7.0 IRE TEST METHOD RELIABILITY

An assessment of test method reliability (intralaboratory repeatability and intra- and inter-laboratory reproducibility) is an essential element of any evaluation of the performance of an alternative test method (ICCVAM 2003). Repeatability refers to the closeness of agreement between test results obtained within a single laboratory when the procedure is performed on the same substance under identical conditions within a given time period (ICCVAM 1997 2003). Intralaboratory reproducibility refers to the determination of the extent to which qualified personnel within the same laboratory can replicate results using a specific test protocol at different times. Interlaboratory reproducibility refers to the determination of the extent to which different laboratories can replicate results using the same protocol and test chemicals, and indicates the extent to which a test method can be transferred successfully among laboratories. A reliability assessment includes reviewing the rationale for selecting the substances used to evaluate test method reliability, a discussion of the extent to which the substances tested represent the range of possible test outcomes and the properties of the various substances for which the test method is proposed for use, and a quantitative and/or qualitative analysis of repeatability and intra- and inter-laboratory reproducibility. In addition, measures of central tendency and variation are summarized for historical control data (negative, vehicle, and positive), where applicable.

Due to the lack of quantitative IRE test method data for replicate experiments within an individual laboratory, an evaluation of the intralaboratory repeatability and reproducibility of the IRE test method could not be conducted. However, multilaboratory qualitative and quantitative IRE test data were available for a collaborative study by the CEC (1991) and a validation study conducted by Balls et al. (1995). Three laboratories participated in the CEC (1991) collaborative study and four laboratories participated in the Balls et al. (1995) validation effort. In the CEC (1991) study, each substance tested was assigned a EU classification (R41, R36, or nonirritant [EU 2001]) based on *in vivo* rabbit eye test results. However, due to the lack of individual rabbit *in vivo* Draize scores, a reliability assessment for the CEC (1991) study using the GHS (UN 2003) or EPA (EPA 1996) classification criteria was not possible. The Balls et al. (1995) data were used for an evaluation of the interlaboratory reproducibility of the IRE test method according to the GHS (UN 2003), EPA (EPA 1996), and EU (EU 2001) classification systems.

7.1 Selection Rationale for the Substances Used to Evaluate the Reliability of the IRE Test Method

The quality of a reliability evaluation depends on the extent to which the substances tested adequately represent the range of physicochemical characteristics and response levels that the test method must be capable of evaluating. The only sources of data for conducting an assessment of interlaboratory reproducibility were the CEC (1991) collaborative study and the EC/HO validation study reported on by Balls et al. (1995).

The CEC (1991) collaborative pilot study evaluated the reproducibility of the IRE test method using 21 substances. These substances were provided by FRAME via Aldrich Chemical Company Limited and were selected to cover a full range of eye irritation

potential. A single supplier provided the substances and each chemical originated from a single batch. All of the substances were > 96% pure. The authors also intended the list of substances to be representative of a variety of chemical structures and representative of currently used industrial chemicals. Furthermore, the authors chose substances with *in vivo* data to which a EU risk phrase could be assigned and, where possible, those that had been used in previous validation studies.

The Balls et al. (1995) study evaluated the performance and reproducibility of the IRE test method using 60 "substances" (i.e., there were 52 different substances with four substances tested at two different concentrations and two substances tested at three concentrations, for a total of 60 possible ocular irritation outcomes). To be selected for inclusion in this study, the substances had to be single chemicals (no mixtures) available at high purity and stable when stored, and the reference *in vivo* rabbit eye data had to have been generated since 1981 according to OECD TG 405 following GLP guidelines. In addition, substances were selected to ensure an adequately diverse group of physicochemical characteristics and levels of irritancy severity. One substance (thiourea) was tested *in vitro* in the IRE assay but, due to its excessive toxicity *in vivo*, excluded from the comparison of *in vitro* and *in vivo* test results.

7.2 Analyses of Repeatability and Reproducibility

7.2.1 Quantitative and Qualitative Assessments of Intralaboratory Repeatability

Generally, analyses of intralaboratory repeatability have included approaches such as:

- a coefficient of variation (CV) analysis, which is a statistical measure of the deviation of a variable from its mean (e.g., Holzhütter et al. 1996)
- analysis of variance (ANOVA) methods (e.g., Holzhütter et al. 1996; ASTM 1999)

Due to the lack of available IRE test data for replicate enucleated rabbit eyes within individual experiments and for experiments conducted on the same substance under exactly the same conditions, an evaluation of the intralaboratory repeatability of the IRE test method could not be conducted.

7.2.2 Quantitative and Qualitative Assessments of Intralaboratory Reproducibility

Generally, analyses of intralaboratory reproducibility have included approaches such as:

- a CV analysis, which is a statistical measure of the deviation of a variable from its mean (e.g., Holzhütter et al. 1996)
- ANOVA methods (e.g., Holzhütter et al.[1996; ASTM 1999)

Due to the lack of available IRE test data for experiments conducted multiple times in the same laboratory, an evaluation of IRE test method intralaboratory reproducibility could not be conducted.

7.2.3 Assessment of Interlaboratory Reproducibility

Generally, analyses of interlaboratory variability have included approaches such as:

- determination of the extent of concordance among laboratories in assigning the same regulatory classification for a particular substance (e.g., Holzhütter et al. 1996)
- a CV analysis, which is a statistical measure of the deviation of a variable from its mean (e.g., Holzhütter et al. 1996)
- ANOVA methods (e.g., Holzhütter et al. 1996; ASTM 1999)
- bivariant scatter diagrams/correlation analyses for pairs of laboratories to assess the extent possibility of divergence (e.g., Holzhütter et al. 1996)

Several of the studies discussed in **Section 6.0** included interlaboratory data for at least a subset of the substances evaluated. Using this data, the ability of the IRE test method to reproducibly identify ocular corrosives and severe irritants versus nonsevere irritants and nonirritants was evaluated using two approaches.

In the first approach, a qualitative assessment of reproducibility was conducted. In this evaluation, the individual laboratory *in vitro* ocular irritation classification for each substance was used to evaluate the extent of agreement among the participating laboratories in their ability to identify ocular corrosives/severe irritants versus nonsevere irritants/nonirritants. The reliability of the IRE test method was assessed separately for each study (i.e., publication) with multiple laboratory data reviewed in **Sections 4.0** and **5.0**.

Substances classified, based on IRE test data, as corrosive/severe irritants or nonsevere irritants/nonirritants were further classified by their *in vivo* rabbit eye test results, as determined within the GHS (UN 2003), EPA (EPA 1996), and EU (EU 2001) classification systems. Because the focus of this reliability assessment is on the interlaboratory reproducibility of IRE in identifying corrosives/severe irritants versus nonsevere irritants/nonirritants, considerable variability could exist among laboratories in their classification of substances as nonsevere irritants or nonirritants. For example, three laboratories could classify a chemical as a nonirritant and one laboratory could classify the same chemical as a moderate irritant. Within this analysis, this distribution of classification calls would be considered as 100% agreement between laboratories.

In the second approach, a quantitative assessment of reproducibility was determined. CVs where laboratory scores were available for substances tested were reported or determined. The reproducibility of the IRE test method was assessed for studies (i.e., publication) reviewed in **Sections 4.0** and **5.0** where individual testing laboratory data was available.

7.2.3.1 *Interlaboratory Reproducibility of Hazard Classification Category Using the GHS Classification System*

For this classification system (UN 2003), one study could be used to assess the interlaboratory reproducibility of the IRE test method (Balls et al. 1995). The four participating laboratories in this EC/HO validation study were in agreement in regard to the ocular irritancy classification (corrosive/severe irritant or nonsevere irritant/nonirritant) of 35 (59%) of the 59 substances tested.

As shown in **Table 7-1:**

Table 7-1 Interlaboratory Variability of Balls et al. (1995) for Substances Classified as Ocular Corrosives/Severe Irritants or Nonsevere Irritants/Nonirritants Using the GHS Classification System

Classification (*in vivo*/*in vitro*)[1]	Number of Substances	Number of Testing Laboratories	Substances with 100% Agreement Among Laboratories (%)	Substances with 75% Agreement Among Laboratories (%)	Substances with 50% Agreement Among Laboratories (%)
+/+	14	4	14 (100)	0 (0)	0 (0)
+/-	9	4	5 (56)	4 (44)	0 (0)
-/+	20	4	8(40)	3 (15)	9 (45)
-/-	14	4	6 (43)	8 (57)	0 (0)
?/-	1	4	1 (100)	0 (0)	0 (0)
?/+	1	4	1 (100)	0 (0)	0 (0)
TOTAL	**59**	**4**	**35 (59)**	**15 (25)**	**9 (15)**

[1]A "+" indicates that the substance was assigned an overall classification of corrosive or a severe irritant (Category 1); a "-" indicates that the substance was assigned an overall classification of nonsevere irritant (Category 2A, 2B) or nonirritant; a "?" indicates that, due to the lack of appropriate *in vivo* data (e.g., (e.g., studies were terminated too early to assess reversibility of effects; insufficient dose volume), a GHS classification (UN 2003) could not be made. See **Section 6.1** for a description of the rules followed to classify the ocular irritancy of substances tested multiple times *in vitro*.

- All four participating laboratories agreed on the classification of 14 (100%) of the 14 substances that were GHS corrosives/severe irritants[1].
- Five (56%) of the nine substances classified according to the GHS based on *in vivo* rabbit eye data as corrosives/severe irritants were incorrectly classified by all four participating laboratories as nonsevere irritants (i.e., Category 2A and 2B irritants) or nonirritants whereas four of the nine substances (44%) had 75% agreement among the laboratories. The five substances incorrectly classified by all four laboratories were Captan 90 concentrate, dibenzoyl-L-tartaric acid, 2,5-dimethylhexanediol, 15% sodium lauryl sulfate, and sodium perborate tetrahydrate.
- Eight (40%) of the 20 substances classified according to the GHS based on *in vivo* rabbit eye data as nonsevere irritants or nonirritants were incorrectly classified by the four laboratories as corrosives or severe irritants. Of the 12 substances (60%) with discordant *in vitro* classification results among the four laboratories, three (15%) (ethyl acetate, iso-propanol, and methyl acetate) were incorrectly classified by three of the four laboratories and nine (45%) (acetone, 0.1% cetylpyridinium bromide, ethyl-2-methylacetoacetate, Fomesafen, Maneb, methylisobutylketone, n-octanol, polyethylene glycol 400, and toluene) were incorrectly classified by two of the four laboratories.

[1] As described in **Section 6.1**, the overall *in vitro* classification for each substance was determined based on the most frequent individual laboratory classification, or in the case of an even number of discordant responses, the most severe classification.

- All four laboratories agreed on the classification of six (43%) of the 14 substances classified as GHS nonsevere irritants/nonirritants. Of the eight substances (57%) with discordant classification results, all eight substances (ammonium nitrate, butyl acetate, 4-carboxybenzaldehyde, dibenzyl phosphate, 2,6-dichlorobenzoyl chloride, tetra-aminopyrimidine sulfate, 3% trichloroacetic acid, and Tween 20) were correctly classified by three of the four laboratories.
- Due to the lack of appropriate *in vivo* data (e.g., studies were terminated too early to assess reversibility of effects; insufficient dose volume), two of the 59 substances tested could not be classified according to the GHS classification scheme (UN 2003). All four laboratories were in agreement with the classification of one of these substances as a nonsevere irritant/nonirritant and of one substance as a corrosive/severe irritant.

7.2.3.2 *Interlaboratory Reproducibility of Hazard Classification Category Using the EPA Classification System*

The four participating laboratories in the EC/HO study (Balls et al. 1995) were in 100% agreement with the ocular irritancy classification (corrosive/severe irritant or nonsevere irritant/nonirritant) of 36 (61%) of the 59 substances tested. As shown in **Table 7-2:**

Table 7-2 Interlaboratory Variability of Balls et al. (1995) for Substances Classified as Ocular Corrosives/Severe Irritants or Nonsevere Irritants/Nonirritants Using the EPA Classification System

Classification (*in vivo*/*in vitro*)[1]	Number of Substances	Number of Testing Laboratories	Substances with 100% Agreement Among Laboratories (%)	Substances with 75% Agreement Among Laboratories (%)	Substances with 50% Agreement Among Laboratories (%)
+/+	18	4	18 (100)	0 (0)	0 (0)
+/-	7	4	4 (57)	3 (43)	0 (0)
-/+	20	4	8(40)	3 (15)	9 (45)
-/-	14	4	6 (43)	8 (57)	0 (0)
?/-	0	4	0 (0)	0 (0)	0 (0)
?/+	0	4	0 (0)	0 (0)	0 (0)
TOTAL	59	4	36 (61)	14 (24)	9 (15)

[1]A "+" indicates that the substance was assigned an overall classification of corrosive or a severe irritant (Category I); a "-" indicates that the substance was assigned an overall classification of nonsevere irritant (Category II, III) or nonirritant (category IV); a "?" indicates that, due to the lack of appropriate *in vivo* data (e.g., studies were terminated too early to assess reversibility of effects; insufficient dose volume), an EPA classification could not be made. See **Section 6.1** for a description of the rules followed to classify the ocular irritancy of substances tested multiple times *in vitro*.

- All four participating laboratories[2] agreed on the classification of eighteen (100%) of the 18 substances that were EPA (EPA 1996) corrosives/severe irritants.
- Four (57%) of the seven substances classified according to the EPA (EPA 1996) based on *in vivo* rabbit eye data as corrosives or severe irritants were incorrectly classified by the four participating laboratories as nonsevere irritants (i.e., Category II or III) or nonirritants (Category IV). Three substances (43%) were shown to have discordant *in vitro* classification results among the four participating laboratories (Captan 90 concentrate, 2,5-dimethylhexanediol, and sodium lauryl sulfate [15%]). These substances were incorrectly identified by three of the four laboratories.
- Eight (40%) of the 20 substances classified according to the EPA (EPA 1996) based on *in vivo* rabbit eye data as nonsevere irritants or nonirritants were incorrectly classified by the four laboratories as corrosives or severe irritants. Of the 12 remaining substances (60%), three substances (15%) (ethyl acetate, iso-propanol, and methyl acetate) were incorrectly classified by three of the four laboratories and nine substances (45%) (acetone, cetylpyridinium bromide, ethyl-2-methylacetoacetate, Fomesafen, Maneb, methylisobutyl ketone, n-octanol, polyethylene glycol 400, and toluene) by two of the four laboratories.
- Six (43%) of the 14 substances classified according to the EPA (EPA 1996) based on *in vivo* rabbit eye data as nonsevere irritants/nonirritants were correctly classified by all four laboratories. All eight substances (57%) with discordant classification results (ammonium nitrate, butyl acetate, 4-carboxybenzaldehyde, dibenzyl phosphate, 2,6-dichlorobenzoyl chloride, tetra-aminopyrimidine sulfate, 3% trichloroacetic acid, and Tween 20) were correctly classified by three of the four laboratories.

7.2.3.3 *Interlaboratory Reproducibility Based on In Vitro Irritancy Classification Relative to the In Vivo Classification Using the EU Classification System*

A reliability analysis of the IRE test method in terms of the EU classification system could be conducted for the CEC (1991) collaborative study and the Balls et al. (1995) validation study.

In the CEC (1991) collaborative study, the participating laboratories were in 100% agreement in regard to the ocular irritancy classification (corrosive/severe irritant or nonsevere irritant/nonirritant) of 17 (81%) of the 21 substances tested.

As shown in **Table 7-3:**

[2] As described in **Section 6.1**, the overall *in vitro* classification for each substance was determined based on the most frequent individual laboratory classification, or in the case of an even number of discordant responses, the most severe classification.

Table 7-3 Interlaboratory Variability of CEC Collaborative Study (1991) for Substances Classified as Ocular Corrosives/Severe Irritants or Nonsevere Irritants/Nonirritants Using the EU Classification System

Classification (*in vivo*/*in vitro*)[1]	Number of Substances	Number of Testing Laboratories	Substances with 100% Agreement Among Laboratories (%)	Substances with 67%[2] Agreement Among Laboratories (%)	Substances with 33%[3] Agreement Among Laboratories (%)
+/+	5	3	3 (60)	1 (20)	1 (20)
+/-	0	3	0 (0)	0 (0)	0 (0)
-/+	2	3	2 (100)	0 (0)	0 (0)
-/-	8	3	6 (75)	2(25)	0 (0)
?/-	2	2[5]	2 (100)	0 (0)	0 (0)
?/+	4	3	4 (100)[6]	0 (0)	0 (0)
TOTAL	21	3[6]	17 (81)	3 (14)	1 (5)

[1] A "+" indicates that the substance was assigned an overall classification of corrosive or severe irritant (Category R41); a "-" indicates that the substance was assigned an overall classification of nonirritant (Category R36); a "?" indicates that, due to the lack of appropriate *in vivo* data (i.e., insufficient dose volume), an EU classification (EU 2001) could not be made. See **Section 6.1** for a description of the rules followed to classify the ocular irritancy of substances tested multiple times *in vitro*.
[2] When two of three laboratories were concordant.
[3] When one of three laboratories was concordant.
[4] With the exception of the two (+/-) substances.
[5] Two of the three testing laboratories evaluated these two substances.
[6] One of the four substances was tested in two laboratories with severe classifications assigned.

- Three (60%) of five substances classified according to *in vivo* rabbit eye data as corrosives/severe irritants were identified correctly by all three laboratories[3]. One discordant substance (sodium dodecyl sulfate) was correctly classified by two of the three laboratories, and one (dibutyltin chloride) was correctly classified by one of three laboratories.
- Of the 21 substances evaluated, none were identified as false negative (i.e., as a corrosive/severe irritant *in vivo* and as a nonsevere irritant *in vitro*).
- Two of two substances (100%) were incorrectly classified as corrosives/severe irritants by all three laboratories (100%). There were no discordant substances.
- Six of eight (75%) substances were in complete agreement among laboratories for identification of nonsevere irritants/nonirritants. Two discordant substances (25%) (Brij 35 and 2-butoxyethylacetate) were identified as nonsevere irritants/nonirritants by two of the three testing laboratories.
- Both laboratories (only two of three laboratories tested these substances) agreed in the identification of two substances as nonsevere irritants/nonirritants (100%), although no *in vivo* classification could be assigned to these substances.

[3] As described in **Section 6.1**, the overall *in vitro* classification for each substance was determined based on the most frequent individual laboratory classification, or in the case of an even number of discordant responses, the most severe classification.

- All three laboratories agreed in the identification of four substances as severe irritants (100%), although no *in vivo* classification could be assigned to these substances.

Using the Balls et al. (1995) validation data set, the participating laboratories were in 100% agreement with the ocular irritancy classification (corrosive/severe irritant or nonsevere irritant/nonirritant) of 37 (63%) of the 59 substances tested. As shown in **Table 7-4:**

Table 7-4 Interlaboratory Variability of Balls et al. (1995) for Substances Classified as Ocular Corrosives/Severe Irritants or Nonsevere Irritants/Nonirritants Using the EU Classification System

Classification (*in vivo*/*in vitro*)[1]	Number of Substances	Number of Testing Laboratories	Substances with 100% Agreement Among Laboratories (%)	Substances with 75% Agreement Among Laboratories (%)	Substances with 50% Agreement Among Laboratories (%)
+/+	12	4	12 (100)	0 (0)	0 (0)
+/-	6	4	3 (50)	3 (50)	0 (0)
-/+	18	4	7(39)	2 (11)	9 (50)
-/-	12	4	6 (50)	6 (50)	0 (0)
?/-	6	4	4 (67)	2 (33)	0 (0)
?/+	5	4	5 (100)	0 (0)	0 (0)
TOTAL	59	4	37 (63)	13 (22)	9 (15)

[1]A "+" indicates that the substance was assigned an overall classification of corrosive or severe irritant (Category R41); a "-" indicates that the substance was assigned an overall classification of nonirritant (Category R36); a "?" indicates that, due to the lack of appropriate *in vivo* data (i.e., insufficient dose volume), an EU classification could not be made. See **Section 6.1** for a description of the rules followed to classify the ocular irritancy of substances tested multiple times *in vitro*.

- All four participating laboratories agreed on the classification of 12 (100%) of the 12 substances that were EU corrosives/severe irritants[4].
- Three (50%) of the six substances classified according to the EU (EU 2001) based on *in vivo* rabbit eye data as corrosives/severe irritants were incorrectly classified by all four laboratories as nonsevere irritants/nonirritants. Of the three substances (50%) with discordant *in vitro* classification results among the four participating laboratories, all three substances (Captan 90 concentrate, dibenzoyl-L-tartaric acid, and 2,5-dimethylhexanediol) were incorrectly classified by three of the four laboratories.
- Seven (39%) of the 18 substances classified according to the EU (EU 2001) based on *in vivo* rabbit eye data as nonsevere irritants/nonirritants were incorrectly classified by all four participating laboratories as corrosives/severe irritants. Of the 11 substances (61%) with discordant *in vitro* classification results among the four participating laboratories, two substances (44%), ethyl

[4] As described in **Section 6.1**, the overall *in vitro* classification for each substance was determined based on the most frequent individual laboratory classification, or in the case of an even number of discordant responses, the most severe classification.

acetate and methyl acetate, were incorrectly classified by three of the four laboratories and nine (50%) were incorrectly classified by two of the four laboratories (acetone,γ-butyrolactone, 0.1% cetylpyridinium bromide, ethyl-2-methylacetaoacetate, fomesafen, methylisobutylketone, n-octanol, polyethylene glycol 400, and toluene).

- All four participating laboratories agreed on the classification of six (50%) of the 12 substances classified as EU (EU 2001) nonsevere irritants/nonirritants. Three of the four laboratories were in agreement for the six substances (50%) with discordant *in vitro* classification results (ammonium nitrate, 4-carboxybenzaldehyde, dibenzyl phosphate, tetra-aminopyrimidine sulfate, 3% trichloroacetic acid, and Tween 20).
- Four of six (67%) of substances were classified *in vitro* as nonirritants by all four laboratories, but could not be classified *in vivo* due to lack of sufficient data. Two of the six (33%) were classified as nonsevere irritants/nonirritants *in vitro* by three of the four laboratories.
- Five of five (100%) substances were classified *in vitro* as corrosives/severe irritants by all four laboratories, but could not be classified *in vivo* due to lack of appropriate data.

7.2.3.4 *Common Chemical or Product Classes Among Substances with Discordant Interlaboratory Results Using the GHS, EPA, and EU Classification Systems*

In the CEC (1991) study, GHS and EPA classifications were not available due to lack of *in vivo* rabbit eye data. Using the EU (EU 2001) classification system, four discordant substances that were incorrectly classified *in vitro* had no commonality with respect to chemical or product class. There were no false negative or false positive discordant substances in this analysis.

Twenty-four, 23, and 22, substances, respectively, for the GHS (UN 2003), EPA (EPA 1996), and EU (EU 2001) classification systems exhibited interlaboratory differences in *in vitro* classification in the Balls et al. (1995) study. Six esters, four alcohols, three carboxylic acids, and three ketones exhibited discordant results. Four substances (Captan 90 concentrate [pesticide], dibenzoyl-L-tartaric acid, 2,5-dimethylhexanediol [pesticide], and 15% sodium lauryl sulfate [surfactant]) were consistently found in at least two of three classification systems to be underpredicted with 75% concordance among the laboratories. Three substances (ethyl acetate, methyl acetate [acetates/solvents], and iso-propanol [alcohol/solvent]) were consistently overpredictive in all three classification systems with 75% concordance between laboratories. Nine substances (acetone, 0.1% cetylpyridinium bromide, ethyl-2-methylacetoaceate, Fomesafen, Maneb, methyl isobutylketone, n-octanol, polyethylene glycol 400, and toluene) were consistently found to be overpredictive with 50% concordance among the testing laboratories in at least two of the three classification systems. Solvent (nonaqueous water miscible and nonmiscible) was the product class appearing most frequently among all of these discordant substances. Eight of the discordant substances belonged to this product class. Surfactants/soaps (3) and pesticides (4) were other product classes for which discordant results were observed.

7.2.4 Coefficient of Variation Analysis

7.2.4.1 *CEC Collaborative Study (CEC 1991)*

Mean endpoint values (i.e., corneal opacity, corneal swelling, and fluorescein retention at one and four hours; one laboratory used a 1.25 hour endpoint) for each substance tested were provided from each of the three laboratories participating in the CEC (1991) study. These values were used to calculate the SD and %CV values for each IRE test method endpoint for each substance to provide a quantitative assessment of interlaboratory variability (**Table 7-5**)

Mean and median %CV values for the entire dataset were also calculated to provide an assessment of overall variability. Traditionally, mean/median %CV values of less than 35% have been considered satisfactory for biologically based test methods (Fentem et al. 1998). For IRE, there is moderate interlaboratory variability for each of the four endpoints, with fluorescein retention at four hours representing the largest %CV (59%) and corneal swelling at four hours representing the lowest %CV value (33%). When only severe irritants (EU Category R41[5] [EU 2001], based on *in vivo* data) are considered, the interlaboratory variability is lower for all endpoints. Corneal swelling at 1.25 hours retains the highest variability (CV of 37%) and 4-hour corneal opacity the lowest (CV of 16%). It should be noted that this analysis was performed without using a correction factor to normalize corneal swelling values, a practice that has been suggested if different depth measuring devices were used among the different laboratories (Prinsen M, personal communication). The overall median CV of the 4-hour corneal swelling values was 70% (40/57.3) of the mean, whereas all other parameters ranged from 47% (28/58.9) to 81% (43.0/53.3) of their respective means. The overall median CV of the 4-hour corneal swelling for severe irritants was 100% (35.5/35.4) of the mean with the other parameters ranging from 83% (30.5/36.6) to 99% (15.4/15.5) of their respective means. These values suggest that efforts to increase the interlaboratory reproducibility of the test method might be warranted.

There do not appear to be physicochemical characteristics that are common to most of the substances with the most variable responses (defined as > 100% CV in any of the endpoints). All of the substances in the CEC study were tested as liquids (some were diluted to concentrations used in the *in vivo* studies). Of nine substances with significant variability in at least one endpoint, there were no obvious chemical or product classes that appear to be responsible for the variability. Four substances (sodium fluorescein, glycerol, triethanolamine, and n-hexane) had variability in more than one endpoint.

[5] GHS classification (UN 2003) was not available for this dataset.

Table 7-5 Quantitative Evaluation of the Interlaboratory Variability of the IRE Test Method (CEC 1991)

Substance	CS(1)				FR (1)		CO(2)	
	1.25 Hour	1.25 Hour	4 Hour	4 Hour	4 Hour	4 Hour	4 Hour	4 Hour
	Mean	(%CV)	Mean	(%CV)	Mean	(%CV)	Mean	(%CV)
Acetic acid	20.7	56	40.7	30	2.33	25	1.70	25
Brij 35	7.67	38	12.3	25	0.87	93	0.50	141
Benzalkonium chloride	40.3	31	82.7	38	2.67	22	3.00	0.00
Dimethylsulfoxide	8.00	66	11.7	95	1.33	87	0.50	141
Sodium fluorescein	2.33	138	4.70	173	0.67	172	0.00	0.00
Glycerol	3.33	92	5.33	43	0.33	175	0.40	141
Triacetin	2.67	43	0.67	172	0.00	0.00	0.00	0.00
Mercury chloride	19.0	30	76.0	35	2.50	28	2.40	24
Silver nitrate	14.0	7.1	16.7	40	1.00	100	1.75	20
Sodium hydroxide	38.7	19	67.3	22	3.00	0.00	3.00	0.00
Toluene	9.00	22	10.7	30	1.73	37	0.60	141
Triethanolamine	3.33	148	6.33	97	0.07	165	0.00	0.00
n-Hexane	4.00	132	7.00	108	0.33	175	0.00	0.00
Chloroform	17.0	60	30.7	47	3.00	0.00	1.70	25
2-Methoxy ethanol	12.7	54	42.7	7.2	2.67	22	2.40	24
n-Butanol	31.3	6.7	60.3	16	3.00	0.00	2.50	28
Acetaldehyde	12.3	21	34.7	20	2.93	3.9	1.25	28
2-Butoxy ethylacetate	10.0	20	23.0	68	1.67	35	0.95	7.4
Sodium dodecylsulfate	15.7	47	24.0	61	2.07	78	1.40	40
Dibutyltin chloride	11.0	26	29.5	41	2.00	0.00	1.00	0.00
Tributyltin chloride	22.3	63	97.0	36	2.53	20	2.10	6.7
Mean for All Substances	14.5	53.3	32.6	57.3	1.7	58.9	1.3	37.7
Median for All Substances		43.0		40.0		28.0		24.0
Range for All Substances	2.3-40	6.7-148	0.7-97	7.2-173	0-3.0	0.0-175	0-3.0	0-141
Mean for Severe Irritants (EU)	22.5	36.6	56.5	35.4	2.5	22.1	2.0	15.5
Median for Severe Irritants (EU)		30.5		35.5		21.0		15.4
Range for Severe Irritants (EU)	11-40	19-63	24-97	20-61	2.0-3.0	0-78	1.0-3.0	0 -40

CO = Corneal opacity; CS = Corneal swelling; FR = Fluorescein retention, SD = Standard deviation; %CV = Percent coefficient of variation

[1]Substances listed in bolded italics are classified *in vivo* as severe irritants (Category 1) according to GHS (UN 2003).

7.2.4.2 *Balls et al. (1995)*

Mean endpoint values (i.e., corneal opacity and corneal swelling at 1 and 4 hours) for each substance tested were provided from each of the four laboratories participating in the EC/HO study. These values were used to calculate the standard deviation and CV for each IRE test method endpoint for each substance to provide a quantitative assessment of interlaboratory variability (**Table 7-6**). Mean and median %CV values for the entire dataset were also calculated to provide an assessment of overall variability. Traditionally, mean/median %CV values of less than 35% have been considered satisfactory for biologically-based test methods (Fentem et al. 1998; ICCVAM 2003). For IRE, there is moderate interlaboratory variability for each of the four endpoints, with corneal opacity at 1 hour representing the largest %CV (84%) with a range spanning 0 to 200% and corneal swelling at 4 hours representing the lowest %CV (53%) with a range of 10 to 118%. When only severe irritants (GHS Category 1[6], based on *in vivo* data [UN 2003]) are considered, the interlaboratory variability is lower for all endpoints, although corneal opacity at 1 hour retains the highest variability (47%CV) with a range of 0 to 200% and 4-hour corneal swelling the lowest (37%CV) spanning a range of 11 to 118%. The overall median of the 4-hour corneal opacity values was 68% (43.4/63.79) of the mean, whereas all other parameters ranged from 89% (74.6/84.1) to 93% (49.7/53.47) of their respective means. The overall median of the 4-hour corneal opacity for severe irritants was 83% (33.6/40.5) of the mean with the other parameters ranging from 87% (40.6/46.6) to 96% (35.5/36.9) of their respective means. These values suggest that efforts to increase the interlaboratory reproducibility of the test method might be warranted.

There do not appear to be physicochemical characteristics that are common to most of the substances with the most variable responses (defined arbitrarily as > 100%CV) in any of the endpoints). Of the 36 substances with significant variability in at least one endpoint, 17 are solids (of 19 tested) and 19 are liquids (of 40 tested). However, there are some chemical classes that predominate among the variable results with seven acetates/esters (of 7 tested), six surfactants (of 12 tested), six acids (of 6 tested), three heterocyclic compounds (of 6 tested), three alcohols (of 7 tested), and three pesticides (of 4 tested) represented among the 36 substances. However, in the absence of a larger dataset, the significance of these findings is not clear.

[6] One of these substances (sodium lauryl sulfate, 15%) is classified as R36 according to EU (EU 2001). Two other substances (cetylpyridinium bromide, 6% and dibenzoyl-L-tartaric acid) were not classified according to EPA system due to inadequate *in vivo* data with which to follow the EPA-specific classification rules (EPA 1996). Therefore, substances classified as severe irritants according to the GHS system (UN 2003) were used for this subanalysis in order to include the largest dataset.

Table 7-6 Quantitative Evaluation of the Interlaboratory Variability of the IRE Test Method (Balls et al. 1995)

Substance	CO 1 Hour Mean	CO 1 Hour (%CV)	CO 4 Hour Mean	CO 4 Hour (%CV)	CS 1 Hour Mean	CS 1 Hour (%CV)	CS 4 Hour Mean	CS 4 Hour (%CV)
1-Naphthalene acetic acid[1]	0.25	200	0.90	114	11.98	73	13.7	64
1-Naphthalene acetic acid, Na salt	1.00	115	2.68	18	57.03	51	107.6	45
2,2-Dimethylbutanoic acid	2.75	18	2.74	12	33.58	17	68.0	22
2,5-Dimethylhexanediol	0.33	142	0.42	120	15.30	87	16.4	79
2,6-Dichlorobenzoyl chloride	0.75	67	1.90	32	8.53	103	21.1	56
2-Ethyl-1-hexanol	0.25	200	1.43	35	10.7	44	20.3	17
4-Carboxybenzaldehyde	0.25	200	0.43	119	6.20	56	13.0	70
Acetone	0.43	119	1.05	105	15.3	78	31.9	95
Ammonium nitrate	0.00	0	0.00	0	7.30	43	10.2	111
Benzalkonium chloride (1 %)	0.93	90	2.43	28.0	23.9	23.0	52.8	48
Benzalkonium chloride (10%)	1.67	28	2.50	23.0	36.4	50	73.1	43
Benzalkonium chloride (5%)	1.33	71	3.00	0.00	32.3	40	99.2	23
Dibenzoyl-L-tartaric acid	1.00	141	1.90	60.0	18.2	118	24.5	70
Captan 90 concentrate	0.75	128	1.01	98.0	6.50	80	18.7	51
Cetylpyridinium bromide (0.1%)	0.00	0.00	0.00	0.00	14.7	47.0	19.8	50
Cetylpyridinium bromide (10%)	0.83	106	1.92	43	17.9	36	43.5	68

Substance	CO 1 Hour Mean	CO 1 Hour (%CV)	CO 4 Hour Mean	CO 4 Hour (%CV)	CS 1 Hour Mean	CS 1 Hour (%CV)	CS 4 Hour Mean	CS 4 Hour (%CV)
Cetylpyridinium bromide (6%)	0.58	88.0	1.75	43	21.4	41	32.0	31
Chlorhexidine	1.25	101	2.68	35	26.8	56	69.2	59
Cyclohexanol	1.08	77	2.50	23	24.3	41	82.1	26
Dibenzyl phosphate	0.50	115	1.08	64	9.5	44	16.4	55
Ethanol	1.72	45.0	2.58	20	26.8	60	52.6	18
Ethyl acetate	0.00	0.00	1.43	47	14.6	41	30.6	46
Ethyl trimethyl acetate	0.00	0.00	0.83	108	6.6	79	12.0	49
Ethyl-2-methylacetoacetate	0.42	120	1.68	50	16.3	68	21.2	67
Fomesafen	0.83	175	1.18	124	9.2	115	16.3	84
Gammabutyrolactone	0.25	200	1.67	63	21.4	19	38.3	26
Glycerol	0.00	0.00	0.33	145	7.7	40	7.6	47
Imidazole	2.50	23.0	2.75	18	44.8	11	74.7	11
Isobutanol	1.33	71.0	2.50	23	25.1	44	75.5	26
Isopropanol	1.34	68.0	1.92	51	16.0	70	35.8	57
L-aspartic acid	0.25	200	0.25	200	5.1	76	6.08	107
Maneb	1.00	141	1.00	115	24.0	82	26.6	87
Methyl acetate	0.50	115	1.59	77	15.1	28	30.6	43
Methyl cyanoacetate	0.08	200	0.66	138	5.0	29	6.9	21
Methyl ethyl ketone	0.92	91.0	2.41	18	21.2	30	61.2	34
Methyl isobutyl ketone	0.25	200	1.58	80	18.2	90	34.2	70
Methylcyclopentane	0.00	0.00	0.00	0	8.2	80	9.5	82
n-Butyl acetate	0.00	0.00	0.34	116	6.6	74	14.7	74
n-Hexanol	0.66	115	2.68	18	18.6	18	48.3	21
n-Octanol	0.00	0.00	1.45	36	11.8	48	21.7	34
Parafluoraniline	1.24	71	2.29	21	27.8	15	64.3	11
Polyethylene glycol 400	0.25	200	0.50	115	15.0	81	17.6	84
Potassium cyanate	0.00	0	0.00	0	5.2	59	5.3	113

Substance	CO 1 Hour Mean	CO 1 Hour (%CV)	CO 4 Hour Mean	CO 4 Hour (%CV)	CS 1 Hour Mean	CS 1 Hour (%CV)	CS 4 Hour Mean	CS 4 Hour (%CV)
Promethazine HCl	1.50	38	2.33	20	44.1	67	89.7	36
Pyridine	1.83	31	2.83	12	25.9	54	54.9	26
Quinacrine	0.00	0	0.18	200	7.1	82	81.0	89
Sodium hydroxide (1%)	0.99	72	2.75	18	50.2	22	93.5	26
Sodium hydroxide (10%)	2.93	24	4.00	0	101.6	13	138.3	18
Sodium lauryl sulfate (3 %)	0.00	0	0.50	115	9.8	37	15.4	35
Sodium lauryl sulfate (15 %)	0.08	200	1.33	63	16.3	21	23.4	10
Sodium oxalate	0.00	0	0.00	0	7.3	97	9.7	85
Sodium perborate	0.00	0	0.00	0	3.2	57	5.5	118
Tetraaminopyrimidine sulfate	0.75	128	0.75	128	4.3	129	10.3	98
Toluene	0.43	119	0.50	115	14.4	65	22.8	61
Trichloroacetic acid (3%)	0.68	70	0.75	128	8.1	34	18.4	72
Trichloroacetic acid (30%)	3.43	15	3.68	13	24.0	118	77.4	43
Triton X-100 (10 %)	0.67	141	2.33	20	27.1	51	56.8	64
Triton X-100 (5 %)	0.58	164	1.95	39	19.7	35	33.0	26
Tween 20	0.00	0	0.25	200	13.5	75	15.8	66

Substance	CO 1 Hour Mean	CO 1 Hour (%CV)	CO 4 Hour Mean	CO 4 Hour (%CV)	CS 1 Hour Mean	CS 1 Hour (%CV)	CS 4 Hour Mean	CS 4 Hour (%CV)
Mean for All Substances	**0.72**	**84.1**	**1.47**	**63.79**	**19.19**	**56.18**	**37.08**	**53.47**
Median for All Substances		**74.6**		**43.4**		**50.8**		**49.7**
Range for All Substances	**0-3.4**	**0-200**	**0-3.7**	**0-200**	**5-102**	**11-129_**	**6-108**	**10-118**
Mean for Severe Irritants (GHS)	**32.4**	**46.6**	**1.94**	**40.5**	**33.2**	**37.6**	**33.3**	**36.9**
Median for Severe Irritants		**40.6**		**33.6**		**36.0**		**35.5**
Range for Severe Irritants	**0-3.4**	**0-200**	**0-2.4**	**0-200**	**5-102**	**11-118**	**6-108**	**11-118**

CO = Corneal opacity; CS = Corneal swelling; SD = Standard deviation; %CV = Percent coefficient of variation
[1]Substances listed in bolded italics are classified *in vivo* as severe irritants (Category 1) according to GHS (UN 2003).

7.2.5 Additional Analysis of Interlaboratory Reproducibility

In the EC/HO validation study, Balls et al. (1995) determined the interlaboratory correlation between four specific IRE endpoints (corneal opacity at 1 and 4 hours; corneal swelling at 1 and 4 hours) as well as the summary endpoint generated by four independent laboratories. Correlation analyses were conducted for the total data set, along with specific subsets of substances (water-soluble, water-insoluble, surfactants, solids, solutions, and liquids). This analysis yielded a range of correlation coefficients provided in **Table 7-7** (see **Appendix E** for all correlation coefficients derived from comparing each laboratory with every other laboratory).

Interlaboratory correlation coefficients varied considerably depending on the endpoint assessed and the subset of substances tested. In general, when the different endpoints were considered, the highest correlation and the most consistent data was produced with the 4-hour opacity and swelling measurements. Also, in general, compared to the individual 4-hour opacity and swelling measurements, the IRE summary score exhibited greater variability and a lower maximum correlation. The highest correlation was obtained for surfactants (0.696-0.853; 4-hour opacity, and 0.532-0.677; 4-hour swelling) and for liquids (0.402-0.759; 4-hour opacity, and 0.527-0.763; 4-hour swelling). For solids, the highest correlation was only 0.566 and the range of correlation values was increased considerably. Much of the discordance can be attributed to a single laboratory (laboratory b) for the entire range of substances. In general, there was good correlation between three of the four laboratories, including the lead laboratory. The other laboratories (laboratories c and d) contributed more to the discordance when the substances were solids or those insoluble in water.

7.3 Historical Positive and Negative Control Data

As noted in **Section 2.0**, positive controls have not been employed in the IRE test method publications or submitted data, and therefore, historical positive control data is not available. In addition, although negative/vehicle controls (isotonic saline) are traditionally run on at least one test eye with each experiment, these data have not been published and/or provided with data submitted for this BRD. Therefore, an analysis of historical negative control data also is not possible.

Table 7-7 Interlaboratory Correlation Ranges Determined for Various Subsets of Tested Substances in Balls et al. (1995)

Index Score	Interlaboratory Pearson's Correlation (r) of the *In Vitro* Data
Full set of substances (60)	
IREA-Mean Opacity Score, 1 Hour	0.407-0.502
IREB-Mean Opacity Score, 4 Hour	0.485-0.606
IREC-Corneal Swelling, 1 Hour	0.247-0.528
IRED-Corneal Swelling, 4 Hour	0.447-0.611
IRESUM-Summary Score	0.399-0.483
Chemicals soluble in water (30)	
IREA-Mean Opacity Score, 1 Hour	0.422-0.514
IREB-Mean Opacity Score, 4 Hour	0.341-0.516
IREC-Corneal Swelling, 1 Hour	0.246-0.492
IRED-Corneal Swelling, 4 Hour	0.329-0.552
IRESUM-Summary Score	0.471-0.560
Chemicals insoluble in water (18)	
IREA-Mean Opacity Score, 1 Hour	0.104-0.706
IREB-Mean Opacity Score, 4 Hour	0.422-0.730
IREC-Corneal Swelling, 1 Hour	0.177-0.762
IRED-Corneal Swelling, 4 Hour	0.342-0.763
IRESUM-Summary Score	0.156-0.502
Surfactants (12)	
IREA-Mean Opacity Score, 1 Hour	0.466-0.833
IREB-Mean Opacity Score, 4 Hour	0.696-0.853
IREC-Corneal Swelling, 1 Hour	0.204-0.690
IRED-Corneal Swelling, 4 Hour	0.532-0.677
IRESUM-Summary Score	0.513-0.666
Solids (20)	
IREA-Mean Opacity Score, 1 Hour	0.001-0.403
IREB-Mean Opacity Score, 4 Hour	0.231-0.564
IREC-Corneal Swelling, 1 Hour	-0.056-0.487
IRED-Corneal Swelling, 4 Hour	0.112-0.566
IRESUM-Summary Score	0.033-0.293
Solutions (14)	
IREA-Mean Opacity Score, 1 Hour	0.502-0.718
IREB-Mean Opacity Score, 4 Hour	0.657-0.763
IREC-Corneal Swelling, 1 Hour	0.157-0.564
IRED-Corneal Swelling, 4 Hour	0.240-0.686
IRESUM-Summary Score	0.631-0.770
Liquids (26)	
IREA-Mean Opacity Score, 1 Hour	0.197-0.595
IREB-Mean Opacity Score, 4 Hour	0.402-0.759
IREC-Corneal Swelling, 1 Hour	0.115-0.709
IRED-Corneal Swelling, 4 Hour	0.527-0.763
IRESUM-Summary Score	0.203-0.514

7.4 Conclusions

Evaluation of the intralaboratory repeatability and reproducibility of the IRE test method could not be conducted. Interlaboratory reproducibility was assessed based on a qualitative analysis (correct classification as a severe irritant or as a nonsevere irritant) of the individual laboratory test results obtained for the EC/HO validation study (Balls et al. 1995). However, it must be noted that the protocols for these studies were not always identical. This data suggested that the IRE test method may be generally reproducible with respect to identification of severe irritants (and ocular corrosives). For example, in the Balls et al. (1995) validation study, when *in vivo* data from four laboratories was assigned a regulatory classification and compared to irritancy defined using the IRE test method with decision criteria targeted for identification of severe irritants (i.e., Guerriero et al. 2004), 100% of the laboratories correctly identified the 14, 18, and 12 substances, respectively, tested as Category 1 GHS (UN 2003), Category I EPA (EPA 1996), or R41 EU (EU 2001) severe irritants. Discordance was greatest for false positives where only 45-83% of the substances were concordant among three of the four testing laboratories, and 45-50% were concordant among two of the four testing laboratories. By chemical class, the substances with the greatest levels of interlaboratory variability in all studies included alcohols, carboxylic acids, esters, and ketones. Solvent was the most common product class exhibiting a greater level of interlaboratory variability.

An evaluation of IRE interlaboratory variability using a CV analysis of corneal swelling, corneal opacity, and fluorescein retention also indicated generally reproducible results across laboratories when testing severe irritants (%CVs for severe irritants were approximately 40% for studies where the recommended protocol was not used). When all substances tested were considered, the %CV increased to 84%.

Based on the results from this limited dataset, the IRE test method appears to be generally reproducible among different laboratories with respect to the identification of severe irritants and false positives. However, there is not enough reliability data to draw definitive conclusions based on the limited available data. Reliability needs to be assessed using the standardized test method protocol (with all four ocular parameters) against an appropriate set of substances of varying levels of irritancy, physicochemical properties, chemical classes and product classes.

[This Page Intentionally Left Blank]