6.0 BCOP TEST METHOD ACCURACY

6.1 Accuracy of the BCOP Test Method

A critical component of an ICCVAM evaluation of the validation status of a test method is an assessment of the accuracy of the proposed test method when compared to the current reference test method (ICCVAM 2003). This aspect of assay performance is typically evaluated by calculating:

- accuracy (Concordance): the proportion of correct outcomes (positive and negative) of a test method
- sensitivity: the proportion of all positive substances that are classified as positive
- specificity: the proportion of all negative substances that are classified as negative
- positive predictivity: the proportion of correct positive responses among substances testing positive
- negative predictivity: the proportion of correct negative responses among substances testing negative
- false positive rate: the proportion of all negative substances that are falsely identified as positive
- false negative rate: the proportion of all positive substances that are falsely identified as negative

The ability of the BCOP test method to correctly identify ocular corrosives and severe irritants, as defined by the EPA (1996), the EU (2001), and the GHS (UN 2003)[1], was evaluated using two approaches. In the first approach, the performance of the BCOP assay was assessed separately for each *in vitro-in vivo* comparative study (i.e., publication or data submission) reviewed in **Sections 4.0** and **5.0**. In the second approach, the performance of the BCOP was assessed after pooling data across comparative studies that used the same method of data collection and analysis. The three ocular hazard classification systems considered during this analysis use different classification schemes and decision criteria to identify ocular corrosives and severe irritants based on *in vivo* rabbit eye test results (see **Section 1.0**). All three regulatory classification systems are based on individual animal data in terms of the magnitude of the response and, for the EPA and GHS, the amount of time it takes for the ocular lesions to clear. Thus, to evaluate the accuracy of the BCOP test method for identifying ocular corrosives and severe irritants, individual rabbit data collected at the different observation times were needed for each substance. However, these data were not consistently available in the studies considered, which limited the number of test results that could be used to assess test method accuracy. Furthermore, most of the *in vivo* classifications used for the analyses presented in this section are based on the results of a single study. Unless otherwise indicated, variability in the *in vivo* classification is unknown.

[1] For the purposes of this analysis, an ocular corrosive or severe irritant was defined as a substance that would be classified as Category 1 according to the GHS classification system (UN 2003), as Category I according to the EPA classification system (EPA 1996), or as R41 according to the EU classification system (EU 2001) (see **Section 1.0**).

In addition, the accuracy assessments conducted were based on BCOP data that were evaluated differently. As discussed in **Section 2.2.12**, a majority of BCOP studies used the mean opacity and mean permeability values (OD_{490}) for each treatment group to calculate an *In Vitro* Irritancy Score for each test substance. However, Casterton et al. (1996) assigned irritation classes based on the endpoint (opacity or permeability) with the highest score for its respective range. Conversion of the BCOP data in Casterton et al. (1996) to an *In Vitro* Irritancy Score was not attempted since opacity was measured with a UV/VIS spectrophotometer instead of an opacitometer; the author's classifications were used for this analysis. Gettings et al. (1996) used the *In Vitro* Irritancy Score and permeability score alone to classify the 25 surfactant-based formulations evaluated in the CTFA Phase III study, and it was found that the permeability score alone better predicted the *in vivo* ocular classification according to the FHSA classification system. Thus, for this accuracy analysis, only permeability scores are considered for Gettings et al. (1996).

Accuracy of BCOP for Individual Studies: For the "per study" accuracy analysis, two different types of analyses were performed. In the first analysis, the BCOP ocular irritancy potential of each test substance in each study was determined (**Appendix C**). For the three studies where the same test substance was evaluated in multiple laboratories within the same study (i.e., Gautheron et al. 1994; Balls et al. 1995; Southee 1998), the BCOP ocular irritancy classification for each independent test result was determined. Subsequently, an overall BCOP ocular irritancy classification was assigned for each chemical in the study based on the majority of ocular irritancy classification calls (e.g., if two tests classified a substance as a moderate irritant and three tests classified a substance as a severe irritant, the overall *in vitro* irritancy classification for the substance would be severe irritant). When there was an even number of different irritancy classifications for test substances (e.g., two tests classified a substance as a moderate irritant and two tests classified a substance as a severe irritant), the more severe irritancy classification was used for the overall classification for the substance (severe irritant, in this case). Once the ocular irritancy potential classification was determined for each substance in each of the studies, the ability of the BCOP test method to identify ocular corrosives and severe irritants, based on the three different classification systems, was determined for each study (**Appendix D**).

The second analysis conducted in the "per study" evaluation used each independent test result for each substance that had been tested in multiple laboratories (Gautheron et al. 1994, Balls et al. 1995, and Southee 1998). Each *in vitro* classification obtained when a test substance was evaluated in multiple laboratories was used separately to assess test method accuracy (i.e., results were not combined across multiple testing laboratories to develop an overall BCOP ocular irritancy classification). The ability of the BCOP test method to identify ocular corrosives and severe irritants, based on the three different regulatory classification systems, was then determined.

Accuracy of BCOP for Pooled Studies: For an overall analysis of accuracy for BCOP, results from the six different comparative studies that used the same BCOP analysis approach (i.e., calculation of an *In Vitro* Irritancy Score = opacity + (15 x OD_{490}) or use of permeability value only for substances that produce permeability without appreciable opacity) were combined and an overall ocular classification was determined for each substance. When the same test substance was evaluated in multiple studies, the overall BCOP ocular irritancy potential was based on the majority of calls among all of the studies (see **Appendix C**). Once the overall *in vitro* ocular irritancy classification was determined for each test substance, the classification was compared to the *in vivo* ocular irritancy classification (**Appendix D**).

6.1.1 GHS Classification System: BCOP Test Method Accuracy

Accuracy analyses for ocular corrosives and severe irritants, as defined by the GHS classification system[2] (UN 2003), were performed for the following eight studies: Gautheron et al. (1994), Balls et al. (1995), Swanson et al. (1995), Casterton et al. (1996), Gettings et al. (1996), Southee (1998), Swanson and Harbell (2000), and Bailey et al. (2004). The GHS classification assigned to each test substance is presented in **Appendix D**. The performance characteristics (i.e., accuracy, sensitivity, specificity, positive predictivity, negative predictivity, false positive rate, and false negative rate) were determined for each of the eight studies based on the available *in vivo* reference data for the substances tested in these studies (**Table 6-1**). Of the eight studies, Gautheron et al. (1994), Balls et al. (1995), and Southee (1998) provided BCOP data for substances tested in multiple laboratories; the first set of accuracy calculations for these studies in **Table 6-1** represents the results obtained using the consensus call for each test substance, while the second set of accuracy calculations for each study represents the results obtained when each independent test result from each laboratory was considered separately.

Based on the data provided in the eight studies, when a single call was used per test substance per study, the BCOP test method has an accuracy of 67% to 100%, a sensitivity of 48% to 100%, a specificity of 66% to 100%, a false positive rate of 0% to 34%, and a false negative rate of 0% to 52% (**Table 6-1)**.

Using the first accuracy analysis approach (single call per test substance), the three BCOP studies that evaluated test substances in multiple laboratories (Gautheron et al. 1994; Balls et al. 1995; Southee 1998) have an accuracy of 70% to 74%, a sensitivity of 57% to 77%, a specificity of 66% to 88%, a false positive rate of 12% to 34%, and a false negative rate of 23% to 43%. In contrast, when BCOP study results from multiple laboratories are considered separately rather than being combined to provide an overall classification for each substance, the BCOP test method has an accuracy of 70% to 79%, a sensitivity of 69% to 77%, a specificity of 66% to 83%, a false positive rate of 17% to 34%, and a false negative rate of 24% to 31%. These performance characteristics are provided in **Table 6-1**. The values obtained for the second analysis approach changed little in comparison to the first accuracy analysis approach for the Balls et al. (1995) study, but changed more substantially for the Gautheron et al. (1994) and the Southee (1998) studies.

---

[2] For the purpose of this accuracy analysis, *in vivo* rabbit study results were used to identify GHS Category 1 irritants (i.e., severe irritants); substances classified as GHS Category 2A and 2B irritants were identified as nonsevere irritants.

Table 6-1 Evaluation of the Performance of the BCOP Test Method In Predicting Ocular Corrosives and Severe Irritants Compared to *In Vivo* Findings, as Defined by the GHS Classification System, by Study and Overall

Data Source	Anal.[1]	N[2]	Accuracy		Sensitivity		Specificity		Positive Predictivity		Negative Predictivity		False Positive Rate		False Negative Rate	
			%	No.[3]	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.
Gautheron et al. 1994	IVIS	47/52	74[4]	35/47	71	5/7	75	30/40	33	5/15	94	30/32	25	10/40	29	2/7
			77[e]	432/558	69	62/90	79	370/468	39	62/160	93	370/398	21	98/468	31	28/90
Balls et al. 1995[6]	IVIS	54/59	70[4]	38/54	77	17/22	66	21/32	61	17/28	81	21/26	34	11/32	23	5/22
			70[5]	190/270	77	85/110	66	105/160	61	85/140	81	105/130	34	55/160	24	26/110
Swanson et al. 1995	IVIS	8/20	100	8/8	100	6/6	100	2/2	100	6/6	100	2/2	0	0/2	0	0/6
Gettings et al. 1996	Perm	23/25	87	20/23	75	6/8	93	14/15	86	6/7	88	14/16	7	1/15	25	2/8
Casterton et al. 1996	O/P	55/97	67	37/55	48	13/27	86	24/28	76	13/17	63	24/38	14	4/28	52	14/27
Southee 1998	IVIS	15/16	73[4]	11/15	57	4/7	88	7/8	80	4/5	70	7/10	12	1/8	43	3/7
			79[5]	110/139	76	57/75	83	53/64	84	57/68	75	53/71	17	11/64	24	18/75
Swanson & Harbell 2000	IVIS	9/13	78	7/9	100	1/1	75	6/8	33	1/3	100	6/6	25	2/8	0	0/1
Bailey et al. 2004	IVIS	14/16	93	13/14	67	2/3	100	11/11	100	2/2	92	11/12	0	0/11	33	1/3
Pooled Studies[7]		147/203	81	119/147	84	36/43	80	83/104	63	36/57	92	83/90	20	21/104	16	7/43

[1]Anal. = Analytical method used to transform the sample data into BCOP classification; IVIS = *In Vitro* Irritancy Score developed by Gautheron et al. (1994); Perm = Permeability value only used to classify *in vitro* ocular irritancy in the BCOP assay (an OD_{490} value > 0.600 was considered a severe irritant); O/P = Irritation class based on the endpoint (opacity or permeability) with the highest score for its respective range (Casterton et al. 1996).
[2]n = Number of substances included in this analysis/the total number of substances evaluated in the study.
[3]The data on which the percentage calculation is based.
[4]Performance calculated using the overall *in vitro* classification based on the majority and/or most severe classification among the multiple testing laboratories and tests (for substances tested multiple times in a laboratory).
[5]Performance calculated using each individual *in vitro* classification from each testing laboratory and test.
[6]The test substance 1% benzalkonium chloride was tested in two different *in vivo* studies, producing discordant results with respect to GHS classification (study 1 = Category 2B and study 2 = Category 1). The analysis was performed using the Category 1 classification.
[7]Data from Gautheron et al. (1994), Balls et al. (1995), Swanson et al. (1995), Gettings et al. (1996), Southee (1998), Swanson and Harbell (2000), and Bailey et al. (2004) were pooled together and an overall *in vitro* classification was assigned for each test substance based on the majority and/or most severe classification obtained across tests and testing laboratories. Data from Casterton et al. (1996) were not included in this analysis since the protocol used to generate BCOP data differed considerably from the other studies (e.g., A spectrophotometer was used to measure opacity instead of an opacitometer, and solids were applied neat instead of as a 20% solution or suspension).

In terms of an overall accuracy analysis, using data from Gautheron et al. (1994), Balls et al. (1995), Swanson et al. (1995), Gettings et al. (1996), Southee (1998), Swanson and Harbell (2000) and Bailey et al. (2004), the BCOP test method has an accuracy of 81%, a sensitivity of 84%, a specificity of 80%, a false positive rate of 20%, and a false negative rate of 16%. The performance characteristics for the pooled studies are provided in **Table 6-1**.

As described in **Sections 3.0** and **4.0**, appropriate *in vivo* data were not available for all of the substances evaluated in some of the studies. For example, in the Swanson et al. (1995) study, only eight of the 20 substances had appropriate *in vivo* data to assign a GHS classification.

6.1.1.1 *Discordant Results According to the GHS Classification System*

In order to evaluate discordant responses of the BCOP test method relative to the *in vivo* hazard classification, several accuracy sub-analyses were performed. These included specific classes of chemicals with sufficiently robust numbers of substances (n ≥ 5), as well as certain properties of interest considered relevant to ocular toxicity testing (e.g., pesticides, surfactants, pH, physical form).

As indicated in **Table 6-2**, there were some notable trends in the performance of the BCOP test method among these subgroups of substances. The chemical class of substances that was most consistently overpredicted according the GHS classification system (i.e., were false positives) by the BCOP test method is alcohols. Eight out the 21 overpredicted substances were alcohols. Additional chemical classes represented among the overpredicted substances were ketones (4), carboxylic acids (3), heterocyclic compounds (2), esters (1), and hydrocarbons (1). Among the 35 substances labeled as surfactants only 5% (1/21, a surfactant-containing formulation) were overpredicted by the BCOP test method.

With regard to physical form of the substances overpredicted by the BCOP test method, 18 were liquids and two were solids. Considering the proportion of the total available database, liquids (92/124; 74%) appear more likely than solids (32/124; 26%) to be overpredicted by the BCOP test method.

Although there were a relatively small number (7) of substances represented, alcohols (2) were most often underpredicted (i.e., were false negatives[3]) by the BCOP test method according to the GHS classification system (see **Appendix D**). As can be seen in **Table 6-2**, the 35 substances labeled as surfactants were rarely underpredicted by the BCOP test method (7% [1/14] false negative rate).

With regard to physical form of the substances underpredicted by the BCOP test method, five were solids and one was a liquid. Despite the proportion of the total available database, solids (32/124; 26%) appear more likely than liquids (92/124; 74%) to be underpredicted by the BCOP test method.

[3] False negative in this context refers to a substance that was classified as a nonsevere (mild or moderate) irritant or nonirritant by the BCOP test method, but as a severe irritant based on *in vivo* data.

Table 6-2 False Positive and False Negative Rates of the BCOP Test Method, by Chemical Class and Properties of Interest, for the GHS[1] Classification System

Category	N[2]	False Positive Rate[3]		False Negative Rate[4]	
		%	No.[5]	%	No.
Overall	147	20	21/104	16	7/43
Chemical Class[6]					
Alcohol	18	53	8/15	67	2/3
Amine/Amidine	8	0	0/4	0	0/4
Carboxylic acid	15	38	3/8	14	1/7
Ester	12	12	1/8	0	0/4
Ether/Polyether	6	0	0/5	0	0/1
Heterocycle	12	33	2/6	17	1/6
Hydrocarbon	12	8	1/12	-	0/0
Inorganic salt	5	0	0/3	0	0/2
Ketone	10	40	4/10	-	0/0
Onium compound	11	0	0/3	0	0/8
Properties of Interest					
Liquids[7]	92	26	18/68	4	1/24
Solids[7]	32	10	2/20	42	5/12
Pesticide	8	33	1/3	40	2/5
Surfactant – Total[8]	35	5	1/21	7	1/14
-nonionic	5	0	0/4	0	0/1
-anionic	3	0	0/2	100	1/1
-cationic	6	0	0/1	0	0/5
pH – Total[9]	28	-	-	21	5/24
- acidic (pH < 7.0)	11	-	-	18	2/11
- basic (pH > 7.0)	15	-	-	23	3/13
- equals 7	2	-	-	-	-
Category 1 Subgroup[10] **- Total**	38[11]	-	-	18	7/38
- 4 (CO=4 at any time)	20	-	-	15	3/20
- 3 (severity/persistence)	1	-	-	0	0/1
- 2 (severity)	4	-	-	25	1/4
- 2-4 combined[12]	25	-	-	16	4/25
- 1 (persistence)	13	-	-	23	3/13

[1]GHS = Globally Harmonized System (UN 2003).
[2]N = Number of substances.
[3]False Positive Rate = The proportion of all negative substances that are falsely identified as positive *in vitro*.
[4]False Negative Rate = The proportion of all positive substances that are falsely identified as negative *in vitro*.
[5]Data used to calculate the percentage.
[6]Chemical classes included in this table are represented by at least five substances tested in the BCOP test method and assignments are based on the MeSH categories (www.nlm.nih.gov/mesh)
[7]Physical form (i.e., solid or liquid) not known for some substances, and therefore the overall number does not equal the sum of the solid and liquid substances.
[8]Combines single chemicals labeled as surfactants along with surfactant-containing formulations.
[9]Total number of GHS Category 1 substances for which pH information was obtained.
[10]NICEATM-defined subgroups assigned based on the lesions that drove classification of a GHS Category 1 substance. 1: based on lesions that are persistent; 2: based on lesions that are severe (not including Corneal Opacity [CO]=4); 3: based on lesions that are severe (not including CO=4) and persistent; 4: CO = 4 at any time.
[11]The number of substances evaluated in the Category 1 subgroup analysis may be less than the total number of *in vivo* Category 1 substances evaluated, since some substances could not be classified into the subgroups used in the evaluation.
[12]Subcategories 2 to 4 combined to allow for a direct comparison of GHS Category 1 substances classified *in vivo* based on some lesion severity component and those classified based on persistent lesions alone.

There was no definitive difference among the underpredicted substances for which pH information was available, as two were acidic (pH < 7.0) and three were basic (pH > 7.0), and considering the comparable proportion of acidic and basic underpredicted substances

(2/11; 18% vs. 3/13; 23%). Finally, the seven underpredicted substances were more likely to be substances classified *in vivo* based on persistent lesions (3/13; 23%) rather than on severe lesions (4/25; 16%), as evidenced by an analysis of NICEATM-defined GHS Category 1 sub-groupings (**Table 6-2**).

Table 6.3 shows the effects on the BCOP test method performance statistics of excluding from the data set problematic classes (i.e., that gave the most discordant results, according to the GHS classification system). In general, exclusion of alcohols, ketones or solids individually resulted in small changes in the performance statistics, with the exception that the exclusion of solids from the data set caused a four-fold decrease in the false negative rate from 16% (7/43) to 4% (1/29). When both alcohols and ketones were excluded from the data set, changes in the performance statistics were noted, with accuracy increasing from 81% (119/147) to 88% (103/117), and the false positive rate decreasing from 20% (21/104) to 12% (9/77). The largest changes were observed when all three discordant classes were excluded from the data set; accuracy increased from 81% (119/147) to 92% (78/85), the false positive rate decreased from 20% (21/104) to 12% (7/58), and the false negative rate decreased from 16% (7/43) to 0% (0/27).

Table 6-3 Effect of Exclusion of Discordant Classes on False Negative and False Positive Rates of the BCOP Test Method, for the GHS[1] Classification System

Data Set	*Accuracy*		*False Positive Rate[2]*		*False Negative Rate[3]*	
	%	*No.[4]*	*%*	*No.*	*%*	*No.*
Overall	*81*	*119/147*	*20*	*21/104*	*16*	*7/43*
w/o Alcohols	*86*	*109/126*	*14*	*12/86*	*13*	*5/40*
w/o Ketones	*81*	*113/138*	*19*	*18/95*	*16*	*7/43*
w/o Solids	*82*	*93/113*	*23*	*19/84*	*4*	*1/29*
w/o Alcohols & Ketones	*88*	*103/117*	*12*	*9/77*	*13*	*5/40*
w/o Alcohols & Ketones & Solids	*92*	*78/85*	*12*	*7/58*	*0*	*0/27*

[1]GHS =- Globally Harmonized System (UN 2003).
[2]False Positive Rate = The proportion of all negative substances that are falsely identified as positive *in vitro*
[3]False Negative Rate = The proportion of all positive substances that are falsely identified as negative *in vitro*
[4]Data used to calculate the percentage.

6.1.2 EPA Classification System: BCOP Test Method Accuracy

Accuracy analyses for ocular corrosives and severe irritants, as defined by the EPA classification system[4] (EPA 1996), were performed for the following eight studies: Gautheron et al. (1994), Balls et al. (1995), Swanson et al. (1995), Casterton et al. (1996), Gettings et al. (1996), Southee (1998), Swanson and Harbell (2000), and Bailey et al. (2004). The EPA classification assigned to each test substance is presented in **Appendix D.** The performance characteristics of the eight studies are shown in **Table 6-4** and are based on the available *in vivo* reference data for each study. Of the eight studies, Gautheron et al. (1994), Balls et al. (1995), and Southee (1998) provided BCOP data for substances tested in multiple laboratories; the first set of accuracy calculations for these studies in **Table 6-4** represents the results obtained using the consensus call for each test substance, while the second set of accuracy calculations for each study represents the results obtained when each independent test result from each laboratory was considered separately.

Based on the data provided in these eight studies, when a single call was used per test substance per study, the BCOP test method has an accuracy of 62% to 92%, a sensitivity of 40% to 100%, a specificity of 50% to 100%, a false positive rate of 0% to 50%, and a false negative rate of 0% to 100% (**Table 6-4)**.

Using the first accuracy analysis approach (single call per test substance), the three BCOP studies that evaluated test substances in multiple laboratories (Gautheron et al. 1994; Balls et al. 1995; Southee 1998) have an accuracy of 64% to 73%, a sensitivity of 40% to 72%, a specificity of 63% to 78%, a false positive rate of 22% to 37%, and a false negative rate of 28% to 60%. In contrast, when BCOP study results from multiple laboratories are considered separately rather than being combined to provide an overall classification for each substance, the BCOP test method has an accuracy of 66% to 75%, a sensitivity of 60% to 72%, a specificity of 63% to 76%, a false positive rate of 24% to 37%, and a false negative rate of 28% to 40% (**Table 6-4**). The values obtained for the second analysis approach changed little in comparison to the first accuracy analysis approach for the Balls et al. (1995) study, but changed more substantially for the Gautheron et al. (1994) and the Southee (1998) studies.

In terms of an overall accuracy analysis, using data from Gautheron et al. (1994), Balls et al. (1995), Swanson et al. (1995), Gettings et al. (1996), Southee (1998), Swanson and Harbell (2000) and Bailey et al. (2004), the BCOP test method has an accuracy of 79%, a sensitivity of 75%, a specificity of 81%, a false positive rate of 19%, and a false negative rate of 25%. The performance characteristics for the pooled studies are provided in **Table 6-4**.

As described in **Section 4.0**, *in vivo* data were not available for all of the substances evaluated in some of the studies. For example, for the Swanson et al. (1995) study, only eight of the 20 substances had sufficient *in vivo* data to assign an EPA classification.

---

[4] For the purpose of this accuracy analysis, *in vivo* rabbit study results were used to identify EPA Category I irritants (i.e., severe irritants); substances classified as EPA Category II, III, or IV irritants were defined as nonsevere irritants.

6.1.2.1 *Discordant Results According to the EPA Classification System*

In order to evaluate discordant responses of the BCOP test method relative to the *in vivo* hazard classification, several accuracy sub-analyses were performed. These included specific classes of chemicals with sufficiently robust numbers of substances ($n \geq 5$), as well as certain properties of interest considered relevant to ocular toxicity testing (e.g., pesticides, surfactants, pH, physical form).

Table 6-4 Evaluation of the Performance of the BCOP Test Method In Predicting Ocular Corrosives and Severe Irritants Compared to *In Vivo* Findings, as Defined by the EPA Classification System, by Study and Overall

Data Source	Anal.[1]	N[2]	Accuracy		Sensitivity		Specificity		Positive Predictivity		Negative Predictivity		False Positive Rate		False Negative Rate	
			%	No.[3]	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.
Gautheron et al. 1994[4]	IVIS	48/52	73[5]	35/48	71	5/7	73	30/41	31	5/16	94	30/32	27	11/41	29	2/7
			75[6]	427/571	65	51/79	76	376/492	30	51/167	93	376/404	24	116/492	35	28/79
Balls et al. 1995[4]	IVIS	53/59	66[5]	35/53	72	13/18	63	22/35	50	13/26	82	22/27	37	13/35	28	5/18
			66[6]	175/265	72	65/90	63	110/175	50	65/130	82	110/135	37	65/175	28	25/90
Swanson et al. 1995	IVIS	8/20	88	7/8	100	6/6	50	1/2	86	6/7	100	1/1	50	1/2	0	0/6
Gettings et al. 1996	Perm	25/25	80	20/25	60	6/10	93	14/15	86	6/7	78	14/18	7	1/15	40	4/10
Casterton et al. 1996	O/P	56/97	62	35/56	41	11/27	83	24/29	69	11/16	60	24/40	17	5/29	59	16/27
Southee 1998[4]	IVIS	14/16	64[5]	9/14	40	2/5	78	7/9	50	2/4	70	7/10	22	2/9	60	3/5
			70[6]	80/115	60	27/45	76	53/70	61	27/44	75	53/71	24	17/70	40	18/45
Swanson & Harbell 2000[4]	IVIS	9/13	89	8/9	75	3/4	100	5/5	100	3/3	83	5/6	0	0/5	25	1/4
Bailey et al. 2004	IVIS	13/16	92	12/13	0	0/1	100	12/12	-	0/0	92	12/13	0	0/12	100	1/1
Pooled Studies[7]		143/203	79	113/143	75	30/40	81	83/103	60	30/50	89	83/93	19	20/103	25	10/40

[1]Anal. = Analytical method used to transform the sample data into BCOP classification; IVIS = *In Vitro* Irritancy Score developed by Gautheron et al. (1994); Perm = Permeability value only used to classify *in vitro* ocular irritancy in the BCOP assay (an OD_{490} value > 0.600 was considered a severe irritant); O/P = Irritation class based on the endpoint (opacity or permeability) with the highest score for its respective range (Casterton et al. 1996).
[2]n = Number of substances included in this analysis/the total number of substances in the study.
[3]The data on which the percentage calculation is based.
[4]The test substance ethanol was evaluated in two different *in vivo* studies (ECETOC 1998; Swanson and Harbell 2000), producing discordant results with respect to EPA classification (study 1 = Category III and study 2 = Category I). The analysis was performed using the Category I classification.
[5]Performance calculated using the overall *in vitro* classification based on the majority and/or most severe classification among the multiple testing laboratories and tests (for substances tested multiple times in a laboratory).
[6]Performance calculated using each individual *in vitro* classification from each testing laboratory and test.
[7]Data from Gautheron et al. (1994), Balls et al. (1995), Swanson et al. (1995), Gettings et al. (1996), Southee (1998), Swanson and Harbell (2000), and Bailey et al. (2004) were pooled together and an overall *in vitro* classification was assigned for each test substance based on the majority and/or most severe classification obtained across tests and testing laboratories. Data from Casterton et al. (1996) were not included in this analysis, since the protocol used to generate BCOP data differed considerably from the other studies (e.g., a spectrophotometer was used to measure opacity instead of an opacitometer, and solids were applied neat instead of as a 20% solution or suspension).

As indicated in **Table 6-5**, there were some notable trends in the performance of the BCOP test method among these subgroups of substances. The chemical class of substances that was most consistently overpredicted according the EPA classification system (i.e., were false positives) by the BCOP test method is alcohols. Nine out the 20 overpredicted substances were alcohols. Additional chemical classes represented among the overpredicted substances were ketones (4), carboxylic acids (3), heterocyclic compounds (2), esters (2), hydrocarbons (1), inorganic salts (1), and onium compounds (1). Among the 35 substances labeled as surfactants only 9% (2/22) were overpredicted by the BCOP test method (10% Triton X-100 and a surfactant-containing formulation).

Table 6-5 False Positive and False Negative Rates of the BCOP Test Method, by Chemical Class and Properties of Interest, for the EPA[1] Classification System

Category	N[2]	False Positive Rate[3]		False Negative Rate[4]	
		%	No.[5]	%	No.
Overall	143	19	20/103	25	10/40
Chemical Class[6]					
Alcohol	18	56	9/16	100	2/2
Amine/Amidine	8	0	0/6	0	0/2
Carboxylic acid	14	38	3/8	17	1/6
Ester	9	22	2/9	-	0/0
Ether/Polyether	6	0	0/5	100	1/1
Heterocycle	11	33	2/6	20	1/5
Hydrocarbon	12	8	1/12	-	0/0
Inorganic salt	5	25	1/4	0	0/1
Ketone	10	40	4/10	-	0/0
Onium compound	9	25	1/4	0	0/5
Properties of Interest					
Liquids[7]	90	29	18/70	5	1/20
Solids[7]	31	10	2/21	50	5/10
Pesticide	9	25	1/4	40	2/5
Surfactant – Total[8]	35	9	2/22	23	3/13
-nonionic	5	20	1/5	-	0/0
-anionic	3	0	0/2	0	0/1
-cationic	4	0	0/1	0	0/3
pH – Total[9]	25	-	-	32	6/19
- acidic (pH < 7.0)	9	-	-	30	3/10
- basic (pH > 7.0)	14	-	-	33	3/9
- equals 7	2	-	-	-	-

[1]EPA = U.S. Environmental Protection Agency (EPA 1996).
[2]N = Number of substances.
[3]False Positive Rate = The proportion of all negative substances that are falsely identified as positive *in vitro*.
[4]False Negative Rate = The proportion of all positive substances that are falsely identified as negative *in vitro.*
[5]Data used to calculate the percentage.
[6]Chemical classes included in this table are represented by at least five substances tested in the BCOP test method and assignments are based on the MeSH categories (www.nlm.nih.gov/mesh)
[7]Physical form (i.e., solid or liquid) not known for some substances, and therefore the overall number does not equal the sum of the solid and liquid substances.
[8]Combines single chemicals labeled as surfactants along with surfactant-containing formulations.
[9]Total number of EPA Category I substances for which pH information was obtained.

With regard to physical form of the substances overpredicted by the BCOP test method, 18 were liquids and two were solids. Considering the proportion of the total available database, liquids (90/121; 74%) appear more likely than solids (31/121; 26%) to be overpredicted by the BCOP test method

Although there were a relatively small number (10) of substances represented, alcohols (2) were most often underpredicted (i.e., were false negatives) by the BCOP test method according to the EPA classification system (see **Appendix D**). As can be seen in **Table 6-5**, some of the 35 substances labeled as surfactants were underpredicted by the BCOP test method (23% [3/13] false negative rate).

With regard to physical form of the substances underpredicted by the BCOP test method, five were solids and one was a liquid. Despite the proportion of the total available database, solids (31/121; 26%) appear more likely than liquids (90/121; 74%) to be underpredicted by the BCOP test method.

There was no definitive difference among the underpredicted substances for which pH information was available, as three were acidic (pH < 7.0) and three were basic (pH > 7.0), and considering the comparable proportion of acidic and basic underpredicted substances (3/10; 30% vs. 3/9; 33%).

6.1.3 EU Classification System: BCOP Test Method Accuracy

Accuracy analyses for ocular corrosives and severe irritants, as defined by the EU (2001) classification system[5], were performed for the following eight studies: Gautheron et al. (1994), Balls et al. (1995), Swanson et al. (1995), Casterton et al. (1996), Gettings et al. (1996), Southee (1998), Swanson and Harbell (2000) and Bailey et al. (2004). Of these reports, Gautheron et al. (1994), Balls et al. (1995), and Southee (1998) provided BCOP data for substances tested in multiple laboratories. The EU classification assigned to each test substance is presented in **Appendix D.**

Based on the data provided in these eight studies, when a single call was used per test substance per study, the BCOP test method has an accuracy of 68% to 92%, a sensitivity of 52% to 100%, a specificity of 64% to 100%, a false positive rate of 0% to 36%, and a false negative rate of 0% to 48% (**Table 6-6)**.

Using the first accuracy analysis approach (single call per test substance), the three BCOP studies that evaluated test substances in multiple laboratories (Gautheron et al. 1994; Balls et al. 1995; Southee 1998) have an accuracy of 68% to 79%, a sensitivity of 67% to 74%, a specificity of 64% to 88%, a false positive rate of 12% to 36%, and a false negative rate of 26% to 33%. In contrast, when BCOP study results from multiple laboratories are considered separately rather than being combined to provide an overall classification for each substance, the BCOP test method has an accuracy of 69% to 83%, a sensitivity of 69% to 83%, a specificity of 65% to 83%, a false positive rate of 17% to 35%, and a false negative rate of 17% to 31% (**Table 6-6)**.

[5] For the purpose of this accuracy analysis, *in vivo* rabbit study results were used to identify R41 irritants (i.e., severe irritants); substances classified as R36 were defined as nonsevere irritants.

Table 6-6 Evaluation of the Performance of the BCOP Test Method in Predicting Ocular Corrosives and Severe Irritants Compared to *In Vivo* Findings, as Defined by the EU Classification System, by Study and Overall

Data Source	Anal.[1]	N[2]	Accuracy		Sensitivity		Specificity		Positive Predictivity		Negative Predictivity		False Positive Rate		False Negative Rate	
			%	No.[3]	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.
Gautheron et al. 1994[4]	IVIS	48/52	73[5]	35/48	71	5/7	73	30/41	31	5/16	94	30/32	27	11/41	29	2/7
			77[6]	437/570	69	62/90	78	375/480	37	62/167	93	375/403	22	105/480	31	28/90
Balls et al. 1995	IVIS	50/59	68[5]	34/50	74	14/19	64	20/31	56	14/25	80	20/25	36	11/31	26	5/19
			69[6]	171/248	75	71/95	65	100/153	57	71/124	81	100/124	35	53/153	25	24/95
Swanson et al. 1995	IVIS	9/20	89	8/9	100	6/6	67	2/3	86	6/7	100	2/2	33	1/3	0	0/6
Gettings et al. 1996	Perm	23/25	87	20/23	75	6/8	93	14/15	86	6/7	88	14/16	7	1/15	25	2/8
Casterton et al. 1996	O/P	54/97	70	38/54	52	13/25	86	25/29	76	13/17	68	25/37	14	4/29	48	12/25
Southee 1998	IVIS	14/16	79[5]	11/14	67	4/6	88	7/8	80	4/5	78	7/9	12	1/8	33	2/6
			83[6]	110/133	83	57/69	83	53/64	84	57/68	82	53/65	17	11/64	17	12/69
Swanson & Harbell 2000	IVIS	9/13	78	7/9	100	1/1	75	6/8	33	1/3	100	6/6	25	2/8	0	0/1
Bailey et al. 2004	IVIS	13/16	92	12/13	67	2/3	100	10/10	100	2/2	91	10/11	0	0/10	33	1/3
Pooled Studies[7]		143/203	80	114/143	82	33/40	79	81/103	60	33/55	92	81/88	21	22/103	18	7/40

[1]Anal. = Analytical method used to transform the sample data into BCOP classification; IVIS = *In Vitro* Irritancy Score developed by Gautheron et al. (1994); Perm = Permeability value only used to classify *in vitro* ocular irritancy in the BCOP assay (an OD_{490} value > 0.600 was considered a severe irritant); O/P = Irritation class based on the endpoint (opacity or permeability) with the highest score for its respective range (Casterton et al. 1996).
[2]n = Number of substances included in this analysis/the total number of substances in the study.
[3]The data on which the percentage calculation is based.
[4]Accuracy analysis based on EEC (1984) classifications in Gautheron et al. (1994).
[5]Performance calculated using the overall *in vitro* classification based on the majority and/or most severe classification among the multiple testing laboratories and tests (for substances tested multiple times in a laboratory).
[6]Performance calculated using each individual *in vitro* classification from each testing laboratory and test.
[7]Data from Gautheron et al. (1994), Balls et al. (1995), Swanson et al. (1995), Gettings et al. (1996), Southee (1998), Swanson and Harbell (2000), and Bailey et al (2004) were pooled together and an overall *in vitro* classification was assigned for each test substance based on the majority and/or most severe classification obtained across tests and testing laboratories. Data from Casterton et al. (1996) were not included in this analysis, since the protocol used to generate BCOP data differed considerably from the other studies (e.g., a spectrophotometer was used to measure opacity instead of an opacitometer, and solids were applied neat instead of as a 20% solution or suspension).

The values obtained for the second analysis approach changed slightly in comparison to the first accuracy analysis approach for the Balls et al. (1995) and the Gautheron et al. (1994) studies, but changed more substantially for the Southee (1998) study.

In terms of an overall accuracy analysis, using data from Gautheron et al. (1994), Balls et al. (1995), Swanson et al. (1995), Gettings et al. (1996), Southee (1998), Swanson and Harbell (2000) and Bailey et al. (2004), the BCOP test method has an accuracy of 80%, a sensitivity of 82%, a specificity of 79%, a false positive rate of 21%, and a false negative rate of 18%. The performance characteristics for the pooled studies are provided also in **Table 6-6**.

As described in **Section 4.0**, appropriate *in vivo* data were not available for all of the substances evaluated in some of the studies. For example, in Swanson et al. (1995), only nine of the 20 substances evaluated in this study had sufficient *in vivo* data to assign an EU classification.

6.1.3.1 *Discordant Results According to the EU Classification System*

In order to evaluate discordant responses of the BCOP test method relative to the *in vivo* hazard classification, several accuracy sub-analyses were performed. These included specific classes of chemicals with sufficiently robust numbers of substances ($n \geq 5$), as well as certain properties of interest considered relevant to ocular toxicity testing (e.g., pesticides, surfactants, pH, physical form).

As indicated in **Table 6-7**, there were some notable trends in the performance of the BCOP test method among these subgroups of substances. The chemical class of substances that was most consistently overpredicted according the EU classification system (i.e., were false positives) by the BCOP test method is alcohols. Seven out the 22 overpredicted substances were alcohols. Additional chemical classes represented among the overpredicted substances were carboxylic acids (4), ketones (4), heterocyclic compounds (2), esters (1), and hydrocarbons (1). Among the 35 substances labeled as surfactants only 9% (2/22) were overpredicted by the BCOP test method (15% sodium lauryl sulfate and a surfactant-containing formulation).

With regard to physical form of the substances overpredicted by the BCOP test method, 19 were liquids and two were solids. Considering the proportion of the total available database, liquids (90/120; 75%) appear more likely than solids (30/120; 25%) to be overpredicted by the BCOP test method

Although there were a relatively small number (7) of substances represented, alcohols (2) were most often underpredicted (i.e., were false negatives) by the BCOP test method according to the EU classification system (see **Appendix D**). As can be seen in **Table 6-7**, the 35 substances labeled as surfactants were rarely underpredicted by the BCOP test method (8% [1/13] false negative rate).

With regard to physical form of the substances underpredicted by the BCOP test method, five were solids and one was a liquid. Despite the proportion of the total available database,

Table 6-7 False Positive and False Negative Rates of the BCOP Test Method, by Chemical Class and Properties of Interest, for the EU[1] Classification System

Category	N[2]	False Positive Rate[3]		False Negative Rate[4]	
		%	No.[5]	%	No.
Overall	143	21	22/103	18	7/40
Chemical Class[6]					
Alcohol	17	50	7/14	67	2/3
Amine/Amidine	7	0	0/4	0	0/3
Carboxylic acid	14	44	4/9	20	1/5
Ester	12	12	1/8	0	0/4
Ether/Polyether	6	0	0/5	0	0/1
Heterocycle	12	33	2/6	17	1/6
Hydrocarbon	12	8	1/12	-	0/0
Inorganic salt	5	0	0/3	0	0/2
Ketone	10	40	4/10	-	0/0
Onium compound	11	0	0/3	0	0/8
Organic salt	7	0	0/3	0	0/4
Properties of Interest					
Liquids[7]	90	28	19/67	4	1/23
Solids[7]	30	10	2/20	50	5/10
Pesticide	7	33	1/3	50	2/4
Surfactant – Total[8]	35	9	2/22	8	1/13
-nonionic	5	0	0/4	0	0/1
-anionic	3	33	1/3	-	0/0
-cationic	6	0	0/1	0	0/5
pH – Total[9]	26	-	-	27	6/22
- acidic (pH < 7.0)	14	-	-	25	3/12
- basic (pH > 7.0)	10	-	-	30	3/10
- equals 7	2	-	-	-	-

[1]EU = European Union (EU 2001).
[2]N = Number of substances.
[3]False Positive Rate = The proportion of all negative substances that are falsely identified as positive *in vitro*.
[4]False Negative Rate = The proportion of all positive substances that are falsely identified as negative *in vitro*.
[5]Data used to calculate the percentage.
[6]Chemical classes included in this table are represented by at least five substances tested in the BCOP test method and assignments are based on the MeSH categories (www.nlm.nih.gov/mesh)
[7]Physical form (i.e., solid or liquid) not known for some substances, and therefore the overall number does not equal the sum of the solid and liquid substances.
[8]Combines single chemicals labeled as surfactants along with surfactant-containing formulations.
[9]Total number of EU Category R41 substances for which pH information was obtained.

solids (30/120; 25%) appear more likely than liquids (90/120; 75%) to be underpredicted by the BCOP test method.

There was no definitive difference among the underpredicted substances for which pH information was available, as three were acidic (pH < 7.0) and three were basic (pH > 7.0), and considering the comparable proportion of acidic and basic underpredicted substances (3/12; 25% vs. 3/10; 30%).

6.2 Accuracy of the BCOP Test Method for Identifying Ocular Corrosives and Severe Irritants – Summary of Results

While there were some differences in results among the three hazard classification systems evaluated (i.e, EPA [EPA 1996], EU [EU 2001], and GHS [UN 2003]), the accuracy analysis revealed that BCOP test method performance was comparable among the three systems. As can be seen in **Tables 6-1**, **6-4**, and **6-6**, the overall accuracy of the BCOP test method ranged from 79% to 81%, depending on the classification system used. Sensitivity and specificity ranged from 75% to 84% and 79% to 81%, respectively. The false positive rate ranged from 19% to 21%, while the false negative rate ranged from 16% to 25%. Given the relatively homogeneous performance of the BCOP test method among the three classification systems, the discussion below encompasses all three hazard classification systems, unless otherwise indicated.

6.2.1 Discordance Among Chemical Classes

The accuracy analysis indicated that alcohols are often overpredicted (50% to 56% [7/14 to 9/16] false positive rate, depending on the classification system used) in the BCOP test method. Ketones (40% [4/10]), carboxylic acids (38% to 44% [3/8 to 4/9]), and heterocyclic compounds (33% [2/6]) also had high false positive rates. The numbers of substances among the remaining chemical classes were too few to resolve any definitive trends in overprediction by the BCOP test method. For the purposes of these analyses, NICEATM considered five substances to be the threshold number per chemical class for consideration, and thus chemical classes represented by fewer than five substances were not considered.

Although there were a small number of underpredicted substances (4 to 5), alcohols (2) were most often underpredicted by the BCOP test method. The other chemical classes represented were carboxylic acids (1), ethers/polyethers (1), and heterocyclic compounds (1).

6.2.2 Discordance Among Physical or Chemical Properties of Interest

With regard to physical form of the substances overpredicted by the BCOP test method, 18 to 20 were liquids and two were solids. Considering the proportion of the total available database, liquids (90/120 to 92/124) appear more likely than solids (30/120 to 32/124) to be overpredicted by the BCOP test method.

With regard to physical form of the substances underpredicted by the BCOP test method, five were solids and one was a liquid. Despite the proportion of the total available database indicated above, solids (42% to 50% false negative rate) appear more likely than liquids (4% to 5% false negative rate) to be underpredicted by the BCOP test method.

Exclusion of three discordant classes (i.e., alcohols, ketones and solids) from the data set resulted in an increased accuracy (from 81% to 92%), a decreased false positive rate (from 20% to 12%) and a decreased false negative rate (from 16% to 0%).

The 35 substances labeled as surfactants were rarely underpredicted by the BCOP test method for substances classified as severe by the EU (EU 2001) and GHS (UN 2003) classification systems (i.e., R41 or Category 1) as evidence by the false negative rates ranging from 7% to 8%. Substances classified as severe (i.e., Category I) by the EPA classification system (EPA 1996) were more often underpredicted (false negative rate of 23%). However, although the available database was smaller (n = 7 to 9), substances labeled as pesticides were more often underpredicted by the BCOP test method (false negative rates ranging from 40% to 50%).

Considering the comparable proportion of acidic and basic underpredicted substances (18% to 30% [2/11 to 3/10] vs. 23% to 33% [3/13 to 3/9]), there was little difference among the underpredicted substances for which pH information was available. However, it is noted that pH information was available for only a portion of the 40 to 43 severe irritant substances (i.e., Category 1, Category I, or R41) in the database for each classification system.

Finally, with respect to the GHS classification system only, the seven underpredicted substances were more likely to be substances classified *in vivo* based on persistent lesions (false negative rate of 23% [3/13]), rather than on severe lesions (false negative rate of 17% [4/24]), as evidenced by an analysis of NICEATM-defined GHS Category 1 sub-groupings (**Table 6-2**).

[This Page Intentionally Left Blank]