published IRE studies were included in the BRD.

The IRE Panel felt that the conclusions reached on the report summaries were adequate and complete.

Appropriate measures were taken as described in the BRD. A *FR* notice was sent and study authors were contacted to request original IRE data and *in vivo* reference data. The IRE Panel acknowledged that obtaining this data was a difficult process.

10.0 Animal Welfare Considerations (Refinement, Reduction, Replacement)

This was appropriately covered in the BRD. It is important to determine the availability of rabbits to be used for this purpose. Furthermore, rabbits should not be raised or sacrificed specifically for use in this test. Currently, most U.S. Federal regulatory agencies do not permit prior use of animals for ocular testing. Therefore, the availability of eyes from an abattoir may be a factor for further development of this test method. The test method could be considered a partial replacement, under the 3R's, if eyes indeed were available.

11.0 Practical Considerations

The IRE Panel felt that transferability of the IRE test method could be readily achieved.

Training needs to be conducted with experienced personnel. Training videos and visual aids would be useful as discussed at other sessions.

This was adequately described. However, the cost for conducting the tests in each country should be obtained to reflect differences at the current exchange rates. For example, a laboratory in the United Kingdom may run the test with controls at $1074.00 using the current exchange rate for both *in vitro* and *in vivo* eye tests, but the actual cost in the U.S. may be significantly higher due to labor costs and other factors.

The *in vivo* test may take up to 21 days, whereas the *in vitro* IRE test takes up to 4 hours.

12.0 Proposed IRE Test Method Recommendations

The IRE Panel concurred with the BRD recognizing that the recommended version of the protocol was only conducted in one laboratory, and limited data were generated using that protocol. The most appropriate version of the IRE test method was selected using the additional endpoints identified.

The appropriate source of rabbit eyes needs to be defined. Currently, not all U.S. regulatory agencies will accept ocular data from studies in which the rabbits were used for other experimental purposes. The IRE Panel recommends that the U.S. practice be revised and updated, if possible.

Although the decision criteria for identification of ocular corrosives/severe irritants are defined, a rationale for them and/or a discussion of statistical algorithm(s) used in their development should be provided.

A standardized scoring system for histopathology should be more clearly defined to maximize the likelihood of obtaining reproducible results. In addition, reference photographs for all subjective endpoints should be developed to aid in training and transferability.

The recommended IRE test method appears to be capable of identifying ocular corrosives and severe irritants in a tiered-testing strategy. However, the database classifiable as GHS is small (n=36) and there is lack of data on reproducibility. In order to accept IRE data for classification purposes, the database needs to be expanded to corroborate the current results. However, it may be possible to look at existing data such as the CEC (1991) data to expand the database, since they used three of the four recommended endpoints and the accuracy was similar to that of Guerriero et al. (2004). The low false negative rate (0%) is encouraging meaning that either few or no animal tests would need to be performed. The false positive rate (33 to 34%) is relatively high. This false positive rate could be reduced by optimization of the decision criteria through employment of appropriate statistical methods (e.g., discriminant analysis).

Additional Expert Panel Discussion/Recommendations

- For consistency in histopathology assessment, it was recommended that the Panel request the Society of Toxicological Pathologists or NICEATM to set

up a standardized ocular histopathology grading system with visual aids and publish the information as soon as possible to provide uniformity in ocular tissue evaluation.

- To provide consistency among the BRDs, text in Section 8.4 concerning the need for laboratory notebooks containing the original raw data, Drs. Green and Freeman, modified the text to read, "…availability and review of raw data would improve the confidence in the data."
- In response to a concern regarding a potential issue of whether data used in the BRD should be subjected to a GLP audit, particularly if data was obtained from a single laboratory, Dr. Scala noted that in the ICCVAM documents the criteria for validation are presented (i.e., page 3) and indicate that, "All data supporting the assessment of the validity of the test method must be available for review." The final language should, therefore, be consistent with that statement.
- A change in wording was recommended for Section 6.3 of the IRE BRD to "Also recognize that the variability of the Draize test for corrosives or severe irritants is lower."
- Concern was raised that the IRE test method was not being endorsed for use in a tiered-testing strategy as the other test methods were and that there was a recommendation for further testing. It was pointed out that the test appears to be useful for identification, but the numbers to support it are low and the data needs to be corroborated. There was also concern that the data came from a single laboratory and no intra- or inter-laboratory reproducibility was available using the recommended protocol. However, it was stressed that the ICE data also came from a single laboratory, and this was not objected to.
- The current U.S. regulatory policy for acceptance of eye data from animals with prior treatment is inconsistent and it was recommended that this position be clarified and a proposal made to regulatory agencies to reconsider the use of such animals.
- Since consensus for use of the terms Prediction Model and Data Interpretation Procedure were not obtained at OECD or other meetings, Dr. Stokes suggested for now that the term "decision criteria" should be followed by "prediction model" in parentheses.
- Additional data on the IRE should be requested from other companies performing the test.
- The Panel continued to discuss specific points in the IRE BRD to harmonize wording across other BRD recommendations, where necessary and to reword some language of the proposals in IRE presentation for consensus agreement (See Expert Panel transcript for details).

Panel Vote on the IRE Report

The Panel Chair asked the Panel members to indicate whether they accepted the revised document that had been presented, with all the revisions that had been placed in it. A show of hands indicated that all Panel members, except for Dr. Stevens, accepted the revised IRE report. Dr. Stevens indicated stated that he voted no, because of his opinion that animals

should not be used for future validation studies, and because of his minority opinion on Section 4.0. Dr. Stevens indicated that he agreed with the Panel on all other parts of the IRE report. Dr. Guerriero abstained from voting, because he had a conflict of interest as a principal user of the test method.

Public Comment Session

1. Dr. Pauline McNamee (Procter & Gamble, Co; representing The European Cosmetic, Toiletry, and Perfumary Association [COLIPA])

Dr. Pauline McNamee congratulated ICCVAM on the tremendous amount of work, both in terms of scope and depth, which went into compilation of all four BRDs for this very important activity on eye irritation.

Members of COLIPA compiled a list of technical comments on the IRE test method. The first related to use of a statistical method for determination of the decision criteria that was adequately addressed in the Panel recommendations. COLIPA also welcomes consideration of the use of histopathology as an endpoint for consideration in the IRE, as well as in the ICE and BCOP test methods. This effort would build on the initial work by Jim Maurer and Rosemarie Osborne. COLIPA requested that the Panel consider the recommendations supporting the proposed IRE protocol on the number of eyes tested, use of concurrent positive and negative controls, or reference substances to reflect the practical limitations associated with the ability to perfuse and assess all of these eyes at any one time, and also in terms of the ability to appropriately time treatments, measurements, and other functional aspects of the test.

COLIPA recognized that application of the standardized protocol decision criteria to all of the datasets was problematic, because the number and type of endpoints varied from study to study and differences in scoring scales were used. However, COLIPA suggested that an effort be made to contact the authors of the studies to determine if data could be obtained for use in a weight-of-evidence approach for the evaluation of all data in the IRE BRD.

COLIPA reiterated previous public comments that it very much welcomes the continued clarity resulting from this meeting and in further development of the BRDs and encourages further efforts to ensure that 1) the most comprehensive data package available is used in the BRDs, 2) those data sources are used in the overall test evaluation, and 3) additional work needed due to the complexity of the protocols involved and limitations of existing data would be done after examination of the relationship between the experimental protocols, adjustment of the decision criteria, and subsequent interpretation of the data. COLIPA was strongly convinced that a weight-of-evidence approach could be applied to the post-hoc evaluation of these test methods. Furthermore, COLIPA welcomed a retrospective analysis to determine 1) what is needed to move forward, 2) identify specific research needs on mechanisms of chemically-induced eye irritation, and 3) lead to further optimized methods and/or new methods such as those currently being explored and researched in the COLIPA eye-irritation research program.

COLIPA closed by commenting that they stand ready to continue to collaborate with ICCVAM and ECVAM on this and future activities in the area of eye irritation.

Ms. McNamee indicated that Procter & Gamble has used farm-raised a rabbits as a source of eyes for an *in vitro* test for years and would be willing to provide information to ICCVAM on the suitability and acceptability of those eyes for use under the circumstances of removal, transport, and use in the laboratory, if that data is useful to them.

2. Ms. Sadhana Dhruvakumar (PETA)

Ms. Dhruvakumar expressed concern about use of the term accuracy, which she indicated was defined as "the quality of nearness to the truth or the true value." Regardless of the definition in the glossary of the BRD, the term accuracy implies that the rabbit data are the truth. Since we have come a long way from assuming that the animal data are the gold standard, our language should reflect that and use the dictionary definition of accuracy.

The second comment was related to agreement with comments from Ms. Sarah Amundson that the documents far overstep their bounds and led the process. The scope and content of the BRD should have been to present the data to the Expert Panel and allow them to consider that the method were 1) scientifically valid and useful today, 2) unscientifically valid in certain circumstances or with certain limitations, or 3) they are scientifically invalid and are not ready to be useful in any circumstance today. For example, any considerations of improvements or optimization should have been secondary to where these test methods stand in terms of usefulness. Instead, conclusions and additional optimization and validation studies were proposed for every method in the BRDs. The Panel was therefore focused on possible improvement, but did not draw clear conclusions on the current validity of the test methods. This was a step backwards for these test methods and a disservice to them.

These methods have been in use for approximately 20 years. Furthermore, these tests have been accepted for use in a tiered-testing strategy by some European countries for approximately a decade, and are currently accepted by all EU countries because of mutual data acceptance. Several have had extensive laboratory validation studies of them. These methods have stood the test of time, which is also the only test of accuracy that has been applied to the *in vivo* rabbit eye test. These *in vitro* tests have already proved their utility. The fact that these tests were evaluated for use in identification of ocular corrosives/severe irritants should have been a "slam-dunk." If the Panel cannot agree that even one of these test methods is valid for use as a partial replacement right now, without years of additional work, there is little hope of ever getting to a point where mild irritants can be assessed to provide complete replacement of the Draize test. If the process is perceived to take these tests backward, rather than forward, ICCVAM will not be in a position to receive new nominations.

PETA requests that the Panel consider whether or not these tests are scientifically valid and potentially useful as a positive screen for ocular corrosives and severe irritants in any definable set of circumstances, and deemed to be valid for a specific purpose. Also, existing data should be used retrospectively and combined with the scientific judgment of the Expert

Panel to validate this test method. If none of these test methods are judged to be valid for current use, faith in the ICCVAM process will be significantly affected.

3. Dr. Robert Guest (SafePharm Laboratories)

Although an invited expert on the IRE test method, Dr. Guest asked to have time to provide public comment. He indicated that a comment made in a previous session on the statement that use of coded or blinded substances was a GLP requirement and he clarified that, as far as he was aware, it was not.

Dr. Guest commented on the use of rabbit eyes from animals that had been used previously for other tests. Typically, at SafePharm laboratories, rabbits used for skin testing of mild or nonirritant substances are routinely used as a source of eyes in the IRE. Dr. Guest asked that consideration of the use of these animals, as unacceptable sources of rabbit eyes, should be reconsidered. He noted that numerous controls are in place. The animals are allowed sufficient time to recover, certain limits or conditions for *in vitro* use are imposed, and the eyes are carefully examined by slit-lamp *in vivo* and again after enucleation. These eyes remain viable and in good condition for the test and there is data available to support this conclusion. There are certain substances that may have delayed effects on eyes or exhibit systemic toxic effects on local exposure. However, safeguards to avoid use of such substances are in place. Any corneal effects produced by such substances would be identified on examination. The advantage of this method is that the eyes are fresh and animals are spared.

With respect to false positives, the IRE may identify ocular corrosives and severe irritants or even some less irritant substances may be overlabelled. This does not mean that a nonirritant will be classified as an irritant. The company that submits the data for testing must eventually decide on labeling, the level of tolerance they have for false positives, a decision often based on additional weight-of-evidence information.

Yesterday, there was a question pertaining to why more laboratories do not perform the IRE test. Several laboratories do run the IRE test and at least one runs the ICE test as well, but generally the regulators do not ask for the data. If regulators asked for the data, companies would perform the test. If the test is not required, few companies will apply resources unless it is necessary. There is a commitment in most laboratories to reduce the use of animals that drives the development of alternatives. There are pressures to reduce animal use and efforts are underway in the EU to change legislation due to the chemicals-notification programs, such as the Registration, Evaluation and Authorisation of Chemicals (REACH) system, and the 7th amendment to the Cosmetics Directive, to require use *in vitro* tests for testing of these products. Something has to be done now, and this meeting has given us a fantastic opportunity to look at these methods and approve them. Is there a regulatory barrier that would preclude parallel use of an *in vitro* test method with other data to extend or expand the database for at least certain types of products?

Dr. Scala noted that one comment made on regulatory barriers and another regarding classification of false positives were beyond the scope of this Expert Panel, although regulators certainly will consider the implications.

4. Dr. John Harbell (IIVS)

Dr. Harbell noted that the Expert Panel might be considered the supreme court of science, since what is done here and the statements made about how the tests are evaluated will be the models for how this is done in the future. It is the Expert Panel at this public meeting that provides the checks and balances for the executive branch, the ICCVAM/NICEATM. In 1992, a similar meeting was held to design submission criteria for the IRAG program with Drs. Bradlaw, Gupta, Green, and Wilcox. In that meeting, Dr. Scala stressed the importance of continuous evaluation of data rather than categorization, and the importance of looking at individual animals and their performance. Then in 1996-1997, we met as stakeholders to design the ICCVAM program. Before federal funding, this was an informal committee, but the paradigms for validation were laid down at that time by all of the stakeholders. In that document, relevance was defined in the glossary as the extent to which a test is meaningful and useful for a particular purpose, that is the extent to which a test method will correctly predict or measure a biological effect of interest. Although animal welfare is a noble cause, it is not the only driving force for development of an *in vitro* test method. Dr. Scala's work in the predictive capacity of the Draize test, and subsequent work by Dr. Marzulli in 1973 indicate that the predictive capacity of the Draize test is limited. For example, there is a phrase in the paper by Dr. Marzulli, "Furthermore, collaborative results indicate that additional study to identify and eliminate sources of variability is necessary before reproducible results with regard to comparison of degrees of irritancy can be obtained." We have moved from relevance to using the word "accuracy" that leads to a fundamental change. This change is reflected in Criterion 4 in the BRDs that says "the potential for the proposed test method to provide improved prediction of adverse health or environmental effects compared to the current test methods accepted by regulatory agencies." This is a criterion for consideration by ICCVAM. The response of ICCVAM/NICEATM is, "It is proposed that the current animal test provides a suitable assessment for eye-irritation potential in humans." There is not complete agreement on this statement.

The last point regards a paper called *Ophthalmological Perspectives on Eye Irritation Testing*, which says, "We note that two major themes should permeate all future work to further development of alternative tests. First, we unanimously agree that the Draize rabbit eye test method as currently used should not be considered the primary standard for the evaluation of new test methods." Dr. Harbell noted that two of the authors, Drs. Chambers and Edelhauser were members of this Panel. He asked that the Panel consider these points of view and add them to the deliberations.

IV. HEN'S EGG TEST – CHORIOALLANTOIC (HET-CAM) TEST METHOD EVALUATION

Primary reviewers: Drs. Shayne Gad, Donald Fox, Martin Stephens, Frederick Guerriero, Sidney Green, Philippe Vanparys, Nancy Flournoy

Overview of the HET-CAM Test Method Procedure

Dr. Klaus Krauser provided an overview of the HET-CAM test method. Dr. Krauser discussed that the test method was first proposed by Professors Luepke and Kemper both from Germany. In 1988, a validation project was started in Germany, and funded by the government of the Federal Republic of Germany, to evaluate the validity of the HET-CAM test method. In 1992, the protocol used in this validation project was published in INVITTOX (Protocol No. 47), which is in the ERGATT/FRAME databank. Dr. Krauser noted that there had been other validation efforts conducted over the years, as well.

Dr. Krauser stated that a defined dose of a test substance is applied to the chorioallantoic membrane (CAM) of fertilized and incubated hens' eggs and the CAM is then evaluated for up to 300 seconds for the development of defined endpoints. Dr. Krauser then reviewed the evaluation of endpoints either involves determining the time elapsed until the first appearance of the endpoints or the severity of each endpoint is evaluated after a certain time.

Dr. Krauser then provided a brief review of the CAM. He stated that the CAM was a vascularized respiratory membrane that surrounds the developing bird embryo. It is composed of three parts. Dr. Krauser stated that the blood vessels that are present in one of the CAM layers are branches from the embryo-allantoic arteries and veins and they form a capillary bed. He then reviewed drawings to further describe the locations of the various parts of the chicken embryo.

Dr. Krauser stated that most current HET-CAM protocols used White Leghorn hen's eggs that were fresh and not older than 7 days. The eggs are fertile and clean and between 50 and 60 grams. Eggs are typically candled prior to use to ensure viability. Dr. Krauser stated that commercially available incubators with an automatic rotating device are used and that the eggshells are usually opened with small saws or dentist rotary saws.

Dr. Krauser stated that the historical negative control substances were typically saline solution, or other vehicles. Positive control substances were sodium hydroxide or 1% sodium dodecyl sulfate.

Dr. Krauser then reviewed the preparation of the test system. He discussed the age of the eggs used, the incubation temperature, the relative humidity, rotation of the eggs, removal of the eggshells, treatment with the test substance, test substance volume, exposure duration, and endpoint evaluation. Dr. Krauser noted that of all the endpoints that could be evaluated, the evaluation of hyperemia was the most variable. Dr. Krauser stated that the endpoints currently evaluated in the HET-CAM test method are hemorrhage, lysis, and coagulation. He

then provided additional detail about the observations made for each of these endpoints during conduct of the test method.

Dr. Krauser then provided detailed information on the measurements of the endpoints, number of replicates used in the conduct of the test method, number of repeat experiments conducted during studies, calculation of irritancy potential and/scores, decision criteria, and acceptance criteria.

Summary of the HET-CAM Test Method Background Review Document

Dr. Choksi (NICEATM) described the analysis of the HET-CAM test method presented in the HET-CAM BRD. She indicated that ICCVAM agencies were surveyed and to the best of their knowledge HET-CAM data had not been submitted to U.S. regulatory agencies for ocular irritation purposes. Dr. Choksi stated that there were 10 studies that were evaluated in the BRD. Of those studies, there was no information on intralaboratory reproducibility. Three different studies were used to evaluate interlaboratory reliability. Additional data had been received since the draft BRD was published. The information would be included into the revised HET-CAM BRD. In addition to the 10 studies, there were 39 additional studies that were not evaluated because comparative *in vivo* data were not available or the *in vitro* data was qualitatively described. These reports were described in Section 9 of the BRD.

Dr. Choksi stated that there were five different analysis methods that have been described in the literature in the 10 studies evaluated. In the HET-CAM BRD, there were 246 substances evaluated in 253 tests. Most of the substances tested by the IS(A) and IS(B) analysis methods were formulations while the other three analysis methods evaluated mostly chemicals or pharmaceutical intermediates. Chemical classes evaluated, where there were at least three substances or more per class, were alcohols, carboxylic acids, amines, and formulations. Product classes evaluated, where there were at least three substances or more per class, included cosmetics, solvents, hair shampoos, and soaps and surfactants. Dr. Choksi provided a breakdown of the number of testing laboratories that tested the substances.

Dr. Choksi provided variations between the testing protocols used by the studies. Variations included differences in incubation time, temperature and humidity, the amount of volume tested on the CAM, whether the substance was rinsed from the CAM, and the endpoints evaluated. Endpoints evaluated in the studies described in the HET-CAM BRD included hemorrhage, lysis coagulation, hyperemia, and dilation.

Dr. Choksi reviewed of the accuracy of the HET-CAM test method, when compared to the GHS classification scheme. Of the four analysis methods described (IS(A), IS(B), Q-Score, and S-Score), the IS(B) analysis method appeared to be the most accurate. Using the IS(B) analysis method, accuracy statistics versus GHS, EPA, and EU were provided. Dr. Choksi stated that an additional analysis had been conducted the week prior to the Expert Panel meeting, which was given to the Panel as well as was provided to the public. Data in the third phase of the ZEBET evaluation of HET-CAM was evaluated based solely on the irritation score of substances evaluated. The concentrations tested in this evaluation were 10% and 100 %.

Limitations of the accuracy analysis were then discussed; these included:

- the impact of the differences in test method protocols between the studies reviewed was unknown
- the impact of the different end points evaluated in the studies reviewed was unknown
- the lack of severe irritant test substances evaluated
- the lack of solids evaluated
- the limited chemical classes and product classes evaluated

Dr. Choksi then reviewed the reliability described in the BRD. Only one interlaboratory data set with the IS(B) analysis method was available for evaluation. Qualitative and quantitative evaluations of these data were described. Limitations of the reliability analysis in the draft HET-CAM BRD included lack of intralaboratory reliability information and the interlaboratory data was based on a very small number of substances.

Dr. Choksi stated that the standardized version of the HET-CAM test method proposed in the BRD is similar to the one that was provided by ZEBET (INVITTOX 1992), which uses the IS(B) analysis method. The proposed test method protocol requires the use of positive and negative controls. Some proposed additional optimization studies in the HET-CAM BRD included retrospective analysis of the decision criteria used to identify corrosives and the evaluation of additional endpoints, such as trypan blue, which might provide some quantitative information compared to the qualitative currently obtained from the test method.

After the conclusion of the presentation, Dr. Scala requested questions from the Panel. Dr. Green questioned whether there was any data provided to support the states that the hyperemia endpoint was not reliable for use in the test method. Dr. Choksi replied that the information was provided by personal communication and she did not have data that supported that statement.

Proposed Panel Recommendations for the HET-CAM Test Method

1.0 HET-CAM Test Method Rationale

Dr. Green presented the draft recommendations developed by the primary reviewing group ("the Group") for the HET-CAM test method for consideration and concurrence by the full Expert Panel. Dr. Green stated that the Group felt that the CAM was most similar to the conjunctiva and that a deficiency of the method was that the CAM has no structures similar to the iris or cornea. He noted that the method currently was most useful for assessing short-term effects of substances. Dr. Green stated that the ability of HET-CAM to provide improved prediction of adverse health effects in humans would be difficult to achieve unless comparative data for the *in vitro* test method, animals, and humans was generated using substances from a repository. The Group proposed that the method may have the potential to complement others in a tiered testing approach.

The Panel discussed the ability of the HET-CAM test method to assess substances that induce delayed effects in rabbits or produce reversible effects *in vivo*. The Panel proposed that even though the method cannot directly evaluate those things, that a recommendation be included that suggested exploring the relationship between the short-term effects observed in the HET-CAM test method and the long-term effects in the eye. The Panel agreed to include the recommendation into the Panel report.

The Panel also proposed to revise text in the report that stated that the CAM had no structure similar to the iris or cornea. It was discussed that that the CAM did not contain structures similar to the cornea. It was originally proposed to remove the term "cornea". Panel discussion followed on the topic that there were two issues related to the similarity of the CAM to the structures of the eye: vasculature and avascular tissue. The Panel then proposed to revise the statement so that the phrase read that the CAM tissue structure was not similar to the cornea.

2.0 HET-CAM Test Method Protocol Components

Dr. Green stated that the Group believed that the recommendations in the draft BRD appeared to best standardize test method procedures among the various published literature sources and developed a consistent scoring and calculation of irritation indices. The Group agreed with the BRD recommendations on (1) the strain, stock, and age of recommended eggs, (2) the number of replicate and/or repeat experiments, (3) development of a definition of a positive result, and (4) development of a recommended protocol. The Group also agreed that there were no proprietary components of the method. Dr. Green stated that the response to the basis for any modification to the original HET-CAM test method protocol needed to be developed by the Panel.

3.0 Substances Used for Previous Validation Studies of the HET-CAM Test Method

Dr. Green continued that the Group felt that the type and number of substances tested the studies evaluated in the HET-CAM BRD was adequate. The group also stated that it was difficult to determine if coding procedures were appropriate. They noted that not enough information provided was in all studies to make a full assessment. The Group then proposed that as long as the quality and multiplicity of data sources was sufficient to draw conclusions, coding did not matter.

4.0 In Vivo Reference Data Used for an Assessment of Test Method Accuracy

Dr. Green then discussed the *in vivo* data that used in the HET-CAM BRD analysis. The Group proposed that the *in vivo* test method protocols used to generate data used in BRD were appropriate and that interpretation of *in vivo* rabbit eye test results was correct. There was concern expressed by the Panel that the 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. The Group agreed that since original study records were not available for any of the reports evaluated, data quality could not be determined. However, the Group felt that an evaluation of the results could be made and the

quality of the studies otherwise appeared to be adequate. The Group stated that not all studies evaluated for the HET-CAM test method identified the studies as being conducted in compliance with GLP guidelines. The Group noted that there was not a large database on human ocular injury data and that most of the available information originates from accidental exposure. The Group recommended that COLIPA and the International Life Sciences Institute (ILSI) be consulted for information on human eye irritation databases to assess current ocular hazard classification scheme adequacy in protecting human health. The Group stated that the potential variability of the *in vivo* rabbit data had not been adequately discussed in the HET-CAM BRD.

The Panel discussed the use of the phrase "spirit of GLP". The appropriate use of the phrase and the definition of such phrase were discussed. Several Panel members indicated that the use of the phrase generally indicated that most, but not all, GLP guidelines were complied with during the course of conducting a study. The items within the guidelines that may not have been complied with each time may not be the same. However, the lack of compliance with specific items should be identified.

5.0 HET-CAM Test Method Data and Results

Dr. Green stated that the Group noted that the test method protocols used to generate data were adequately described in the BRD. The Group recommended including a description of which type of irritation score (IS(A) or IS(B)) was evaluated by each study in BRD Section 5.4 Additionally, the Group stated that data generated by Gautheron et al. (1994) may be useful in the BRD development. The Group recognized that the lack of consistent evaluation methods complicated BRD evaluations, but that the IS(B) appeared to be the optimal approach. It was proposed that data censoring could be an issue for the method. The Group stated that the BRD documents use of coding and GLP guidelines were adequately discussed and that there was insufficient information on lot-to-lot consistency in studies reviewed.

The Panel discussed a proposal in Section 5.3 that indicated that censoring was an issue for the HET-CAM test method. It was originally proposed that censoring methods would permit use of partial data obtained from studies not fully completed and that development of methods to use partial data should be considered. Several Panel members were unclear as to the intent of the statement. Dr. Lovell indicated that the statement could either refer to making use of incomplete studies or using a "meta-analysis" to combine across a number of different studies to bring them all into sort of a similar framework. There was concern by Panel members that the language was vague and the meaning of the statement was unclear. Panel members attempted to reword the statement to increase clarity. Panel members then proposed to remove the text from the document. A show of hands of Panel members indicated support to remove the text of concern.

The Panel then discussed the statement in Section 5.3 that stated that the approaches used in the evaluation of data were appropriate. Panel members expressed concern that the statement indicated that the Panel members accepted the statistical analysis and non-statistical approaches of evaluation of the data. Members indicated that there were alternative ways to

analyze the data. Initially it was proposed to delete the statement from the document. Upon further consideration, Panel members proposed to add text that indicated:

> *That 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.*

There was unanimous agreement on the modification.

6.0 HET-CAM Test Method Accuracy

Dr. Green continued to discuss the accuracy calculations described in the BRD. The Group stated that it was essential to identify structural classes the test method works well for and poorly for. The Group also recommended replacing the term “accuracy” with “concordance” or “agreement” since HET-CAM accuracy was not being evaluated directly against human data but against the *in vivo* test. The Group stated that the limitations of the analysis were discussed in the BRD. The Group also noted that there was *in vivo* data variability and in cases where false positives and negatives were noted, variability of the *in vivo* responses should be reviewed. The Group agreed that data interpretations were sufficiently described in the BRD and the organization of the document identified the IS(B) method as being the best in identifying most ocular corrosives and severe irritants. The Group noted that the use of a standardized test method protocol was needed to produce more interpretable data. Additionally, the Group recommended development of a table with non-accepted studies (HET-CAM BRD Section 9.0) to evaluate outcomes of these studies.

The Panel discussed the issue of replacing the term "accuracy" with “concordance” or “agreement”. The Panel noted that this issue was relevant to all the test method BRDs. Dr. Stokes stated that the ICCVAM submission guidelines state:

> Accuracy is defined as the closeness of agreement between a test method result and an accepted reference value. The term is often used interchangeably with concordance.

The term “concordance” is defined, in the same document as,:

> The proportion of all chemicals tested that are correctly classified as positive or negative. The term is often used interchangeably with accuracy

Dr. Stokes stated that to ensure consistency in the documents, the term accuracy was used in these reports. Dr. Scala requested discussion on whether the recommendation of replacing the term “accuracy” with another term should be carried to all the BRDs or deleted. A topic of Panel discussion was to whether the term concordance was more descriptive of the analyses presented in the BRDs. Dr. Stokes pointed out that the definition of accuracy states that the closeness of agreement can be between a test method result and a separate reference value. Additional discussion on the topic followed. A vote was taken as to whether the text

should be changed. The proposal to change the text was voted down by a show of eight hands to twelve.

7.0 HET-CAM Test Method Reliability (Repeatability/Reproducibility)

The Group noted that the rationale for substance selection was primarily based on data availability and the quality of the *in vivo* data was a limitation of all studies used in the BRD. The Group noted that even though there were no intralaboratory repeatability and reproducibility evaluations due to lack of data, it should not be a roadblock for use. The Group indicated that the topic of interlaboratory variability was well addressed in the BRD. However, the use of %CV was not an optimal approach and that non-parametric evaluations would be preferred. Additionally, there were several general items for revision noted. The Group agreed that there was an absence of historical data. The Group recommended the use of data from non-accepted studies (BRD Section 9.0) as a source of control data. The Group agreed with the BRD that appropriate recommendations were made for the selection and use of positive and negative controls and that the effect of protocol changes is unknown without having more standardized studies with measures of variability. The Group recommended that a video on the method and scoring be developed to increase test method transferability.

The Panel discussed the potential use of non-parametric analyses to evaluate interlaboratory variability. The Panel noted that the %CV assessment has limitations when evaluating a narrow range of scores. It was recommended to revise the text to indicate that alternative approaches for measuring agreement could be used for evaluating reproducibility. The remainder of the Panel agreed to the revision.

8.0 HET-CAM Test Method Data Quality

Dr. Green stated that the Group agreed with all the items in this section of the BRD. They stated that failure to use GLP guidelines by studies may have had qualitative impact only on borderline classifications. The Panel recommended including coding information (BRD Section 3.4) in Appendix B2.

9.0 Other Scientific Reports and Reviews

Dr. Green stated that the Group agreed that the BRD provided a comprehensive review of available publications made in BRD and adequately and completely provided the conclusions of the publications. The Group noted that it was useful to have information on why the studies were excluded from the evaluation in BRD Sections 6.0 and 7.0. The Group proposed that criteria for data acceptance could be relaxed to allow more studies to be evaluated in the BRD and that information on whom was contacted for additional data and who did and did not respond be included in the document. The Group also proposed that a preferred list of compounds be generated for distribution to companies and a request for additional data on these compounds be provided.

The Panel questioned what types of data were being requested. Dr. Tice (NICEATM) responded that the request was for any kind of data (e.g., *in vivo* human or rabbit data, *in vitro* data) that could be used for evaluation in the BRDs.

10.0 Animal Welfare Considerations

The Group stated that additional discussion on some issues was needed. The Group proposed to define the test as "*ex vivo*" and not "*in vitro*". The Group also proposed that this section should discuss the tiered testing strategy that animals would only be used to confirm negative response, and that HET-CAM should be considered a "partial replacement". The Group also proposed that additional discussion should state: (1) no new animal testing should be conducted, (2) the low false negative rate has the advantage of reducing the exposure of animals to severe irritants and corrosives, and (3) any additional optimization should focus on decreasing the false negative rate.

The Panel discussed the terms "*ex ovo*" and "*in vitro*". It was indicated that the term "*in vitro*" was used broadly to encompass "*ex ovo*". Dr. Stokes stated that subcategories of "*in vitro*" could be made, like "*ex ovo*" and perhaps the subcategories could be included into the glossary. There was no dissent by the Panel on that proposal. The Panel proposed to revise the first proposal to state that in the interest of consistency between the BRDs that the term "*in vitro*" should be used.

11.0 Practical Considerations

Dr. Green stated that the Group agreed with all the statements made in the BRD.

The Panel discussed that use of a training video on the technical aspects of the assay could be applicable to all the test methods being discussed. It was recommended that the language be modified to state that the Panel recommended the use of a training video and development of training approaches in the application of this test method. The Panel agreed that similar language should be included in all the test methods being evaluated at the meeting.

12.0 Proposed HET-CAM Test Method Recommendations

The Group discussed the proposal of using a modified HET-CAM test method protocol to confirm positive results obtained in the proposed, standardized HET-CAM test method protocol provided in the BRD. Discussion as to what type of modifications should be made to the test method protocol to identify potential false positives followed. Inclusion of *in vitro* test method data with various test concentrations into the BRD was proposed. It also was proposed that the report be modified to state that HET-CAM data exists to evaluate the use of a lower concentration and such information and analysis of the data should be included into the BRD.

The Panel then discussed revising or including language similar to the ICE test method report to indicate that even though the Panel recommended the current version of the HET-CAM test method could be used to classify substances as corrosives and/or severe irritants that the

Panel also recommends that the test method could be optimized and validated. The Panel discussed the proposal but determined that the current language in the HET-CAM report was clear since it indicated that the test method, except for test method procedures, were appropriate. The consensus agreement was to leave the language as it currently read.

Panel Vote on the HET-CAM Report

Dr. Scala asked if any individuals on the Panel had a conflict of interest with regard to this method. No hands were raised.

Dr. Scala asked for a vote on whether the Panel concurred with the HET-CAM information and results as presented. There was a show of 19 hands concurring with the presentation with three abstentions.

Dr. Scala asked if any of the three abstentions were for reasons of conflict of interest or for minority opinions. Dr. Stephens stated that his abstention was because of a minority opinion. His minority opinion was the same as for the ICE report regarding whether or not additional animal testing is discouraged or encouraged. Dr. Theran stated that his abstention was because of a minority opinion and was related to the use of "accuracy" in the document. Dr. Theran's opinion was that an *in vitro* test could be more accurate in representing the human experience and yet be not in agreement with the *in vivo* rabbit eye test and therefore did not agree with the use of the term "accuracy" when comparing it to the rabbit eye test. Dr. Yasuo Ohno, stated that the text for Section 8.1 was unclear. He felt that the sentence should state that the lack of GLP compliance would be compensated by using coded samples and appropriate data handling.

Dr. Stokes asked for clarification as to whether those with dissenting views agreed with the rest of the recommendations and conclusions except for the specific item that they expressed a minority opinion on. Dr. Stephens noted that he also dissented on the use of the term "accuracy" and that would be part of his minority opinion. However, there were no other issues that he dissented on. Dr. Theran stated that he did not dissent with any other issues related to the document. Dr. Ohno stated that he did not dissent with any other issues related to the document.

Dr. Fox stated that he believed that they should discuss the topic that Dr. Ohno presented in his minority opinion and potentially modify the statement. Dr. Scala stated that the topic would re-opened for discussion for no more than three minutes. Dr. Spielmann stated that he agreed with the statement, as proposed to be revised by Dr. Ohno. Dr. Scala requested that a Panel member propose that even though the vote was closed that it be re-opened and use Dr. Ohno's language as replacement for the language was present. Dr. Fox so moved. Dr. Scala asked for a show of hands to indicate concurrence with the text as changed. There were no dissenting votes.

Public Comment Session

1. Dr. Dan Marsman (Procter & Gamble)

Dr. Marsman discussed that the *in vivo* data are always requested alongside the *in vitro* methods but they weren't directly called for as part of this assay review and as such there wasn't a complete submission of *in vivo* data. Dr. Marsman stated that he believed that this hampered part of the discussion of these *ex vivo* methods and that submission and inclusion of that *in vivo* data would alter the interpretation of some of the *in vitro* results.

Dr. Marsman stated he believed it was critical that the *in vivo* methodology, its human equivalent, and its quality be reviewed to credibly evaluate the *in vivo* method and then the *ex vivo* method. Since the methods were being evaluated for their utility and classification and labeling for regulatory purposes all relevant *in vivo* data, specifically data from the LVET, should be included. Dr. Marsman stated that the LVET test method represented a minor modification of the traditional Draize method and that it has been accepted for classification and labeling purposes in some regulatory contexts. He stated that the minor modifications of the method yielded improvements. These improvements included refinement of pain and distress in the *in vivo* assay and the relevance to the human experience.

Dr. Marsman stated that all methods should be evaluated in the context for which they are being proposed and the historical LVET data and its associated *in vitro* data sets should be included in the evaluation. Inclusion of such data would likely alter the sensitivity and specificity scores of some of the *ex vivo* methods evaluated. Dr. Marsman concluded that the individual raw historical data for the LVET and the compiled data on the LVET as well as some of the mechanistic information on the pathophysiology, histopathology of the ocular toxicity is data that could further be submitted.

2. Dr. Rodger Curren (IIVS)

Dr. Curren requested that the Panel consider, when drafting the final Expert Panel report, strive for comparability in language between the BRDs for each test method of these as well as clarity of the final conclusions for each evaluated test method. Dr. Curren stated that there were a number of statements made in the Panel presentation on HET-CAM that dealt with many of the same problems or issues that were discussed in the ICE test method. However, Dr. Curren stated that the discussion of them was in many cases was more positive for HET-CAM than for ICE. Therefore, Dr. Curren requested that similar terms be used for similar ideas between the different test methods.

Dr. Curren than noted that Dr. Stokes previously stated that institutional animal care and use committees would have a significant role in how these tests are used. Dr. Curren stated that if at the end of the Panel deliberations there was a positive conclusion, such as stating that the test method could be used as the first stage of a tier-testing process and if it finds the positive results the material can be so labeled, that a statement should be included earlier in the document.

V. REFERENCE SUBSTANCES FOR USE IN VALIDATION STUDIES

Primary Reviewers: Drs. Ih Chu, Sidney Green, Yasuo Ohno, Robert Peiffer

Summary of the Recommended Reference Substances in each BRD

Dr. Allen (NICEATM) presented a summary of the list of reference substances included in each BRD. He summarized the selection criteria for reference substances as outlined in the ICCVAM Submission Guidelines (ICCVAM 2003), which states that to the extent possible, reference substances should:

- Be representative of the range of responses for which the proposed test method is expected to cover (i.e., nonirritants and mild, moderate, or severe eye irritants)
- Represent the classes of chemicals for which the proposed test method is expected to be used.
- Have produced high quality results (i.e., produced in a GLP compliant study) in the reference test method and/or the species of interest
- Have well-defined chemical composition
- Be readily available
- Not be associated with excessive hazard or prohibitive disposal costs

Dr. Allen proceeded to describe that the reference list covers the range of anticipated responses, based on irritancy classification according to the GHS. He highlighted the limitation of the available database that only a fraction of the substances for which *in v*ivo data had been obtained were also commercially available. Also described were the GHS Category I subcategories that have been developed by NICEATM. These subcategories were established to delineate among severe irritants/corrosives based on the type and severity of lesion upon which an irritancy classification was assigned and are as follows:

- Subcategory 4: substances that induce a corneal opacity of 4 at any time in at least one out of three animals.
- Subcategory 3: substances with a positive response, based on mean, within the first three days in two out of three animals, and a persistent lesion (i.e., a response that is severe, as well as persistent).
- Subcategory 2: substances with a positive response, based on mean, within the first three days in two out of three animals that was reversible (i.e., it was not persistent).
- Subcategory 1: substances classified as severe based only on a positive response at day 21.

Dr. Allen then detailed the five different *in vivo* data sources for the substances included in the reference list (the CTFA, the ECETOC, the EPA – TSCA, the FDA, and the Japanese National Institute of Health Sciences).

Finally, Dr. Allen summarized the 25 different chemical classes and 30 different product classes represented among the list. He also indicated the range of responses represented by each chemical and product class, based on GHS ocular irritancy classification.

Proposed Panel Recommendations on Reference Substances

12.4 Recommended Reference Substances for Validation Studies

Dr. Scala summarized the conclusions and recommendations of the primary reviewers of the list of reference substances. He indicated that one person from each of the test groups was included as a primary reviewer of the reference substances.

Dr. Scala then stated that the list of recommended substances is comprehensive in that the three major groups of products to which the eye is exposed (i.e., industrial chemicals, pharmaceuticals, cosmetics) are represented. He also stated that 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. Dr. Scala recognized that the selection of reference substances is in part limited by the availability of *in vivo* reference data. He then detailed the following comments and recommendations for the list:

- The current list has entirely too many substances and is unwieldy.
- Surfactants are over-represented and correspond to an area where the panel can make selective recommendations.
- The list appears to have too few inorganic substances; more should be added to the list if feasible.
- 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; this information should be removed from the list.
- Colored substances that might interfere with the observation of the endpoints should not be included.

Dr. Scala then described an approach to determine the most appropriate numbers and types of substances that should be included in the reference list. He described a two-staged study design to validation studies. During the first stage, a small number of substances from a wide range of chemical classes and spanning the range of severe irritancy should be tested among several laboratories to assess reliability. He stated that substances selected for this stage should:

- have an applicable pre-existing *in vivo* database
- cover a broad range of chemical classes that are representative of substances that are most likely to come in contact with the eye (e.g., 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)
- encompass the range of GHS Category 1 responses (i.e., GHS Category 1 subcategories described above)
- include a reasonable range of molecular weights, but no formulations, prototypes or products should be included

- include only liquid substances as these represent the majority of chemicals in the "real world" that will come in contact with the eye (using only liquids minimizes the inclusion of additional variables in the first stage of validation)

If deemed adequately reliable, an expanded set of substances would be tested in a second stage that would include multiple representatives of each chemical classes, diverse physicochemical characteristics, and the full range of irritancy responses to assess accuracy. Substances included in this stage should include:

- multiple representatives from each chemical class
- multiple representatives from each GHS Category 1 subcategory
- within each chemical class, compounds of different physical properties (solubility, molecular weight, pH) where feasible

Dr. Scala noted that for all validation studies, Material Safety Data Sheets (MSDS) for the recommended substances should be provided (e.g., a coded MSDS), in concert with a prestudy safety briefing

During the Panel discussion, Dr. Stitzel voiced concern over including a large proportion of Category 1 chemicals that were classified based only on a persistent response, given the fact that these "less severe" substances represent the area for which the Draize test is most variable. Dr. Spielmann concurred and stated that no substances classified based on a single animal should be included. Dr. Allen noted that some of these substances may have been classified based on all three animals showing the same response, which could potentially alleviate some of these concerns of variability.

Dr. Spielmann also voiced concern over the fact that the function of the list was not clear. He questioned if all substances were intended for use, or only a subset. He stated that the list should be preceded by a statement of precisely what the list was to be used for. Dr. Stokes responded by citing Section 12.4 of each BRD and the fact that a statement is included that the list has multiple purposes. He stated that for a full validation study, a large set of substances would be used. However, for performance standards, a smaller subset of the list could be used, and for proficiency chemicals, which a laboratory can use to demonstrate its proficiency with a test method, even a further subset of the reference chemicals in performance standards could be used.

Dr. Stephens stated that with regard to the BCOP, this validation effort might be satisfied by the submission of additional historical data. Therefore, he believed that with respect to BCOP, there may be a way to avoid a complete validation study.

In reference to the proposed two-staged validation study, Dr. Stitzel recalled earlier discussions among the Panel in which they concluded that validation against a grading system (i.e., classification) was not appropriate. She stated that a new method should be evaluated against a ranking of the severity of each chemical, rather than trying to evaluate new methods based on grading systems, which are very difficult even for the *in vivo* test to get right every time. Therefore she questioned why the substances were apparently still being selected based on a grading system. Dr. Scala responded by summarizing the selection

criteria used to generate the outline of the types of substances recommended by the Panel. He stated that the initial intention was to select from the list of the 50 most common industrial chemicals (which would serve as a surrogate for exposure). However, *in vivo* data were available for only five of these chemicals and therefore a list of the most common chemical classes was constructed from which a list of substances could be derived.

Dr. Stokes raised the issue of whether substances known to induce severe lesions, *in vivo*, in the eyes of humans should be included, even in the absence of rabbit data. The Panel agreed that such substances should be included.

Dr. Spielmann noted that in order to compare the different tests, it would be nice if a few standard positive controls or reference chemicals were available that could be used across the tests, so that a comparison of the efficiency of the tests could be conducted. He noted that there has been no agreement on what is acceptable as a positive control. Dr. Tice responded that each BRD proposed the type of positive control that would be useful. He noted that the positive control should be a severe, but on the very borderline or just slightly above the borderline between severe and nonsevere, and that it be either liquid or solid, based on what is being tested. However, because historical positive-control data were not available, reproducibility of such substances were not known, and therefore it would be premature to identify a specific substance that should be used all the time.

Dr. Robert Peiffer queried if it was ICCVAM's intention to use the Panel's recommendations to compile a final list of substances. Dr. Stokes responded that the intention is indeed to recommend a list, taking into consideration the Panel's advice, as well as advice received in public comments and will also receive in public comments on the Panel's report. Dr. Scala clarified that the Panel was not voting on approval of the list, rather on the outline of how to revise the list.

Panel Vote on the Recommended Reference Substances Report

A vote of concurrence on this section was taken and all members except Dr. Stephens agreed. Dr. Stephens's dissention was that additional validation studies involving prospective testing of chemicals in rabbits was not necessary.

Public Comment Session

1. Sadhana Dhruvakumar (PETA)

Ms. Dhruvakumar began by expressing her opinion that the discussion over the validation status of the methods (during the BCOP test method discussion) highlighted the apparent confusion among the Panel as to their ultimate charge in reviewing these methods. She believed that a statement that the validation status has been met should have been used with regard to the BCOP test method.

She then made reference to the reported data gaps for each of the test methods, and that these gaps may have been the result of inclusion that were too restrictive. She also pointed out that

other data sources had been mentioned during the course of the meeting, and therefore a sufficiently diligent search for the data had not been conducted.

Finally, Ms. Dhruvakumar commented on the evaluation of the underprediction rate of the Draize test conducted by Dr. Haseman. She stated that there was a lot of apparent bias in the study design. For example, assuming homogeneity within chemical classes was an assumption she stated would bias the test towards a more favorable outcome. She also stated that the evaluation only looked at intraexperimental variability. She closed by stating that due to the variability in the *in vivo* test, the calculated underprediction rate was a minimum at best, and PETA does not agree with the analysis.

2. Rodger Current (IIVS)

Dr. Curren opened by recognizing the amount of effort that was required of the Panel in reviewing the BRDs. He stated that he hoped that Panel's efforts could be used to yield the greatest potential value of outcome. He asked the practical question of whether it was the Panel's recommendation that these methods were to be used by industry prior to conducting an animal test. He declared if that indeed was the Panel's intention, then they in fact are indicating that the methods are "valid" for that purpose, and they should use this language in their recommendations. He indicated that the term "useful" was ambiguous and only caused confusion about whether the methods should actually be used or not.

Adjournment

Literature Citations

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. 8:307-310.

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.

CPSC. 1995. Test for eye irritants. 16CFR1500.42. Available: http://www.access.gpo.gov/nara/cfr/waisidx_00/16cfr1500_00.html [accessed 7 December 2005].

Draize J, Woodard G, Calvery H. 1944. Methods for the study of irritation and toxicity of substances applied topically to the skin and mucous membranes. J Pharm Exp Ther 82:377-390.

EPA. 1998. Health Effects Test Guideline, OPPTS 870.2400 Acute Eye Irritation. EPA 712-C-98-195. Washington, DC: U.S. Environmental Protection Agency.

EPA. 1996. Label Review Manual: 2nd Ed. EPA737-B-96-001. Washington, DC: U.S. Environmental Protection Agency.

EU. 2004. Manual of Decisions for Implementation of the Sixth and Seventh Amendments to Directive 67/548/EEC on Dangerous Substances (Directives 79/831/EEC and 92/32/EEC). European Chemicals Bureau, IHCP, European Commission DG-JRC, report n. EUR 20519 EN, updated version of July 2004..

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(3):381-392.

Gettings S, Lordo R, Hintze K, Bagley D, Casterton P, Chudkowski M, Curren RD, Demetrulias JL, Dipasquale LC, Earl LK, Feder PI, Galli CL, Glaza SM, Gordon VC, Janus J, Kurtz, Marenus KD, Moral J, Pape WJW, Renskers KJ, Rheins LA, Roddy MT, Rozen MG, Tedeschi JP, Zyracki J. 1996. The CTFA evaluation of alternatives program: An

evaluation of in vitro alternatives to the Draize primary eye irritation test. (Phase III) surfactant-based formulations. Food Chem Toxic 34:79-117.

Guerriero FJ, Seaman CW, Olsen MJ, Guest R, Whittingham A. 2004. Retrospective assessment of the rabbit enucleated eye test (REET) as a screen to refine worker safety studies. [Abstract No. 1282]. Toxicol Sci (The Toxicologist Supplement). 78(1-S).

Hackett RB, McDonald TO. 1991. Eye irritation. In: Advances in Modern Toxicology: Dermatoxicology. 4th ed. (Marzulli F, Maibach H, eds.). Washington, DC: Hemisphere Publishing Corporation, 749-815.

INVITTOX. 1992. Protocol No. 47: HET-CAM Test. Available: http://ecvam-sis.jrc.it/invittox/published/indexed_47.html. [accessed 18 February 2004].

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.

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.

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, 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 - Cut and Ocular Toxicol 14:179-195.

UN. 2003. Globally Harmonized System of Classification and Labelling of Chemicals (GHS). New York & Geneva: United Nations Publications. Available: http://www.unece.org/trans/danger/publi/ghs/officialtext.html. [accessed 26 August 2004].

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

January 11-12, 2005
Expert Panel Meeting to Assess the Current Validation Status of *In Vitro* Test Methods for Identifying Ocular Corrosives and Severe Irritants: Bovine Corneal Opacity and Permeability (BCOP), Hen's Egg Test – Chorioallantoic Membrane (HET-CAM), Isolated Chicken Eye (ICE) and Isolated Rabbit Eye (IRE)

"These Summary Minutes have been read and approved by the Chair of the Expert Panel Meeting on the Evaluation of the Validation Status of *In Vitro* Test Methods for Identifying Ocular Corrosives and Severe Irritants, as certified below."

_____ _____
Dr. Robert Scala Date
Panel Chair

_____ _____
Dr. William Stokes Date
Panel Executive Secretary

[This Page Intentionally Left Blank]

APPENDIX A4
SUMMARY MINUTES FROM EXPERT PANEL TELECONFERENCE ON SEPTEMBER 19, 2005

[This Page Intentionally Left Blank]

Department of Health and Human Services
National Institutes of Health
National Institute of Environmental Health Sciences
Interagency Coordinating Committee on the Validation of Alterative Methods (ICCVAM)
Expert Panel Teleconference

Summary Minutes of the Expert Panel Teleconference to Evaluate Revised Analyses and Proposed Reference Substances

Introduction

A public teleconference of an independent Expert Panel was convened on September 19, 2005 to evaluate several i*n vitro* ocular irritation test methods. The purpose of this meeting was to evaluate (1) revised accuracy and reliability analyses of four *in vitro* test methods proposed for detecting ocular corrosives and severe irritants, and (2) a revised list of proposed reference substances for validation studies on *in vitro* test methods for identifying ocular corrosives and severe irritants. The four *in vitro* test methods under consideration were the (1) Bovine Corneal Opacity and Permeability (BCOP) assay, (2) Hen's Egg Test--Chorioallantoic Membrane (HET-CAM), (3) Isolated Rabbit Eye (IRE) assay, and (4) Isolated Chicken Eye (ICE) assay. The teleconference was organized by ICCVAM and the National Toxicology Program (NTP) Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM), and sponsored by the National Institute of Environmental Health Sciences (NIEHS) and the NTP.

The following scientists served on the Expert Panel:

- Robert Scala, Ph.D., (Panel Chair), Tucson, Arizona, United States
- Sally S. Atherton, Ph.D., Professor, Medical College of Georgia, Augusta, Georgia, United States
- Roger Beuerman, Ph.D., Professor, Louisiana State University, New Orleans, Louisiana, United States
- June Bradlaw, Ph.D., International Foundation for Ethical Research, Rockville, Maryland, United States
- Ih Chu, Ph.D., Health Canada, Ottawa, Canada
- Henry Edelhauser, Ph.D., Professor, Emory University, Atlanta, Georgia, United States
- Donald Fox, Ph.D., Professor, University of Houston, Houston, Texas, United States
- James Freeman, Ph.D., Lab Director, ExxonMobil Biomedical Sciences, Inc., Annandale, New Jersey, United States
- Sidney Green, Ph.D., A.T.S., Graduate Professor, Howard University, Washington, DC, United States
- Frederick Guerriero, M.S., Senior Occupational Toxicologist, GlaxoSmithKline, King of Prussia, Pennsylvania, United States

- 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, United States
- Hiroshi Itagaki, Ph.D., Deputy Director of JSAAE, Manager of Alternative Section, 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 JSAAE, National Institute of Health, Japan
- Robert Peiffer, D.V.M., D.A.C.V.O., Senior Investigator, Merck Research Laboratories, West Point, Ohio, United States
- Lionel Rubin, V.M.D., D.A.C.V.O., Emeritus Professor of Ophthalmology, University of Pennsylvania, Philadelphia, Pennsylvania, United States
- 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, United States
- Katherine Stitzel, D.V.M., Consultant, West Chester, Ohio, United States
- 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, United States
- Scheffer Tseng, M.D., Ph.D., Director, Ocular Surface Research and Education Foundation, Miami, Florida, United States
- Philippe Vanparys, Ph.D., Senior Research Fellow, Johnson and Johnson, Belgium

The following ICCVAM agency representatives participated in the teleconference:

- Dr. Robert Bronaugh, (Ocular Toxicity Working Group - OTWG), U.S. Food and Drug Administration
- Dr. Karen Hamernik, (OTWG), U.S. Environmental Protection Agency
- Dr. Abigail Jacobs, (OTWG), U.S. Food and Drug Administration

The following additional members of the ICCVAM Ocular Toxicity Working Group (OTWG) participated in the teleconference:

- Ms. Donnie Lowther, U.S. Environmental Protection Agency
- Dr. Jill Merrill, U.S. Food and Drug Administration

The following members of the NICEATM Staff participated in the teleconference:

- Dr. David Allen, Integrated Laboratory Systems, Inc.
- Mr. Bradley Blackard, Integrated Laboratory Systems, Inc.
- Mr. Thomas Burns, Integrated Laboratory Systems, Inc.
- Dr. Jeffrey Charles, Integrated Laboratory Systems, Inc.
- Dr. Neepa Choksi, Integrated Laboratory Systems, Inc.

- Ms. Linda Litchfield, Integrated Laboratory Systems, Inc.
- Ms. Debbie McCarley, National Institute of Environmental Health Sciences
- Dr. Raymond Tice, National Institute of Environmental Health Sciences
- Mr. James Truax, Integrated Laboratory Systems, Inc.

The following members of the public participated in the teleconference:

- Dr. Rodger D. Curren, Institute for *In Vitro* Sciences, Inc.
- Dr. Jean Domoradzki, Dow Chemicals
- Dr. John Harbell, Institute for *In Vitro* Sciences, Inc.
- Dr. Pauline M. McNamee, The European Cosmetic, Toiletry, and Perfumary Association (COLIPA)
- Dr. Brooke McManus, Rosenhect
- Dr. Pat Phibbs, Bureau of National Affairs News
- Dr. Kristina Thayer, National Institute for Environmental Health Sciences
- Dr. Sherry L. Ward, Physicians Committee for Responsible Medicine

The purpose of this meeting was to evaluate (1) revised accuracy and reliability analyses of the four *in vitro* test methods proposed for detecting ocular corrosives and severe irritants and (2) a revised list of proposed reference substances for validation studies on *in vitro* test methods for identifying ocular corrosives and severe irritants. The Expert Panel was asked to evaluate an addendum to these four ocular draft background review documents (BRDs) prepared by NICEATM (Available: http://iccvam.niehs.nih.gov).

Call to Order and Introductions

Dr. Robert Scala, Panel Chair, called the teleconference of the Expert Panel (Panel) to order at 9:10 a.m. He asked Mr. Blackard to take attendance of the panel members. NICEATM staff members were then asked to introduce themselves. The operator then provided the names and affiliations of the members of the public that were participating. Dr. Scala stated that there were no requests made to make a public comment during the teleconference.

Dr. Raymond Tice, Designated Federal Official, read the Statement of Conflict of Interest and explained policies and procedures regarding confidentiality and avoidance of conflict of interest, as follows: *"As a Special Emphasis Panel, the members of the ocular expert panel serve as individual scientists and not as representatives of any organization. Each member is to exercise judgment prior to any meeting as to whether a potential conflict of interest might exist relative to agenda topics or concepts for discussion by the Expert Panel due to his or her occupational affiliation, professional activity or financial interest. Should there be a potential conflict of interest, they will be handled in accordance with departmental policies and requirements."*

Dr. Scala asked if any member of the Panel had any potential conflicts of interest. None of the Panel members declared a conflict of interest.

Overview of the Performance Reanalysis

Dr. Tice (Deputy Director, NICEATM, NIEHS) provided a brief overview of the process that led to the public teleconference of the Expert Panel. On November 1, 2004, NICEATM made available four BRDs that provided information and data about the current validation status of four *in vitro* test methods for the ability to detect ocular corrosives and severe irritants. The four test methods evaluated in the BRDs were the BCOP assay, the HET-CAM assay, the IRE assay, and the ICE assay. The analyses in the BRDs were based on published literature and data submitted in response to a 2004 *Federal Register* (*FR*) notice. An Expert Panel was convened on January 11-12, 2005 to assess the validation status of these four *in vitro* test methods to identify ocular corrosives and severe irritants. Public comments at the meeting indicated that additional data were available that had not been provided in response to earlier *FR* notices. The Expert Panel recommended that the additional data be requested and that a reanalysis of the performance of each *in vitro* test method be conducted, where appropriate.

Dr. Tice stated that in response to this recommendation a *FR* notice was published on February 28, 2005 requesting *in vitro* ocular toxicity and corresponding *in vivo* ocular toxicity data be submitted to NICEATM for inclusion in the reanalysis. In response to this notice, data was received for the BCOP, HET-CAM, and ICE test methods. Dr. Tice discussed other additional changes and analyses that were conducted and incorporated in the reanalysis of test method performance.

Dr. Tice stated that the proposed reference list that was included in each draft BRD was revised based on:

- recommendations from the Panel
- additional *in vivo* data received for approximately 300 substances
- reclassification of substances based on clarification of ocular toxicity classification rules
- reclassification of the chemical class of a substance based on Medical Subject Headings (MeSH) chemical classes

The BRD addendum was released on July 26, 2005 for public review and comment.

Dr. Tice stated that the purpose of the teleconference was for the Expert Panel to address the following questions:

- For each test method, is the information provided in the addendum appropriate for inclusion in the accuracy and reliability analysis? Are there any errors or omissions that should be corrected?
- Based on the revised accuracy and reliability analysis, does the new information provide the basis for any change in the Panel's conclusions and recommendations made at the January 11-12, 2005 meeting?
- Is the revised list of proposed reference substances, sufficiently valid and complete for use in *in vitro* test methods to evaluate ocular corrosives and severe irritants?

Dr. Tice stated that the Panel's recommendations and public comments would be considered by ICCVAM when making their final recommendations. These recommendations would be provided to the public and U.S. Federal agencies for consideration.

Organization of the Panel Review

The remainder of the teleconference was devoted to Panel discussion and answering the three questions stated by Dr. Tice.

Prior to the presentations and discussions by each of the four groups, a NICEATM staff member provided a brief summary of the information contained in the test method specific BRD reanalysis addendum, including updated accuracy and reliability analyses for the test method.

Each Panel group discussed its draft response for each of the questions. After each presentation, the entire Panel discussed the draft positions and offered additional comments and suggestions. The Panel Chair summarized the discussion for each question and sought consensus from the Panel on the topic.

I. IRE TEST METHOD EVALUATION

Primary reviewers: Drs. James Freeman (Group Chair), Sally Atherton, David Lovell, Yasuo Ohno, Horst Spielmann, Peter Theran

Summary of the IRE Reanalysis

Dr. Neepa Choksi (NICEATM) stated that Mr. Jim Truax (NICEATM) had conducted the re-analysis on the IRE test method. No additional information and/or *in vitro* and comparative *in vivo* data were received in response to the *FR* notice. However, an additional analysis was conducted (based on a recommendation by the Panel at the January 11-12, 2006 meeting) where a positive score in studies that did not use all four endpoints used by Guerriero et al. (2004) were combined with the results from Guerriero et al. This new database was referred to as the "Expanded Dataset".

Dr. Choksi noted that, for comparative purposes, the addendum contained the results presented in the IRE BRD as well as the results from the reanalysis. For the United Nations Globally Harmonized System (GHS; UN 2003), the accuracy of the IRE test method based on the 38 substances tested by Guerriero et al. (2004) was 79%, with a false positive rate of 30% and false negative rate of 0%. For the Expanded Dataset, the accuracy was 68%, with a false positive rate of 68% and a false negative rate of 0%. Dr. Choksi noted that for the Expanded Dataset, only substances classified as positive based on their response in the IRE test method from the other studies were included into the analysis. This potential bias should be considered when reviewing the data.

Dr. Choksi then reviewed the accuracy analyses conducted for various subgroups of the data (based on chemical class, properties of interest, pH, and irritancy subcategories). Limiting the chemical class evaluation to those with five or more substances per chemical class, the classes with the highest rate of overprediction were ketones, esters, and alcohols. In addition, liquids tended to have a higher false positive rate than solids.

Dr. Choksi stated that, as in the draft IRE BRD, analyses on intralaboratory repeatability and intralaboratory reproducibility could not be conducted due to the lack of data. Dr. Choksi then reviewed the change to the qualitative interlaboratory reproducibility analysis, for the GHS classification system, noting that changes in the U.S. Environmental Protection Agency (EPA 1996) and European Union (EU 2001) classification are similar.

Proposed Panel Recommendation for the IRE Test Method

Dr. Freeman reviewed the questions provided to the Panel and stated what the draft Expert Panel comments were.

Is the information provided in the Addendum to the November 2004 IRE BRD appropriate for inclusion in the accuracy and reliability analysis?

The Panel stated that the answer was yes. There was no dissention or disagreement noted.

Are there any errors or omissions that should be corrected?

Dr. Freeman stated that the draft response was as follows:

> The Panel agreed that there were no errors or omissions. The Panel recognized and supported the rationale for excluding some substances from the evaluation based on lack of adequate *in vivo* rabbit eye test data (i.e., severe ocular irritancy/corrosivity classification based solely on skin corrosivity, pH extremes, etc., or no classification feasible based on eye test data provided to NICEATM). While the pH and/or dermal corrosive effects of a test substance are utilized as substitutes for animal eye irritation data for the purposes of ocular hazard classification, the goal of this evaluation was to determine whether the four *in vitro* test methods can be used to predict the outcome of the *in vivo* rabbit eye test for the same test substance. Therefore, including data on pH extremes and/or dermal corrosivity (in the absence of *in vivo* rabbit eye test data) was judged to be inappropriate due to the uncertainty of its performance in predicting the *in vivo* rabbit eye test outcome. In addition, the Panel recommended that text be included in the final BRD to underscore the fact that, where such information was available, data derived from scientifically acceptable *in vivo* rabbit eye tests terminated based on humane endpoints (e.g., severe discomfort) were included in the accuracy and reliability analysis.

Panel discussion followed in the appropriateness of excluding chemicals that were classified as severe ocular irritants or corrosives on the basis of dermal corrosivity and/or pH extremes from the accuracy and reliability reanalysis. The Panel concurred with the decision to limit the evaluation to substances where appropriate *in vivo* ocular data were available.

Based on the revised accuracy and reliability analysis for the IRE test method for identifying ocular corrosives and severe irritants, does this new information provide the basis for any changes in the Panel's conclusions and recommendations made at the January 11-12, 2005 meeting?

The Panel agreed that there was no basis for changes to the original conclusions and recommendations[1]. There were no dissenting opinions from the Panel.

Panel Vote on the IRE Report

The Panel Chair asked the Panel members if there were any dissenting opinions. No dissenting opinions were indicated.

[1] *Editor's Note:* At the Expert Panel meeting in January 2005, the Panel recommended that the current version of the IRE test method appeared "to be capable of identifying ocular corrosives/severe irritants in a tiered-testing strategy." However, they noted that the available data were "too limited to allow for an adequate judgment of its accuracy and reliability."

A Panel member asked if it was acceptable to compare rabbit eye data to *in vitro* data given concerns with the reliability of the *in vivo* test. The Panel Chair stated that such a question, while an excellent and important point, was currently outside the charge of the Panel. The Panel Chair noted that evaluation of the question and efforts related to such an evaluation were important. However, these are future efforts as the work on evaluation of alternative ocular test method progress. The Panel Chair stated that a statement would be incorporated recommending such activities into the final report.

Dr. Scala concluded this discussion with a vote among the Panel members. He noted that everyone on the Panel was in agreement with the conclusions and recommendations for the IRE test method.

II. ICE TEST METHOD EVALUATION

Primary Reviewers: Drs. Robert Scala, Roger Beuerman, June Bradlaw, Wallace Hayes, Robert Peiffer, Nancy Flournoy

Summary of ICE Reanalysis

Dr. David Allen (NICEATM) stated that additional ICE and *in vivo* data were received. The total database for evaluating ICE performance increased from 92 to 144 substances.

Dr. Allen noted that the overall accuracy, when compared to the GHS classification system, increased from 82% to 83%. The false positive rate decreased from 10% to 8% and the false negative rate increased from 40% to 50%. The numbers observed for the GHS classification system were comparable to those obtained for the EPA and EU classification system.

Dr. Allen then reviewed the accuracy analyses conducted for various subgroups of the data (based on chemical class, properties of interest, pH, and irritancy subcategories). Limiting the chemical class evaluation to those with five or more substances per chemical class, the class with the highest rate of overprediction was alcohols. Surfactants and solids had the highest rate of underprediction. An analysis based on the pH indicated that basic substances tended to be underpredicted. Furthermore, substances that produce persistent lesions (lesions that last at least 21 days) also tend to be underpredicted by the ICE test method.

Dr. Allen stated that, based on newly received data, assessments of intralaboratory repeatability and reproducibility could be conducted. An intralaboratory repeatability coefficient of variation (CV) analysis evaluation, on each ICE endpoint, indicated that corneal thickness was repeatable (CV ranged from 1% to 6%). An intralaboratory reproducibility CV analysis also indicated that the corneal thickness measurement was generally reproducible (CV < 7%). Dr. Allen reviewed the changes to the qualitative interlaboratory reproducibility analysis, for the GHS classification system; he then reviewed the historical negative and positive control results received.

Proposed Panel Recommendations for the ICE Test Method

The Panel Chair opened the discussion on the questions.

Is the information provided in the Addendum to the November 2004 ICE BRD appropriate for inclusion in the accuracy and reliability analysis?

The Panel stated that the answer was yes.

Are there any errors or omissions that should be corrected?

Dr. Scala stated that the language developed for the IRE test methods was appropriate for the ICE test method. The draft response was as follows:

The Panel agreed that there were no errors or omissions. The Panel recognized and supported the rationale for excluding some substances from the evaluation based on lack of adequate *in vivo* rabbit eye test data (i.e., severe ocular irritancy/corrosivity classification based solely on skin corrosivity, pH extremes, etc., or no classification feasible based on eye test data provided to NICEATM). While the pH and/or dermal corrosive effects of a test substance are utilized as substitutes for animal eye irritation data for the purposes of ocular hazard classification, the goal of this evaluation was to determine whether the four *in vitro* test methods can be used to predict the outcome of the *in vivo* rabbit eye test for the same test substance. Therefore, including data on pH extremes and/or dermal corrosivity (in the absence of *in vivo* rabbit eye test data) was judged to be inappropriate due to the uncertainty of its performance in predicting the *in vivo* rabbit eye test outcome. In addition, the Panel recommended that text be included in the final BRD to underscore the fact that, where such information was available, data derived from scientifically acceptable *in vivo* rabbit eye tests terminated based on humane endpoints (e.g., severe discomfort) were included in the accuracy and reliability analysis.

Based on the revised accuracy and reliability analysis for the ICE test method for identifying ocular corrosives and severe irritants, does this new information provide the basis for any changes in the Panel's conclusions and recommendations made at the January 11-12, 2005 meeting?

The Panel agreed that there was no basis for changes to the original conclusions and recommendations[2]. There were no dissenting opinions from the Panel on the topic.

Panel Vote on the ICE Report

Dr. Scala concluded this discussion with a vote among the Panel members. He noted that everyone on the Panel was in agreement with the conclusions and recommendations for the ICE test method.

[2] *Editor's Note:* At the Expert Panel meeting in January 2005, the Panel recommended that the current version of the ICE test method could be used for identifying ocular corrosives/severe irritants in a tiered-testing strategy with the following limitations: (a) alcohols tend to be overpredicted, (b) surfactants tend to be underpredicted, (c) solids and insoluble substances may be problematic since they may not come in adequate contact with the corneal surface.

III. BCOP TEST METHOD EVALUATION

Primary reviewers: Drs. Kathy Stitzel (Group Chair), Ih Chu, Henry Edelhauser, Hiroshi Itagaki, Lionel Rubin, Scheffer Tseng, David Lovell

Summary of BCOP Reanalysis

Dr. Allen stated that additional *in vivo* data was received. The total database for evaluating BCOP performance increased from 120 to 147 substances.

Dr. Allen noted that the overall accuracy, when compared to the GHS classification system, increased from 79% to 81%. The false positive rate increased from 19% to 20% and the false negative rate decreased from 24% to 16%.

Dr. Allen then reviewed the accuracy analyses conducted for various subgroups of the data (based on chemical class, properties of interest, pH, and irritancy subcategories). Limiting the chemical class evaluation to those with five or more substances per chemical class, the classes with the highest rate of overprediction were ketones and alcohols. Solids had the highest rate of underprediction. Removal of ketones, solids, and alcohols from the database increased the accuracy to 92%, decreased the overall false negative rate to 0% and the overall false positive rate to 12%. Furthermore, substances that produce persistent lesions (lesions that last at least 21 days) also tend to be underpredicted by the BCOP test method.

Dr. Allen stated that intralaboratory repeatability and reproducibility and quantitative interlaboratory analyses were not affected by the new data received. Dr. Allen reviewed the changes to the qualitative interlaboratory reproducibility analysis, for the GHS classification system, using three different studies. Dr. Allen noted that the results from the analyses were similar to those previously presented in the BCOP BRD.

Proposed Panel Recommendations for the BCOP Test Method

The Panel Chair opened the discussion on the questions.

Is the information provided in the Addendum to the November 2004 BCOP BRD appropriate for inclusion in the accuracy and reliability analysis?

The Panel stated that the answer was yes.

Are there any errors or omissions that should be corrected?

Dr. Stitzel stated that the language developed for the IRE test method was appropriate for the BCOP test method. The draft response was as follows:

> The Panel agreed that there were no errors or omissions. The Panel recognized and supported the rationale for excluding some substances from

the evaluation based on lack of adequate *in vivo* rabbit eye test data (i.e., severe ocular irritancy/corrosivity classification based solely on skin corrosivity, pH extremes, etc., or no classification feasible based on eye test data provided to NICEATM). While the pH and/or dermal corrosive effects of a test substance are utilized as substitutes for animal eye irritation data for the purposes of ocular hazard classification, the goal of this evaluation was to determine whether the four *in vitro* test methods can be used to predict the outcome of the *in vivo* rabbit eye test for the same test substance. Therefore, including data on pH extremes and/or dermal corrosivity (in the absence of *in vivo* rabbit eye test data) was judged to be inappropriate due to the uncertainty of its performance in predicting the *in vivo* rabbit eye test outcome. In addition, the Panel recommended that text be included in the final BRD to underscore the fact that, where such information was available, data derived from scientifically acceptable *in vivo* rabbit eye tests terminated based on humane endpoints (e.g., severe discomfort) were included in the accuracy and reliability analysis.

Based on the revised accuracy and reliability analysis for the BCOP test method for identifying ocular corrosives and severe irritants, does this new information provide the basis for any changes in the Panel's conclusions and recommendations made at the January 11-12, 2005 meeting?

The Panel agreed that there was no basis for changes to the original conclusions and recommendations[3]. There were no dissenting opinions from the Panel on the topic.

Panel Vote on the BCOP Report

Dr. Scala concluded this discussion with a vote among the Panel members. He noted that everyone on the Panel was in agreement with the conclusions and recommendations for the BCOP test method.

[3] *Editor's Note:* At the Expert Panel meeting in January 2005, the Panel recommended that the current version of the BCOP test method appeared to be capable of identifying ocular corrosives/severe irritants in a tiered-testing strategy with the limitation that further optimization and validation are necessary before alcohols, ketones, and solids can be assessed with this method.

IV. HET-CAM TEST METHOD EVALUATION

Primary reviewers: Drs. Shayne Gad, Donald Fox, Martin Stephens, Frederick Guerriero, Sidney Green, Philippe Vanparys, Nancy Flournoy

Summary of HET-CAM Reanalysis

Dr. Choksi stated that additional *in vivo* data was received relating to Gilleron et al. (1996, 1997) and Spielmann et al. (1996). Dr. Choksi stated that due to the additional data received, several additional analyses were conducted and presented in the addendum.

Dr. Choksi noted that the overall accuracy, when compared to the GHS classification system, for the various IS(B) analysis methods (IS(B)-10 and IS(B)-100; using the decision criteria of Luepke [1985]) ranged from 53% to 68%. The false positive rates ranged from 33% to 61% and the false negative rates ranged from 15% to 33%. Dr. Choksi stated the results for each analysis method were comparable for all three regulatory hazard classification systems. Dr. Choksi then reviewed the accuracy analyses conducted for various subgroups of the data (based on chemical class, properties of interest, pH, and irritancy subcategories).

Dr. Choksi stated that assessments of intralaboratory repeatability and reproducibility could be conducted, using the additional data. An intralaboratory repeatability and interlaboratory CV analysis on each HET-CAM endpoint evaluation indicated that the coagulation endpoint was the lowest of the three endpoints evaluated. Dr. Choksi reviewed the changes to the qualitative and quantitative interlaboratory reproducibility analyses, for the GHS classification system. Dr. Choksi then reviewed the historical negative and positive control results received.

Proposed Panel Recommendations for the HET-CAM Test Method

The Panel Chair opened the discussion on the questions.

Is the information provided in the Addendum to the November 2004 HET-CAM BRD appropriate for inclusion in the accuracy and reliability analysis?

The Panel stated that the answer was yes.

Are there any errors or omissions that should be corrected?

Dr. Gad stated that the language developed for the IRE test method was appropriate for the HET-CAM test method. The draft response was as follows:

> The Panel agreed that there were no errors or omissions. The Panel recognized and supported the rationale for excluding some substances from the evaluation based on lack of adequate *in vivo* rabbit eye test data (i.e., severe ocular irritancy/corrosivity classification based solely on skin corrosivity, pH extremes, etc., or no classification feasible based on eye test

data provided to NICEATM). While the pH and/or dermal corrosive effects of a test substance are utilized as substitutes for animal eye irritation data for the purposes of ocular hazard classification, the goal of this evaluation was to determine whether the four *in vitro* test methods can be used to predict the outcome of the *in vivo* rabbit eye test for the same test substance. Therefore, including data on pH extremes and/or dermal corrosivity (in the absence of *in vivo* rabbit eye test data) was judged to be inappropriate due to the uncertainty of its performance in predicting the *in vivo* rabbit eye test outcome. In addition, the Panel recommended that text be included in the final BRD to underscore the fact that, where such information was available, data derived from scientifically acceptable *in vivo* rabbit eye tests terminated based on humane endpoints (e.g., severe discomfort) were included in the accuracy and reliability analysis.

Based on the revised accuracy and reliability analysis for the HET-CAM test method for identifying ocular corrosives and severe irritants, does this new information provide the basis for any changes in the Panel's conclusions and recommendations made at the January 11-12, 2005 meeting?

The Panel agreed that the IS(B) analysis method (using the decision criteria of Luepke 1985) was not sufficiently predictable to use for identifying ocular corrosives and severe irritants. This conclusion was different from the Panel's conclusions and recommendations made at the January 11-12, 2005 meeting, where the test method was proposed to be sufficiently predictable to use for identifying ocular corrosives and severe irritants. There were no dissenting opinions from the Panel on the topic.

Panel Vote on the HET-CAM Report

Dr. Scala concluded this discussion with a vote among the Panel members. He noted that everyone on the Panel was in agreement with the conclusions and recommendations for the HET-CAM test method.

V. REFERENCE SUBSTANCES FOR USE IN VALIDATION STUDIES

Summary of the Recommended Reference Substances

Dr. Tice presented a summary of the list of reference substances included in the BRD Addendum. He summarized the selection criteria for reference substances were intended to:

- represent a range of ocular responses
- represent a range of chemical/product classes
- represent a range of known or anticipated mechanisms or modes of action
- high quality *in vivo* rabbit eye test method studies exist for these substances
- have a well-defined chemical composition
- have been tested at a defined concentration and purity
- be readily available

Based on the recommendation from the Expert Panel meeting in January, several changes were made to the original list. Overall:

- The number of inorganics on the list was increased from 2 to 11
- The list now included 10 human ocular corrosives/irritants, despite the lack of corresponding individual rabbit eye test data
- Formulations were removed from the list
- The number of surfactants on the list were decreased from 12 to 7

The Panel also recommended that the total number of proposed reference substances be decreased from 89 substances. Dr. Tice noted that the number of substances needed to evaluate the accuracy of an alternative test method depends on several factors including (1) the range of possible responses that the test method is expected to be able to measure, (2) the diversity of the known or anticipated mechanisms or modes of action that are involved in producing a toxic response, and (3) the number of chemical/physical classes and physicochemical properties that the test method is expected to be able to evaluate. Dr. Tice stated that a preliminary statistical evaluation indicates that several hundred substances could potentially be required to evaluate the accuracy of a test method with a high level of confidence

Dr. Tice went on to state that for the detection of ocular corrosives and severe irritants, the list of substances needs to include substances that:

- induce very severe responses within a relatively short period, as well as those where the response is delayed
- adversely affect the cornea, iris, and/or conjunctiva
- induce persistent and/or non-persistent lesions
- represent a diverse population of chemical classes and physicochemical properties

Dr. Tice stated that to meet these needs and to address the recommendations of the Panel, the list was increased from 89 substances to 122 substances. The proposed list includes 79 GHS Category 1 substances, 10 substances classified based on human data, 28 GHS Category 2

substances, and 15 GHS nonirritant substances. There were 34 chemical classes and 29 product classes represented on the list.

Proposed Panel Recommendations on Reference Substances

The Panel Chair opened the discussion to the Panel.

Panel discussion followed regarding assurance that substances with the same purity and quality as those on the proposed reference list would be evaluated by all testing laboratories. The Panel stated that purity of the test substance should be given and that impurities, to the extent possible, should be noted and quantified.

The Panel discussion also indicated that the list of proposed reference substances was too large if the list is intended to be the minimum number of substances that should be used for validation of a new test method. Panel discussion then followed on the proposal to revise the proposed reference substances list so that mechanisms of toxic action would be represented instead of chemical classes. Therefore, chemical classes with similar mechanisms of action could be combined into a single class to decrease the number of substances on the proposed reference substances list.

Adjournment 11:30

Literature Citations

EPA. 1996. Label Review Manual: 2nd Ed. 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.

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. Toxic In Vitro 11:641-644.

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. Toxic In Vitro 10:431-446.

Guerriero FJ, Seaman CW, Olsen MJ, Guest R, Whittingham A. 2004. Retrospective assessment of the rabbit enucleated eye test (REET) as a screen to refine worker safety studies. [Abstract No. 1282]. Toxicol Sci (The Toxicologist Supplement). 78(1-S).

Luepke N. 1985. Hen's egg chorioallantoic membrane test for irritation potential. Food Chem Toxic 23:287-291.

Spielmann H, Liebsch M, Kalweit S, Moldenhauer F, Wirnsberger T, Holzhütter 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.

UN. 2003. Globally Harmonized System of Classification and Labelling of Chemicals (GHS). New York & Geneva: United Nations Publications. Available: http://www.unece.org/trans/danger/publi/ghs/officialtext.html. [accessed 26 August 2004].

September 19, 2005
Expert Panel Teleconference to Assess the Current Validation Status of *In Vitro* Test Methods for Identifying Ocular Corrosives and Severe Irritants: Bovine Corneal Opacity and Permeability (BCOP), Hen's Egg Test – Chorioallantoic Membrane (HET-CAM), Isolated Chicken Eye (ICE) and Isolated Rabbit Eye (IRE)

"These Summary Minutes have been read and approved by the Chair of the Expert Panel Teleconference on the Evaluation of the Validation Status of *In Vitro* Test Methods for Identifying Ocular Corrosives and Severe Irritants, as certified below."

_____ _____
Dr. Robert Scala Date
Panel Chair

_____ _____
Dr. Raymond Tice Date
Designated Federal Official