7.0 TEST METHOD RELIABILITY (REPEATABILITY/ REPRODUCIBILITY)

Studies that did not follow this ASTM FETAX Guideline (1991, 1998), especially in regard to substance identification, data presentation, and analysis, were excluded from consideration of test method reliability.

7.1 Selection Rationale for Substances Used to Evaluate Test Method Reliability

Only limited information is available on the selection rationale for the chemicals/products used to evaluate test method reliability in the five FETAX validation studies. This information is summarized in **Section 3.1**.

7.2 Assessment of Test Method Reliability (Repeatability and Reproducibility)

Five separate but related FETAX validation studies in three phases were conducted. The aim of the validation process was to evaluate the suitability of a defined protocol (ASTM, 1991), assess the inclusion of an MAS in the assay, and to assess FETAX for its reliability within and across laboratories. A total of 26 substances were tested without metabolic activation and 14 substances with metabolic activation, with from three to six different laboratories participating in each validation study. Validation was measured using the four different measurements obtained from FETAX—LC_{50}, EC_{50}, TI, and the MCIG. In some studies, the types and incidence of malformations present in the embryos were considered. The investigators assessed reproducibility and reliability of each FETAX endpoint by calculating coefficients of variation (CV [%]), and conclusions about reproducibility and reliability were made from evaluating the range of CVs for each measure across laboratories. Additionally, in most validation studies, a statistical approach for assessing intra- and inter-laboratory reliability as described in ASTM E691—92 (ASTM, 1992) was used (**Appendix 12**). The ASTM (1992) method formally calculates intra-laboratory variability (*k*) and inter-laboratory variability (*h*). For both *k* and *h*,

95% confidence limits are calculated and values that exceed these limits indicate excess variability. For the validation studies, the intra-laboratory assessment was based on comparing the results of the three identical replicates within each test and not on multiple independent tests for the same substance within the same laboratory. As a single FETAX test result is based on the average of three identical replicates (ASTM, 1991; 1998), the data from these identical replicates may not be entirely appropriate for an analysis of intra-laboratory repeatability.

7.2.1 FETAX Phase I Validation Study (Bantle et al., 1994a)

The Phase I Validation Study was classified as a training and protocol evaluation phase where the identity of the test substances were known (**Appendix 16**). Three substances (6-AN, hydroxyurea, isoniazid) were tested in six laboratories, with one laboratory conducting each study twice using different technicians (i.e., there were seven studies). In the publication, for ease of discussion, the data were considered to have been generated by seven laboratories. Information on the teratogenic activity of these substances can be found in **Appendix 4**. 6-AN is teratogenic in mice. Hydroxyurea is teratogenic in rats, mice, and rabbits. Information on human teratogenic activity for these two substances was not located. Isoniazid is teratogenic in humans but not in rats, mice, or rabbits. All studies were conducted using identical test substance concentrations. For each study, substances were tested in triplicate, in the absence of metabolic activation only, following the standard FETAX protocol (ASTM, 1991).

Hydroxyurea and isoniazid were tested and the data evaluated before 6-AN was tested. Excessive inter-laboratory variability was noted for hydroxyurea and isoniazid. In response, the FETAX protocol was modified from treating each of the two replicate Petri dishes within a dose group separately to using a common treatment scheme (i.e., the test concentration was mixed with culture media prior to adding the media to the cultures). Quantitative information on the types and incidence of malformations observed was not provided.

For hydroxyurea, all seven studies reported a TI value greater than 1.5, while four of seven studies reported an MCIG/LC_{50} ratio less than 0.30. For isoniazid, all seven studies reported a TI value greater than 1.5, while six of seven studies reported an MCIG/LC_{50} ratio less than 0.30.

For 6-AN, six of six studies reported a TI value greater than 1.5, while one of seven studies reported an MCIG/LC_{50} ratio less than 0.3. The reported data, by study, are tabulated in **Table 21;** data for substances tested twice in the same laboratory by different technicians are summarized in **Table 22**. Based on the data provided in **Table 21** and the standard FETAX decision criteria (ASTM, 1991), an assessment was made by NICEATM of the extent of concordance among the participating laboratories in the results obtained. All participating laboratories obtained a TI value greater than 1.5 for all three test substances. However, complete concordance among the participating laboratories in obtaining an MCIG/LC_{50} ratio above or below 0.3 was not obtained for any test substance. Based on the data provided in **Table 22** and the standard FETAX decision criteria (ASTM, 1991), an assessment was made also by NICEATM, of the extent of intra-laboratory concordance for the single laboratory that tested each substance twice using different technicians. Using different technicians within the same laboratory, concordance for the TI value was obtained for all three test substances, while concordance for the MCIG/LC_{50} ratio was obtained for two of the three test substances.

Individual laboratory results were compared by NICEATM using the statistical methodology described in ASTM (1992). The results of this analysis are presented graphically in **Appendix 7**. For hydroyxurea, excessive intra-laboratory variability was present for LC_{50} and EC_{50} values within laboratory three; excessive inter-laboratory variability was not present. For isoniazid, excessive intra-laboratory variability was present for LC_{50}, EC_{50}, and TI values within laboratory three; excessive inter-laboratory variability was present for the same laboratory. For 6-AN, despite the protocol change, excessive intra-laboratory variability occurred for LC_{50} values within laboratory two; excessive inter-laboratory variability was not present.

The overall mean CV(%) for the Phase I Validation Study was 66.3% and the range was 20.5 to 201.5%. These values suggested to the investigators that the protocol needed refinement or that additional technician training was required. The greatest variability, based on CV(%) values, occurred for MCIG data. The investigators in Phase I concluded that the wide variation of results may be due to a lack of consistency of skills in evaluating *X. laevis* embryos for malformations.

Table 21. Phase I Validation Study—Concordance among Laboratories in Obtaining a Significant FETAX Response Based on Single Decision Criteria*

Chemical Tested	TI >1.5 (actual values)	MCIG/LC_{50} <0.3 (actual values)
6-AN	6 of 6 studies (412.5, 620.0, 5541, 432.7, 241.6, no data,465.2)	1 of 7 studies (0.54, 0.41, 0.78, 0.48, 0.86, 0.63, <0.01)
Hydroxyurea	7 of 7 studies (2.8, 3.4, 4.8, 5.7, 2.1, 6.0, 3.4)	4 of 7 studies (0.40, 0.29, 0.48, 0.16, 1.30, 0.18, 0.27)
Isoniazid	7 of 7 studies (43.3, 50.8, 7.3, 72.8, 4.1, 55.5, 41.3)	6 of 7 studies (0.26, 0.01, 0.81, 0.01, 0.14, 0.23, 0.01)
Proportion of study results in agreement	3 of 3 (100%)	0 of 3 (0%)

*Concordance among studies based on agreement in obtaining a TI >1.5 or an MCIG/LC_{50} <0.30.
Data from Bantle et al. (1994a), organized in sequence by laboratory number;
"no data" indicates study not done.

Table 22. Phase I Validation Study—Concordance Within the Same Laboratory in Obtaining a Significant FETAX Response Based on Single Decision Criteria*

Chemical Tested	TI >1.5 (actual values)	MCIG/LC_{50} <0.3 (actual values)
6-AN	2 of 2 studies (412.5, 620.0)	0 of 2 studies (0.41, 0.54)
Hydroxyurea	2 of 2 studies (2.8, 3.4)	1 of 2 studies (0.29, 0.40)
Isoniazid	2 of 2 studies (43.3, 50.8)	2 of 2 studies (0.01, 0.26)
Proportion of study results in agreement	3 of 3 (100%)	2 of 3 (67%)

* Concordance among studies based on agreement in obtaining a TI >1.5 or an MCIG/LC_{50} <0.30.
Data from Bantle et al. (1994a), organized by numeric value.

Based on the results obtained, several modifications to the standard FETAX protocol were recommended by the investigators, including: (1) increasing the acceptable malformation rate in FETAX Solution controls from 7% to 10%; (2) distributing 25-mL volumes of the toxicant solution to 50-mL flasks prior to aliquoting 10 mL into each replicate dish within a test concentration; and (3) potentially eliminating 6-AN as the positive control.

7.2.2 FETAX Phase II Validation Study (Bantle et al., 1994b)

The Phase II Validation Study (**Appendix 17**) followed the 1991 ASTM FETAX Guideline, but used the modifications recommended in the FETAX Phase I Validation Study. Six laboratories participated in the study, with one laboratory conducting each study twice using different technicians (i.e., there were seven studies). In the publication (Bantle et al., 1994b), information on which laboratory conducted the independent replicate studies was not provided, and the within-laboratory results were not discussed. The test substances tested, in the absence of metabolic activation only, were caffeine, 5-fluorouracil, saccharin, and sodium cyclamate. Information on the teratogenic activity of these substances can be found in **Appendix 4**. Caffeine is a teratogen in rats, mice, and rabbits; but not in humans. 5-Fluorouracil is a teratogen in rats, mice, and humans. Sodium cyclamate is not teratogenic in rats, mice, or rabbits. Saccharin is not teratogenic in rats, mice, rabbits, or humans. Where information on the negative or positive teratogenicity of a test substance in a specific species is not provided above, relevant information was not located. Coded substances were used, but all laboratories used the same preset test concentrations. Quantitative information on induced malformations was not provided. Consistent with the ASTM FETAX Guidelines (1991), a concurrent positive control was not included in the study design.

For sodium cyclamate and saccharin, none of the seven studies resulted in a TI value greater than 1.5, or in an MCIG/LC_{50} ratio less than 0.30. For caffeine, all seven studies resulted in a TI value greater than 1.5, while five of the seven studies resulted in an MCIG/LC_{50} ratio less than 0.30. For 5-fluorouracil, all seven studies resulted in a TI value greater than 1.5, while none of the seven studies resulted in an MCIG/LC_{50} ratio less than 0.30. The reported data, by study, are

tabulated in **Table 23**; data responses for substances tested twice in the same laboratory by different technicians are summarized in **Table 24**. Based on the data in **Table 23**, an assessment was made by NICEATM of the extent of concordance among the studies in obtaining a similar response (positive or negative) for each of the substances tested. When the TI value was considered, all participating laboratories obtained the same FETAX response. When the MCIG/LC_{50} ratio was used, inter-laboratory concordance was obtained for three of the four test substances. Based on the data provided in **Table 23** and the standard FETAX decision criteria (ASTM, 1991), an assessment was made by NICEATM of the extent of intra-laboratory concordance for the single laboratory that tested each substance twice using different technicians. Concordance for the TI value and the MCIG/LC_{50} ratio were obtained for all four substances tested.

Individual laboratory results were compared using the statistical methodology described in ASTM (1992). The results of this analysis are presented graphically in **Appendix 7**. For sodium cyclamate, excessive intra-laboratory variability occurred for TI values within laboratory four; excessive inter-laboratory variability was not present. For saccharin, excessive intra- and inter-laboratory variability was not present. For caffeine, excessive intra-laboratory variability occurred for EC_{50} and TI values within laboratory three; excessive inter-laboratory variability was not present. For 5-fluorouracil, excessive intra-laboratory variability occurred for TI and MCIG values within laboratories two and four, respectively; excessive inter-laboratory variability was not present.

Compared to the Phase I Validation Study results, the CVs were much reduced. The overall mean CV(%) for the Phase II Validation Study was 24.4% and the range was 7.3 to 54.7%. The MCIG seemed to consistently be the most variable measure in both Phase I and II, and was considered to be a direct reflection of the difficulty of evaluating *X. laevis* embryos for malformations at the end of the 96-hour treatment period. The investigators concluded that non-teratogens showed the most consistent results.

Table 23. Phase II Validation Study—Concordance among Laboratories in Obtaining a Significant FETAX Response Based on Single Decision Criteria*

Chemical Tested	TI >1.5 (actual values)	MCIG/LC_{50} <0.3 (actual values)
Caffeine	7 of 7 studies (2.6, 3.4, 1.8, 2.3, 1.9, 3.2, 2.5)	5 of 7 studies (0.25, 0.20, 0.29, 0.29, 0.31, 0.20, 0.33)
5-Fluorouracil	7 of 7 studies (18.0, 18.7, 6.7, 12.6, 8.5, 12.3, 12.3)	7 of 7 studies (0.07, 0.07, 0.21, 0.04, 0.16, 0.03, 0.05)
Saccharin	0 of 7 studies (1.0, 0.9, 1.0, 1.1, 1.0, 1.0, 1.0)	0 of 7 studies (1.04, 1.02, 1.02, 0.81, 0.96, 0.80, 1.09)
Sodium cyclamate	0 of 7 studies (1.2, 1.3, 1.1, 1.0, 1.0, 1.3, 1.0)	0 of 7 studies (0.91, 0.77, 1.03, 0.67, 1.05, 0.57, 0.96)
Proportion of study results in agreement	4 of 4 (100%)	3 of 4 (75%)

*Concordance among studies based on agreement in obtaining a TI >1.5 or an MCIG/LC_{50} <0.30.

Data from Bantle et al. (1994b), organized in sequence by laboratory number.

Table 24. Phase II Validation Study—Concordance Within the Same Laboratory in Obtaining a Significant FETAX Response Based on Single Decision Criteria*

Chemical Tested	TI >1.5 (actual values)	MCIG/LC_{50} <0.3 (actual values)
Caffeine	2 of 2 studies (2.6, 3.4)	2 of 2 studies (0.20, 0.25)
5-Fluorouracil	2 of 2 studies (18.0, 18.7)	2 of 2 studies (0.07, 0.07)
Saccharin	0 of 2 studies (0.9, 1.0)	0 of 2 studies (1.02, 1.04)
Sodium cyclamate	0 of 2 studies (1.2, 1.3)	0 of 2 studies (0.77, 0.91)
Proportion of study Results in agreement	4 of 4 (100%)	4 of 4 (100%)

*Concordance among studies based on agreement in obtaining a TI >1.5 or an MCIG/LC_{50} <0.30.

Data from Bantle et al. (1994b), organized by numeric value.

7.2.3 FETAX Phase III.1 Validation Study (Bantle et al., 1996)

The Phase III.1 Validation Study involved the testing of six substances (β-aminopropionitrile, ascorbic acid, copper sulfate, monosodium glutamate, sodium acetate, and sodium arsenate) (**Appendix 18**). Information on the teratogenic activity of these substances can be found in **Appendix 4**. Information on the teratogenicity of β-aminopropionitrile was not located. Ascorbic acid is a non-teratogen in rat s, mice, and humans. Copper sulfate is a teratogen in mice. Monosodium glutamate and sodium acetate are non-teratogens in mice. Sodium arsenate is a teratogen in rats and mice. Where information on the negative or positive teratogenicity of a test substance in a specific species is not provided above, relevant information was not located. Four substances were tested in six laboratories, with one laboratory conducting each study twice using different technicians (i.e., there were seven studies). The remaining two substances were tested in six laboratories. In the publication (Bantle et al., 1996), information on which laboratory conducted the independent replicate studies was not provided and the within-laboratory results were not discussed. All substances were tested without metabolic activation. Consistent with the ASTM FETAX Guideline (1991), a concurrent positive control was not included in the study design. Coded substances were used, and each participant was responsible for dose selection. Quantitative information on induced malformations was not provided. It was stated that the ASTM FETAX Guideline (1991) was followed with the exceptions noted for the FETAX Phase II Validation Study.

All seven studies with β-aminopropionitrile resulted in a TI value greater than 1.5, while six of the seven studies resulted in an MCIG/LC_{50} ratio less than 0.30. For ascorbic acid, three of six studies resulted in a TI value greater than 1.5 and an MCIG/LC_{50} ratio less than 0.30. For copper sulfate, five of seven studies resulted in a TI value greater than 1.5, while four of the seven studies resulted in an MCIG/LC_{50} ratio less than 0.30. For monosodium glutamate, four of six studies resulted in a TI value greater than 1.5, while one of six studies resulted in an MCIG/LC_{50} ratio less than 0.30. For sodium acetate, five of seven studies resulted in a TI value greater than 1.5, while two of seven studies resulted in an MCIG/LC_{50} ratio less than 0.30. For sodium arsenate, six of seven studies resulted in a TI value greater than 1.5, while one of seven studies

resulted in an MCIG/LC_{50} ratio less than 0.30. These data are tabulated in **Table 25**. Data responses for substances tested twice in one laboratory by different technicians is presented in **Table 26**. Based on the data in **Table 25**, an assessment was made by NICEATM of the extent of intra-laboratory concordance for the single laboratory that tested each substance twice using different technicians. Concordance for the TI value was obtained for all six substances tested, while concordance for the MCIG/LC_{50} was obtained for only one of the six test substances.

Table 25. Phase III.1 Validation Study—Concordance among Laboratories in Obtaining a Significant FETAX Response Based on Single Decision Criteria*

Chemical Tested	TI >1.5 (range of values)	MCIG/LC_{50} <0.3 (range of values)
β-Aminopropionitrile	7 of 7 studies (1221.0, 97.9, 4.5, 35.4, 7.1, 40.1, 72.2)	6 of 7 studies (<0.01, 0.40, 0.08, 0.01, <0.01, 0.03, <0.01)
Ascorbic acid	3 of 6 studies (1.7, 2.1, 1.3, 2.5, 1.3, no data, 1.0)	3 of 6 studies (0.76, 0.20, 0.22, 0.20, 0.84, no data, 1.08)
Copper sulfate	5 of 7 studies (5.6, 3.8, 1.9, 2.3, 0.8, 1.1, 1.8)	4 of 7 studies (0.37, 0.35, 0.05, 0.09, 0.42, 0.23, 0.22)
Monosodium glutamate	4 of 6 studies (15.4, 7.4, no data, 1.7, 2.3, 1.2, 1.2)	1 of 6 studies (0.50, 0.20, no data, 0.36, 0.75, 0.47, 1.24)
Sodium acetate	5 of 7 studies (2.6, 7.5, 1.6, 1.6, 0.9, 1.4, 4.4)	2 of 7 studies (0.48, 0.05, 0.80, 0.17, 1.09, 0.47, 0.48)
Sodium arsenate	5 of 6 studies (5.3, 7.0, 1.5, no data, 1.6, 1.4, 4.0)	1 of 6 studies (0.22, 0.33, 0.57, no data, 0.72, 0.41, 0.72)
Proportion of study results in agreement	1 of 6 (17%)	0 of 6 (0%)

* Concordance among studies based on agreement in obtaining a TI >1.5 or an MCIG/LC_{50} <0.30.

Data from Bantle et al. (1996), organized in sequence by laboratory number; "no data" indicates study not done.

Table 26. Phase III.1 Validation Study—Concordance Within the Same Laboratory in Obtaining a Significant FETAX Response Based on Single Decision Criteria*

Chemical Tested	TI >1.5 (range of values)	MCIG/LC_{50} <0.3 (range of values)
β-Aminopropionitrile	2 of 2 studies (97.9, 1221.0)	1 of 2 studies (0.001, 0.40)
Ascorbic acid	2 of 2 studies (1.7, 2.1)	1 of 2 studies (0.20, 0.76)
Copper sulfate	2 of 2 studies (3.8, 5.6)	0 of 2 studies (0.35, 0.37)
Monosodium glutamate	2 of 2 studies (7.4, 15.4)	1 of 2 studies (0.20, 0.50)
Sodium acetate	2 of 2 studies (2.6, 7.5)	1 of 2 studies (0.05, 0.48)
Sodium arsenate	2 of 2 studies (5.3, 7.0)	1 of 2 studies (0.22, 0.33)
Proportion of study results in agreement	6 of 6 (100%)	1 of 6 (17%)

* Concordance among studies based on agreement in obtaining a TI >1.5 or an MCIG/LC_{50} <0.30.
Data from Bantle et al. (1996), organized by numeric value.

(ASTM, 1991), an assessment was made by NICEATM of the extent of intra-laboratory concordance for the single laboratory that tested each substance twice using different technicians. Within laboratory concordance was obtained using a TI value greater than 1.5 for all three substances tested, while the concordance for the MCIG/LC_{50} ratio less than 0.30 was only 17% (one of six studies).

Individual laboratory results were compared using the statistical methodology described in ASTM (1992). The results of this analysis are presented graphically in **Appendix 7**. For β-aminopropionitrile, excessive intra-laboratory variability occurred for LC_{50}, EC_{50}, MCIG, and TI values within multiple laboratories; excessive inter-laboratory variability was present for MCIG and TI values within laboratory one and two, respectively. For ascorbic acid and sodium acetate,

excessive intra- and inter-laboratory variability was not present. For copper sulfate, excessive intra-laboratory variability occurred for EC_{50} and TI values within laboratories five and one, respectively; excessive inter-laboratory variability for EC_{50} values in laboratory five was present also. For monosodium glutamate, excessive intra-laboratory variability occurred for LC_{50} and MCIG values within laboratory one; excessive inter-laboratory variability was not present. For sodium arsenate, excessive intra-laboratory variability occurred for MCIG values within laboratory seven; excessive inter-laboratory variability was not present.

The overall CV(%) for the Phase III.1 Validation Study was relatively high; the overall mean CV(%) was 134.5%, with a range from 21.7% to 991.6%. As reported for the previous validation studies, variability was high among the laboratories for MCIG, but the highest variability was for the TI. Test substance concentration levels chosen by the independent laboratories and the lack of consistent *X. laevis* embryo evaluations may have contributed to the wide variation in results. The investigators recommended that the concentrations tested be standardized. Based on these results, the investigators concluded that FETAX is as repeatable and reliable as other standard bioassays similar to FETAX.

7.2.4 FETAX Phase III.2 Validation Study (Fort et al., 1998)

Two substances (caffeine and CP) were tested, both without and with metabolic activation (Aroclor 1254-induced rat liver S9 obtained from a common source) (**Appendix 19**). Information on the teratogenic activity of these substances can be found in **Appendix 4**. Caffeine is a teratogen in rats, mice, and rabbits; but not in humans, while CP is teratogenic in all four species. CP (when tested both with and without metabolic activation) and caffeine (when tested without metabolic activation), were evaluated in six laboratories, with one laboratory conducting each study twice using different technicians (i.e., there were seven studies). Caffeine (when tested with metabolic activation) was only evaluated in five laboratories. In the publication, information on which laboratory conducted the independent replicate studies was not provided and the within laboratory results were not discussed. Coded substances were used, and each participant was responsible for dose selection. Consistent with the ASTM FETAX Guideline (1991), a concurrent positive control was not included in the study design.

Quantitative information on induced malformations was not provided. The ASTM FETAX Guideline (1991) was adhered to with the exceptions noted for the FETAX Phase II Validation Study, and by the use of 20 and not 25 embryos per dish. This latter modification was necessitated by the use of plastic Petri dishes that were slightly smaller than the usual glass Petri dishes (Bantle et al., 1998). Plastic dishes are preferentially used in studies involving an MAS.

For CP, without metabolic activation, three of seven studies resulted in a TI value greater than 1.5, while none of seven studies resulted in an MCIG/LC_{50} ratio less than 0.30. With metabolic activation, five of seven studies resulted in a TI value greater than 1.5 and an MCIG/LC_{50} ratio less than 0.30. For caffeine, without metabolic activation, all six studies resulted in a TI value greater than 1.5, while four of six studies resulted in an MCIG/LC_{50} ratio less than 0.30. With metabolic activation, all six studies resulted in a TI value greater than 1.5, while none of the six studies resulted in an MCIG/LC_{50} ratio less than 0.30. These data are tabulated in **Table 27;** data responses for substances tested twice in one laboratory by different technicians are presented in **Table 28**. Based on the data in **Table 27**, an assessment was made by NICEATM of the extent of concordance among the studies conducted in obtaining a similar response (positive or negative) for each of the substances tested. When the TI value or the MCIG/LC_{50} ratio were considered, concordance among studies was obtained for two of the four test combinations. Based on the data provided in **Table 28** and the standard FETAX decision criteria (ASTM, 1991), an assessment was made by NICEATM of the extent of intra-laboratory concordance for the single laboratory that tested each substance twice using different technicians. Within laboratory concordance was 50% (two of four studies) for using a TI value greater than 1.5 or an MCIG/LC_{50} ratio less than 0.30.

The investigators averaged the TI values across laboratories and, based on the average value, concluded whether or not a positive teratogenic response was obtained. Within these studies, the control values exceeded those indicated as acceptable in the ASTM FETAX Guideline (1991) in one study investigating CP without metabolic activation, while the TI value for one study of CP with metabolic activation study was based on two replicates only. Data from these studies were included in the overall analysis; no explanation was provided for accepting data from studies that deviated from the 1991 ASTM FETAX Guideline.

Table 27. Phase III.2 Validation Study—Concordance among Laboratories in Obtaining a Significant FETAX Response Based on Single Decision Criteria*

Chemical Tested	TI >1.5 (range of values)	MCIG/LC_{50} <0.3 (range of values)
Caffeine without metabolic activation	6 of 6 studies (3.10, 8.65, 4.92, no data, 3.40, 2.87, 3.94)	4 of 6 studies (0.32, 0.13, 0.16, no data, 0.35, 0.25, 0.16)
Caffeine with metabolic activation	6 of 6 studies (2.34, 2.60, 2.53, no data, 1.76, 1.65, 2.66)	0 of 6 studies (0.55, 0.34, 0.40, no data, 0.56, 0.46, 0.32)
CP without metabolic activation	3 of 7 studies (1.52, 2.31, 1.48, 1.54, 1.29, 1.27, 1.35)	0 of 7 studies (0.69, 0.33, 0.41, 0.41, 0.69, 0.48, 0.67)
CP with metabolic activation	5 of 7 studies (8.37, 8.12, 1.31, 1.48, 1.71, 2.08, 3.15)	5 of 7 studies (0.29, 0.06, 0.12, 0.14, 0.38, 0.27, 0.33)
Proportion of study results in agreement	2 of 4 (50%)	2 of 4 (50%)

* Concordance among studies based on agreement in obtaining a TI >1.5 or an MCIG/LC_{50} <0.30.
Data from Fort et al. (1998), organized in sequence by laboratory number;
"no data" indicates study not done.

Individual laboratory results were compared using the statistical methodology described in ASTM (1992). The results of this analysis are presented graphically in **Appendix 7**. For CP, ,without metabolic activation, excessive intra-laboratory variability occurred for MCIG values in laboratory seven; excessive inter-laboratory variability was present for TI values in laboratory two. For CP, with metabolic activation, excessive intra-laboratory variability occurred for LC_{50} and MCIG values in laboratory three; excessive inter-laboratory variability was present for TI values in laboratories one and two. For caffeine, tested without metabolic activation, excessive

Table 28. Phase III.2 Validation Study—Concordance Within the Same Laboratory in Obtaining a Significant FETAX Response Based on Single Decision Criteria*

Chemical Tested	TI >1.5 (range of values)	MCIG/LC_{50} <0.3 (range of values)
Caffeine without metabolic activation	2 of 2 studies (3.10, 8.65)	1 of 2 studies (0.13, 0.32)
Caffeine with metabolic activation	2 of 2 studies (2.37, 2.6)	0 of 2 studies (0.34, 0.55)
CP without metabolic activation	2 of 2 studies (1.52, 2.34)	0 of 2 studies (0.33, 0.69)
CP with metabolic activation	2 of 2 studies (8.12, 8.37)	2 of 2 studies (0.06, 0.29)
Proportion of study results in agreement	4 of 4 (100%)	3 of 4 (75%)

* Concordance among studies based on agreement in obtaining a TI >1.5 or an MCIG/LC_{50} <0.30.

Data from Fort et al. (1998), organized by numeric value.

intra-laboratory variability occurred for MCIG values for one laboratory; excessive inter laboratory variability was present for TI. For caffeine, with metabolic activation, excessive intra-laboratory variability occurred for MCIG values in laboratory one; excessive inter-laboratory variability was not present.

The overall mean CV(%) for the Phase III.2 Validation Study for FETAX, without metabolic activation, was 26.0% with a range of 15.0 to 47.0%. In contrast, the overall mean CV(%) for FETAX with metabolic activation was 51.0% with a range of 18.0 to 131.0%. Again the MCIG and, hence, a lack of uniformity in evaluating embryo endpoints, seemed to be responsible for much of the variation, especially for FETAX with metabolic activation. The use of an MAS consistently increased the variability of FETAX.

The investigators concluded that bioactivated toxicants may be prone to higher variability due to the greater complexity of FETAX once an MAS is incorporated. However, they also concluded that the variability seen was not more than what would be expected for other aquatic-based bioassays.

7.2.5 FETAX Phase III.3 Validation Study (Bantle et al., 1999)

The Phase III.3 Validation Study (**Appendix 20**) involved the testing of 12 substances (acrylamide, boric acid, dichloroacetate, diethylene glycol, ethylene glycol, glycerol, phthalic acid, sodium arsenite, sodium bromate, sodium iodoacetate, tribromoacetic acid, and triethylene glycol dimethylether) in three laboratories with extensive FETAX experience. Information on the teratogenic activity of these substances can be found in **Appendix 4**. Acrylamide is not teratogenic in rats or mice. Boric acid is a teratogen in rats, mice, rabbits, but not in humans. Dichloroacetate is a teratogen in rats and mice, but not in humans. Diethylene glycol is not a teratogen in rabbits. Ethylene glycol is a teratogen in rats and mice, but not in rabbits. Glycerol is not teratogenic in rats, mice, or rabbits. Tribromoacetic acid is a teratogen in mice. Phthalic acid is tnon-eratogenic in rats and rabbits. Information on the teratogenicity of sodium arsenite, sodium bromate, and sodium iodoacetate in rats, mice, rabbits, or humans was not located. Triethylene glycol dimethylether is teratogenic in mice and rabbits. Where information on the teratogenicity of a test substance in a specific species is not provided above; relevant information was not located. All substances were tested using FETAX without and with metabolic activation. Coded substances were used, and each participant was responsible for dose selection. Consistent with the 1991 ASTM FETAX Guideline, a positive control was not included in the study design. Qualitative but not quantitative data on induced malformations were provided. The ASTM FETAX Guideline (1991) was followed with the exceptions noted for the Phase III.2 Validation Study.

For acrylamide, without metabolic activation, all three laboratories reported a TI value greater than 1.5, while two of three laboratories reported an MCIG/LC_{50} ratio less than 0.30. With metabolic activation, all three laboratories reported a TI value greater than 1.5, while one of three laboratories reported an MCIG/LC_{50} ratio less than 0.30.

For boric acid, without metabolic activation, all three laboratories reported a TI value greater than 1.5, while two of three laboratories reported an MCIG/LC_{50} ratio less than 0.30. With metabolic activation, all three laboratories reported a TI value greater than 1.5, while none of three laboratories reported an MCIG/LC_{50} ratio less than 0.30.

For dichloroacetate, without metabolic activation, one of three laboratories reported a TI value greater than 1.5, while none of three laboratories reported an MCIG/LC_{50} ratio less than 0.30. With metabolic activation, two of three laboratories reported a TI value greater than 1.5, while none of three laboratories reported an MCIG/LC_{50} ratio less than 0.30.

For diethylene glycol, without metabolic activation, all three laboratories reported a TI value greater than 1.5, while one of three laboratories reported an MCIG/LC_{50} ratio less than 0.30. With metabolic activation, all three laboratories reported a TI value greater than 1.5, while one of three laboratories reported an MCIG/LC_{50} ratio less than 0.30.

For ethylene glycol, without metabolic activation, all three laboratories reported a TI value greater than 1.5, while none of three laboratories reported an MCIG/LC_{50} ratio less than 0.30. With metabolic activation, two of three laboratories reported a TI value greater than 1.5, while none of three laboratories reported an MCIG/LC_{50} ratio less than 0.30.

For glycerol, without metabolic activation, one of three laboratories reported a TI value greater than 1.5, while none of three laboratories reported an MCIG/LC_{50} ratio less than 0.30. With metabolic activation, two of three laboratories reported a TI value greater than 1.5, while one of three laboratories reported an MCIG/LC_{50} ratio less than 0.30.

For phthalic acid, without metabolic activation, one of three laboratories reported a TI value greater than 1.5, while none of three laboratories reported an MCIG/LC_{50} ratio less than 0.30. With metabolic activation, one of three laboratories reported a TI value greater than 1.5, while none of three laboratories reported an MCIG/LC_{50} ratio less than 0.30.

For sodium arsenite, without metabolic activation, none of three laboratories reported a TI value greater than 1.5 or an MCIG/LC_{50} ratio less than 0.30. With metabolic activation, one of three laboratories reported a TI value greater than 1.5, while none of three laboratories reported an MCIG/LC_{50} ratio less than 0.30.

For sodium bromate, without metabolic activation, all three laboratories reported a TI value greater than 1.5, while two of three laboratories reported an MCIG/LC_{50} ratio less than 0.30. With metabolic activation, two of three laboratories reported a TI value greater than 1.5, while one of three laboratories reported an MCIG/LC_{50} ratio less than 0.30.

For sodium iodoacetate, without metabolic activation, one of three laboratories reported a TI value greater than 1.5, while two of two laboratories reported an MCIG/LC_{50} ratio less than 0.30. With metabolic activation, one of three laboratories reported a TI value greater than 1.5, while one of three laboratories did not report an MCIG/LC_{50} ratio less than 0.30.

For tribromoacetic acid, without metabolic activation, all three laboratories reported a TI value greater than 1.5, while none of three laboratories reported an MCIG/LC_{50} ratio less than 0.30. With metabolic activation, two of three laboratories reported a TI value greater than 1.5, while none of three laboratories reported an MCIG/LC_{50} ratio less than 0.30.

For triethylene glycol dimethylether, without metabolic activation, all three laboratories reported a TI value greater than 1.5, while two of three laboratories reported an MCIG/LC_{50} ratio less than 0.30. With metabolic activation, all three laboratories reported a TI value greater than 1.5 and an MCIG/LC_{50} ratio less than 0.30.

These data are tabulated in **Table 29a** (without metabolic activation) and **Table 29b** (with metabolic activation). Based on these data, an assessment was made by NICEATM of the extent of concordance among the laboratories in obtaining a similar response (positive or negative) for each of the substances tested. When TI was considered, concordance among studies was obtained for eight of twelve test substances (67%) without metabolic activation and for four

Table 29a. Phase III.3 Validation Study Without Metabolic Activation—Concordance among Laboratories in Obtaining a Significant FETAX Response Based on Single Decision Criteria*

Chemical Tested	TI >1.5 (actual values)	MCIG/LC_{50} <0.3 (actual values)
Acrylamide	3/3 (2.51, 4.68, 5.56)	2/3 (0.27, 0.37, 0.07)
Boric acid	3/3 (2.26, 5.95, 1.93)	2/3 (0.34, 0.09, 0.26)
Dichloroacetate	1/3 (1.13, 3.81, 1.38)	0/3 (0.57, 0.47, 0.93)
Diethylene glycol	3/3 (1.61, 2.28, 3.50)	1/3 (0.44, 0.47, 0.10)
Ethylene glycol	3/3 (1.61, 2.97, 1.71)	0/3 (0.53, 0.48, 0.53)
Glycerol	1/3 (1.35, 1.67, 1.41)	0/3 (0.85, 0.57, 0.39)
Phthalic acid	1/3 (1.11, 1.22, 2.51)	0/3 (0.91, 0.77, 0.94)
Sodium arsenite	0/3 (1.02, 1.32, 0.93)	0/3 (0.75, 0.66, 0.54)
Sodium bromate	3/3 (3.29, 4.12, 3.37)	2/3 (0.17, 0.23, 0.88)
Sodium iodoacetate	1/3 (0.29, 0.67, 2.56)	2/2 (0.06, no data, 0.10)
Tribromoacetic acid	3/3 (2.03, 3.89, 5.66)	0/3 (0.32, 0.47, 0.67)
Triethylene glycol dimethylether	3/3 (2.97, 4.42, 4.42)	2/3 (0.16, 0.26, 0.30)
Proportion of study results in agreement	8 of 12 (67%)	7 of 12 (58%)

* Concordance among studies based on agreement in obtaining a TI >1.5 or an MCIG/LC_{50} <0.30.

Data from Bantle et al. (1999), organized in sequence by laboratory number.

"no data" indicates study not done.

Table 29b. Phase III.3 Validation Study With Metabolic Activation—Concordance among Laboratories in Obtaining a Significant FETAX Response Based on Single Decision Criteria*

Chemical Tested	TI >1.5 (range of values)	MCIG/LC_{50} <0.3 (range of values)
Acrylamide	3/3 (3.55, 5.51, 4.75)	1/3 (0.36, 0.30, 0.09)
Boric acid	3/3 (2.02, 3.30, 1.86)	0/3 (0.54, 0.32, 0.30)
Dichloroacetate	2/3 (1.30, 5.85, 1.53)	0/3 (0.95, 0.84, 0.99)
Diethylene glycol	3/3 (2.46, 1.92, 3.12)	1/3 (0.57, 0.61, 0.09)
Ethylene glycol	2/3 (1.71, 3.78, 1.40)	0/3 (0.49, 0.46, 0.40)
Glycerol	2/3 (0.97, 1.67, 2.33)	1/3 (1.09, 0.57, 0.16)
Phthalic acid	1/3 (1.27, 1.27, 1.76)	0/3 (0.46, 0.78, 1.03)
Sodium arsenite	1/3 (1.31, 1.53, 1.20)	0/3 (0.73, 0.77, 3.55)
Sodium bromate	2/3 (1.21, 3.20, 2.44)	1/3 (0.39, 0.22, 0.98)
Sodium iodoacetate	1/3 (0.27, 0.28, 2.41)	0/1 (no data, no data, 0.47)
Tribromoacetic acid	2/3 (1.36, 3.57, 7.49)	0/3 (0.40, 0.49, 0.57)
Triethylene glycol dimethylether	3/3 (1.99, 3.48, 3.43)	3/3 (0.20, 0.22, 0.18)
Proportion of study results in agreement	4 of 12 (33%)	7 of 11 (64%)

* Concordance among studies based on agreement in obtaining a TI >1.5 or an MCIG/LC_{50} <0.30.

Data from Bantle et al. (1999), organized in sequence by laboratory number; "no data" indicates study not done.

of twelve test substances (33%) with metabolic activation. When the MCIG/LC_{50} ratio was used, concordance was obtained for seven of twelve test substances (58%) tested without metabolic activation and for seven of eleven test substances (64%) tested with metabolic activation. The lack of agreement among three highly experienced laboratories suggests that additional effort is needed in optimizing the FETAX protocol or the decision criteria to classify test substances as positive or negative for teratogenic activity.

In this validation study, the ASTM FETAX Guideline (1991) was not always followed in terms of having three independent replicates per study. The MCIG data (generated without metabolic activation) for sodium iodoacetate in one laboratory was based only on one of three replicates, while MCIG could not be determined in another laboratory. Similarly, the MCIG (generated with metabolic activation) for sodium iodoacetate could not be determined in two of three laboratories. The MCIG data (generated without metabolic activation) for phthalic acid in two laboratories were based on two replicates. In one laboratory, the LC_{50}, EC_{50}, and MCIG (generated with metabolic activation) for dichloroacetate were based only on a single replicate. No explanation was provided for including data from studies that deviated from the 1991 ASTM FETAX Guideline.

The validation study management team averaged the EC_{50}, LC_{50}, TI, and MCIG values across all replicate tests (even in the absence of a fully balanced design) and, based on that average value, concluded whether or not the test substance was positive (TI >1.5 and MCIG/LC_{50} <0.3), equivocal (one parameter was positive), or negative (neither parameter was positive). If equivocal, the types and incidence of malformations present were evaluated to clarify the equivocal nature of the classification. Based on this approach, the investigators concluded, for studies conducted with metabolic activation, that two substances were clearly teratogenic, four substances were non-teratogenic, and six substances were equivocal for teratogenic potential in laboratory mammals. For these 12 substances, based on a consensus evaluation of the available literature and other sources, the investigators concluded that seven substances were positive laboratory mammal teratogens, two were negative laboratory mammal teratogens, and three were equivocal laboratory mammal teratogens. An equivocal laboratory mammal teratogen was defined as having discordant teratogenic results among multiple non-human mammal species.

The conclusions made by each of the three laboratories for FETAX studies conducted with and without metabolic activation are shown in **Tables 30** and **31**, respectively. The distribution of NICEATM final conclusions for all substances tested with and without metabolic activation, as compared to the consensus call of the investigators for laboratory mammal teratogenicity, are provided in **Table 32**.

Based on the multiple decision criteria approach, there was agreement between FETAX studies conducted with and without metabolic activation for eight of 12 substances (67% concordance with two positive, one negative, and five equivocal classifications). Compared to the laboratory mammal results provided in the report, FETAX conducted without metabolic activation agreed five of 12 times (42% concordance with two positive, one negative, and two equivocal classifications). For studies conducted with metabolic activation, the FETAX classifications agreed with the laboratory mammal results for four of 12 times (33% concordance with one positive, one negative, and two equivocal classifications). These data do not support the expected increase in performance accuracy predicted for FETAX by the addition of metabolic activation, and suggest that the substances selected for testing with an MAS do not require metabolic activation.

Subsequent to comparing the results from studies conducted using metabolic activation against the laboratory mammal teratogenicity calls, the investigators concluded that basing FETAX conclusions on TI values greater than 1.5 resulted in better accuracy for identifying laboratory mammal teratogens than did the use of multiple decision criteria. These data, along with the results from FETAX conducted without metabolic activation, are provided in **Table 33**.

Using TI as the single criterion for assessing teratogenicity, there was concordance among FETAX studies conducted with and without metabolic activation for ten of 12 substances (83% with eight positive and two negative classifications). Compared to the laboratory mammal calls provided in the report, the studies conducted with and without metabolic activation both agreed

Table 30. Conclusions by Laboratory for Substances Tested Without Metabolic Activation as Determined Using Multiple Criteria (TI >1.5 plus MCIG/LC_{50} <0.3)

Chemical	Laboratory 1	Laboratory 2	Laboratory 3
Acrylamide	+	E	+
Boric acid	E	+	+
Dichloroacetate	-	E	-
Diethylene glycol	E	E	+
Ethylene glycol	E	E	E
Glycerol	-	E	-
Phthalic acid	-	-	E
Sodium arsenite	-	-	-
Sodium bromate	+	+	E
Sodium iodoacetate	E	No data	+
Tribromoacetic acid	E	E	E
Triethylene glycol Dimethylether	+	+	E
Proportion of study results in agreement		3 of 12 (25%)	

+ = positive for FETAX teratogenicity based on TI >1.5, MCIG/LC_{50} <0.3, and the presence of malformations; consensus positive for laboratory mammal teratogenicity as concluded in Bantle et al. (1999).
- = negative for FETAX teratogenicity based on TI <1.5, MCIG/LC_{50} >0.3, and the lack of malformations; consensus negative for laboratory mammal teratogenicity as concluded in Bantle et al. (1999).
E = equivocal for FETAX teratogenicity based on having a positive response in at least one but not all three FETAX or two parameters (TI >1.5, MCIG/LC_{50} <0.3, presence of malformations); consensus equivocal for laboratory mammal teratogenicity as concluded in Bantle et al. (1999), based on species differences in response.
No data=data not provided for the MCIG/LCC_{50}, and thus the multiple criterion could not be evaluated.

six of 12 times (50% with five positive and one negative classifications). If the equivocal laboratory mammal conclusions are re-classified as mammal teratogens, FETAX studies conducted with and without metabolic activation agreed with the consensus laboratory mammal teratogenicity results nine of 12 times (75% with eight positive and one negative classification). In reviewing these data, the investigators argued that substances with TI values in the range of

Table 31. Conclusions by Laboratory for Substances Tested With Metabolic Activation as Determined Using Multiple Criteria (TI>1.5 plus MCIG/LC_{50} <0.3)

Chemical	Laboratory 1	Laboratory 2	Laboratory 3
Acrylamide	E	E	+
Boric acid	E	E	+
Dichloroacetate	-	E	E
Diethylene glycol	E	E	+
Ethylene glycol	E	E	-
Glycerol	-	E	+
Phthalic acid	-	-	E
Sodium arsenite	-	E	-
Sodium bromate	-	+	E
Sodium iodoacetate	E or -	E or -	E
Tribromoacetic acid	-	E	E
Triethylene glycol Dimethylether	+	+	+
Proportion of study results in agreement		1 of 12 (8%)	

+ = positive for FETAX teratogenicity based on TI >1.5, MCIG/LC_{50} <0.3, and presence of malformations; consensus positive for laboratory mammal teratogenicity as concluded in Bantle et al. (1999).
- = negative for FETAX teratogenicity based on TI <1.5, MCIG/LC_{50} >0.3, and lack of malformations; consensus negative for laboratory mammal teratogenicity as concluded in Bantle et al. (1999).
E = equivocal for FETAX teratogenicity based on having a positive response in at least one but not all three FETAX parameters (TI >1.5, MCIG/LC_{50} <0.3, presence of malformations); consensus equivocal for laboratory mammal teratogenicity as concluded in Bantle et al. (1999), based on species differences in response.

1.5 to 2.5 make identification of teratogenicity difficult. These data again do not support the expected increase in performance accuracy for FETAX by the addition of metabolic activation.

Individual laboratory results were compared by NICEATM using the statistical methodology described in ASTM (1992). The results of this analysis are presented graphically in **Appendix 7**.

For studies conducted without metabolic activation, excessive inter- and/or intra-laboratory variability for at least one endpoint was present for nine of 12 test substances. In terms of

Table 32. Using Multiple Criteria, Comparison of Consensus FETAX Conclusions, With or Without Metabolic Activation, to the Consensus Non-Human Mammalian Teratogenicity Conclusions

Chemical	Without Metabolic Activation Classification	With Metabolic Activation Classification	Mammalian Consensus Classification
Acrylamide	+	+	E
Boric acid	+	E	+
Dichloroacetate	E	E	+
Diethylene glycol	E	E	E
Ethylene glycol	E	E	E
Glycerol	-	E	-
Phthalic acid	E	-	-
Sodium arsenite	-	-	+
Sodium bromate	E	E	+
Sodium iodoacetate	E	-	+
Tribromoacetic acid	E	E	+
Triethylene glycol Dimethylether	+	+	+

+ = positive for FETAX teratogenicity based on TI >1.5, MCIG/LC_{50} <0.3, and presence of malformations; consensus positive for laboratory mammal teratogenicity as concluded in Bantle et al. (1999).
- = negative for FETAX teratogenicity based on TI <1.5, MCIG/LC_{50} >0.3, and lack of malformations; consensus negative for laboratory mammal teratogenicity as concluded in Bantle et al. (1999).
E = equivocal for FETAX teratogenicity based on having a positive response in at least one but not all three FETAX parameters (TI >1.5, MCIG/LC_{50} <0.3, presence of malformations); consensus equivocal for laboratory mammal teratogenicity as concluded in Bantle et al. (1999), based on species differences in response.

repeatability, only laboratory one did not exhibit excessive variability for any endpoint; laboratory two exhibited excessive variability for LC_{50} values (one test substance), TI values (one test substance), and MCIG values (three test substances); and laboratory three exhibited excessive variability for LC_{50} values (one test substance), TI values (one test substance), and

Table 33. Distribution of TI Values >1.5 for FETAX, With or Without Metabolic Activation, Compared to the Consensus Non-Human Mammalian Teratogenicity Conclusions

Chemical	Without Metabolic Activation Consensus Conclusion	With Metabolic Activation Consensus Conclusion	Laboratory Mammal Consensus Conclusion
Acrylamide	+ (4.25)*	+ (4.60)	E
Boric acid	+ (3.38)	+ (2.39)	+
Dichloroacetate	+ (2.11)	+ (2.89)	+
Diethylene glycol	+ (2.47)	+ (2.50)	E
Ethylene glycol	+ (2.10)	+ (2.30)	E
Glycerol	- (1.48)	+ (1.66)	-
Phthalic acid	+ (1.61)	- (1.43)	-
Sodium arsenite	- (1.09)	- (1.35)	+
Sodium bromate	+ (3.59)	+ (2.28)	+
Sodium iodoacetate	- (1.17)	- (0.99)	+
Tribromoacetic acid	+ (3.86)	+ (4.14)	+
Triethylene glycol dimethylether	+ (3.94)	+ (2.97)	+

+ = positive for FETAX teratogenicity based on TI >1.5; consensus positive for laboratory mammal teratogenicity as concluded in Bantle et al. (1999);
- = negative for FETAX teratogenicity based on TI <1.5; consensus negative for laboratory mammal teratogenicity as concluded in Bantle et al. (1999).
* Mean TI value, based on individual replicate definitive tests across laboratories.

MCIG values (one test substance). In terms of reproducibility, laboratory one exhibited excessive variability for MCIG values (three test substances); laboratory two exhibited excessive

variability for TI values (four test substances) and MCIG values (one test substance); and laboratory three exhibited excessive variability for LC_{50} values (one test substance), EC_{50} values (two test substances), TI values (one test substance), and MCIG values (three test substances).

For studies conducted with metabolic activation, excessive inter- and/or intra-laboratory variability for at least one endpoint was present for 11 of 12 test substances. In terms of repeatability, laboratory one exhibited excessive variability for EC_{50} values (two test substances) and MCIG values (one test substance); laboratory two exhibited excessive variability for TI values (one test substance) and MCIG (one test substance); laboratory two exhibited excessive variability for TI values (one test substance) and MCIG values (one test substance); and laboratory three exhibited excessive variability for TI values (one test substance) and MCIG values (two test substances). In terms of reproducibility, laboratory one exhibited excessive variability for LC_{50} values (one test substance), EC_{50} values (two test substances), TI values (one test substance), and MCIG values (one test substance); laboratory two exhibited excessive variability for EC_{50} values (one test substance), TI values (three test substances), and MCIG values (one test substance); and laboratory three exhibited excessive variability for LC_{50} values (two test substances), EC_{50} values (two test substances), TI values (two test substances), and MCIG values (two test substances).

The overall mean CV(%) for the Phase III.3 Validation Study for FETAX without metabolic activation was 38.0%, with a range of 9.5 to 87.2%. In contrast, the overall mean CV(%) for FETAX with metabolic activation was 51.1%, with a range of 2.3 to 166.6%. As occurred during the Phase III.2 FETAX Validation Study, incorporation of metabolic activation resulted in more variability than studies without metabolic activation, and MCIG values exhibited the largest variation.

Conclusions made by the participants in this most recent validation study were:

- There was difficulty in producing an adequate decision process for classifying FETAX results as positive, negative, or equivocal.

- Further research was needed to establish procedures for obtaining a more accurate MCIG.

- Using an MCIG/LC_{50} ratio less than 0.3 as a criterion for a positive response may be too strict and needs further evaluation.

- Classification of *Xenopus* malformations as a criterion for evaluating teratogenic potential in FETAX was too subjective and needs further consideration.

- An MAS was essential in using FETAX to predict developmental hazard in mammals but required further development.

- FETAX intra- and inter-laboratory variability were very low and the assay yielded repeatable and reliable data as long as care was taken during the range-finding assay and technicians were adequately trained.

NICEATM is in agreement with the first five conclusions while the last conclusion does not appear to take into account the extent of variability among laboratories in obtaining similar FETAX results (i.e., negative or positive) based on the decision criteria used.

7.3 Additional Evaluations Conducted by NICEATM

7.3.1 Inter-Laboratory CV Data for All FETAX Validation Studies

For visual comparative purposes, the inter-laboratory CV data for all FETAX validation studies are summarized in **Table 34a** (without metabolic activation) and **Table 34b** (with metabolic activation). Where studies were conducted with and without metabolic activation, inter-laboratory CV values were higher with metabolic activation than without metabolic activation for the same test substances. The possible source(s) of this increased variability warrants investigation. The inter-laboratory CV for MCIG values, except for the first validation study, were generally no greater than that observed for TI values.

To place the inter-laboratory CV values obtained for FETAX in perspective, corresponding CV values for three *in vitro* corrosivity assays are provided. It is fully appreciated that these assays do not use aquatic organisms, nor do they involve developmental endpoints; these differences may alter expectations for what constitutes reasonable CV values. However, all three assays were evaluated for inter-laboratory reproducibility in the same ECVAM-sponsored validation study (Fentem et al., 1998). This increases the comparability of the CV data for these three assays. Appropriate CV data for assays more directly comparable to FETAX is being sought by NICEATM.

The rat skin Transcutaneous Electrical Resistance (TER) assay, the Episkinä assay, and Corrositex® have been evaluated as potential replacement assays for *in vivo* corrosivity testing (Fentem et al., 1998). In the TER assay, test materials are applied up to 24 hours to the epidermal surfaces of skin discs taken from the skin of humanely killed young rats. Corrosive materials are identified by the ability to produce a loss of normal stratum corneum integrity and barrier function, which is measured as a reduction of the inherent transcutaneous electrical resistance below a predetermined threshold level. Episkinä is a three dimensional human skin model comprised of a reconstructed epidermis and a functional stratum corneum. For use in corrosivity testing, the test material is topically applied to the surface of the skin for 3, 60, and 240 minutes, with subsequent assessment of their effects on cell viability. Corrositex® is based on the ability of a corrosive chemical or chemical mixture to pass through a biobarrier, by diffusion and/or destruction/erosion, and to elicit a color change in the underlying liquid Chemical Detection System.

In the ECVAM validation study, three laboratories each tested 60 test chemicals in three independent tests (Fentem et al., 1998). The median inter-laboratory CV was 34.7% (range of 3.8% to 322%) for TER, 11.3% (range 3.9% to 148.8%) for Episkinä, and 30.3% (range 7.7% to 252.5%) for Corrositex®. These values are not greatly different from the overall median CV values and ranges obtained for FETAX in the Phase III.3 Validation Study, with (51.1%, with a range of 2.3% to 166.6%) and without metabolic activation (38.0%, with a range of 9.5% to 87.2%).

Table 34a. Comparison of Coefficient of Variation (CV) Results for All Validation Studies—FETAX Without Metabolic Activation

FETAX Without Metabolic Activation	Phase I (Bantle et al., 1994a)	Phase II (Bantle et al., 1994b)	Phase III.1 (Bantle et al., 1996)	Phase III.2 (Fort et al., 1998)	Phase III.3 (Bantle et al., 1999)
Number of Chemicals	3	4	6	2	12
Number of Participating Laboratories	7[a]	7[a]	7[a,b]	7[a]	3
Inter-laboratory LC_{50} CV mean (range) (%)	48.5 (20.5-75.2)	21.0 (8.7-44.8)	56.6 (21.7-108.2)	23.0 (15.0-31.0)	26.6 (9.5-69.4)
Inter-laboratory EC_{50} CV mean (range) (%)	49.0 (32.7-70.1)	23.1 (10.7-41.0)	83.9 (53.0-134.9)	17.0 (15.0-18.0)	35.6 (19.3-70.3)
Inter-laboratory TI CV mean (range) (%)	58.4 (39.2-82.9)	26.8 (12.1-41.6)	290.0 (46.3-991.6)	36.0 (25.0-47.0)	41.6 (15.0-87.2)
Inter-laboratory MCIG CV mean (range) (%)	109.6 (63.0-201.5)	26.5 (7.3-54.7)	107.4 (44.5-261.1)	30.0 (29.0-31.0)	48.0 (13.2-84.8)
Overall CV mean And range (%)	66.3 (20.5-201.5)	24.4 (7.3-54.7)	134.5 (21.7-991.6)	26.0 (15.0-47.0)	38.0 (9.5-87.2)

Abbreviations: CV = Coefficient of Variation, EC_{50} = Effective Concentration (i.e., Concentration Inducing Malformation in 50% of Exposed Embryos), LC_{50} = Lethal Concentration (i.e., Concentration Inducing Death in 50% of Exposed Embryos), MCIG = Minimum Concentration to Inhibit Growth, TI = Teratogenic Index.

[a] Six laboratories participated with one laboratory conducting each study twice using different technicians.

[b] Six studies instead of seven carried out evaluations for three of the six substances tested.

Table 34b. A Comparison of Coefficient of Variation (CV) Results for All Validation Studies—FETAX With Metabolic Activation

FETAX With Metabolic Activation	Phase I (Bantle et al., 1994a)	Phase II (Bantle et al., 1994b)	Phase III.1 (Bantle et al., 1996)	Phase III.2 (Fort et al., 1998)	Phase III.3 (Bantle et al., 1999)
Number of substances	0	0	0	2	12
Number of Participating Laboratories	7[a]	7[a]	7[a,b]	7[a]	3
Inter-laboratory LC_{50} CV mean (range) (%)	N/A	N/A	N/A	36.0 (18.0-53.0)	41.9 (19.9-114.0)
Inter-laboratory EC_{50} CV mean (range) (%)	N/A	N/A	N/A	42.0 (19.0-64.0)	54.5 (26.7-166.6)
Inter-laboratory TI CV mean (range) (%)	N/A	N/A	N/A	52.0 (21.0-83.0)	51.4 (22.2-111.5)
Inter-laboratory MCIG CV mean (range) (%)	N/A	N/A	N/A	76.0 (20.0-131.0)	56.5 (2.3-79.0)
Overall CV mean (range) (%)	N/A	N/A	N/A	51.0 (18.0-131.0)	51.1 (2.3-166.6)

Abbreviations: CV = Coefficient of Variation, EC_{50} = Effective Concentration (i.e., Concentration Inducing Malformation in 50% of Exposed Embryos), LC_{50} = Lethal Concentration (i.e., Concentration Inducing Death in 50% of Exposed Embryos), MCIG = Minimum Concentration to Inhibit Growth, TI = Teratogenic Index.

[a] Six laboratories participated with one laboratory conducting each study twice using different technicians.

[b] Six laboratories instead of seven carried out evaluations for three of the six substances tested.

7.3.2 Inter- and Intra-Laboratory Reliability of FETAX Studies on Caffeine

One substance, caffeine, has been tested, without metabolic activation, in two FETAX validation studies—Phase II and Phase III.2. The same six laboratories (with one laboratory conducting replicate studies) participated in each validation study. After obtaining the laboratory codes from the investigators, NICEATM evaluated the inter- and intra-laboratory repeatability and reproducibility for caffeine across both validation studies. Excessive inter-laboratory variability was found for TI values (one laboratory) and MCIG values (one laboratory) (**Figure 5**). Excessive intra-laboratory variability was found for LC_{50} values within one laboratory (**Figure 6**).

7.3.3 Assessment of the Effect of Malformation Identification Expertise on FETAX Performance

In some of the FETAX validation studies, it was suggested that the excess inter-laboratory variability may be a direct reflection of the difficulty of evaluating *X. laevis* embryos for malformations and that the level of expertise in identifying malformations may have varied widely among the participating laboratories. NICEATM attempted to assess the effect of expertise on performance by comparing the performance characteristics for FETAX data, with and without metabolic activation, generated by the two most highly experienced laboratories (i.e., the laboratories of Drs. J. Bantle and D. Fort) against that collected for all laboratories (including Drs. Bantle and Fort). The database was limited to those substances tested by Drs. Bantle and Fort and also by laboratories not associated with these two investigators. These data were compared to both combined laboratory mammal (i.e., rat, mouse, and rabbit) and human teratogenicity data (**Table 35**). Because FETAX performance characteristics were not found to be significantly altered when either single decision criteria (i.e. TI >1.5, TI >3.0, MCIG/LC_{50} <0.3) or multiple decision criteria (TI >1.5 plus MCIG/LC_{50}<0.3, TI > 3.0 plus MCIG/LC50 <0.3) were used, this analysis focused on performance characteristics using single decision criteria only. As was done in the other performance analyses, classification of the FETAX results as positive or negative for each of the single decision criteria were based on a weight-of-evidence approach. The number of substances contributing to the performance calculations are

different for the two data sets because of the presence of some substances with equivocal (i.e., an equal number of positive and negative) responses in the data set limited to only experienced laboratory results. Also, in this analysis, a substance tested with and without metabolic activation was classified as positive in FETAX if a consensus positive response was obtained either with or without metabolic activation. A test substance tested with and without metabolic activation was classified as a FETAX negative only if a positive response was not obtained using either exposure condition.

With very few exceptions, performance (i.e., accuracy, sensitivity, specificity, positive predicitivity, negative predictivity, and false positive and false negative rates) for FETAX, with and without metabolic activation, compared to either laboratory mammal or human teratogenicity results, were altered by only one to two percentage points when the analysis was limited to the two most experienced laboratories. Based on these results, it does not appear that the level of expertise is significantly different among the participating laboratories. Alternatively expertise is playing, at best, only a minor role in the variability of the assay and other factors should be investigated further.

7.4 Summary of Historical Positive and Negative Control Data

The recommended solvent for FETAX is FETAX Solution (i.e., medium for culturing *Xenopus* embryos). If a solvent other than FETAX Solution is used, its concentration in the FETAX Solution must be demonstrated to not adversely affect *Xenopus* embryo growth and survival. Because of its low toxicity, low volatility, and high ability to dissolve many organic substances, triethylene glycol is often a good organic solvent for preparing stock solutions. Other water-miscible organic solvents such as dimethyl sulfoxide and acetone also may be used. If a solvent other than dilution-water or FETAX Solution is used, at least one solvent control test group, using solvent from the same batch used to make the stock solution, must be included in the test. A dilution-water or FETAX Solution control should also be included in the test. If no solvent other than dilution-water or FETAX Solution is used, then a dilution-water or FETAX Solution control must be included in the test. The 1991 ASTM Guideline states that for negative or solvent controls, the percentage of malformed embryos must not exceed 7%, while mean

survival must be greater than 90% (ASTM, 1991). However, in the FETAX Phase I Validation Study (Bantle et al., 1994a), the investigators concluded that the negative control percentage of malformed embryos should not exceed 10% and this change has been reflected in the revised 1998 ASTM FETAX Guideline. In the published FETAX literature, quantitative negative/solvent control data were included only sporadically. In almost all cases, general statements were made that suitable negative control data were obtained but no supporting data were provided.

Based on the ASTM FETAX Guideline (1991, 1998), concentration-response experiments without metabolic activation should be performed at least quarterly and the results of these tests compared with historical tests to judge the laboratory quality of FETAX data. The reference toxicant test must produce data within two standard deviations of the historical mean values. The recommended reference substance for studies conducted without metabolic activation is 6-AN (ASTM, 1991; 1998), as this substance presents a mortality and malformation database convenient for reference purposes. However, in the FETAX Phase I Validation Study (Bantle et al., 1994a), the investigators concluded that 6-AN may not be suitable as the positive control based on the extensive variability observed among the participating laboratories. A replacement reference control has not been designated (ASTM, 1998). In the published FETAX literature, quantitative 6-AN (or any other reference agent) control data were not included; general statements were made that suitable positive control data were obtained.

The recommended concurrent bioactivation positive control for studies conducted with metabolic activation is CP at a concentration of 4 mg/mL. The metabolic activation-only control and the CP only control should result in less than 10% mortality and malformations. With metabolic activation, bioactivated CP should kill 100% of the embryos within 96 hours. The appropriateness of using CP at a concentration that results in 100% mortality raises concern. A response of this magnitude limits a statistical consideration of historical data. Also, as the TI is considered a primary measure of teratogenic potential, it may be more informative if a concentration of CP is used that allows for an assessment of malformations, as well as mortality. In the published FETAX literature, quantitative CP control data were not included; general statements were made that suitable positive control data were obtained.

To evaluate historical FETAX data, appropriate data needs to be obtained from multiple laboratories.

7.5 Limitations of FETAX in Regard to Test Method Reliability (as determined by NICEATM)

Limitations associated with FETAX in regard to test method reliability include:

- Excessive variability in LC_{50}, EC_{50}, TI, and MCIG values among highly experienced laboratories, especially in regard to MCIG.

- Lack of agreement among highly experienced laboratories in FETAX study results, based on the single decision criteria set forth in the ASTM FETAX Guideline (1991, 1998) and multiple decision criteria used in various validation studies.

- The lack of readily available historical negative and positive control data for FETAX.

- The limited database for studies with metabolic activation.

7.6 Data Interpretation Issues

The ASTM FETAX Guideline (1991, 1998) specifies the calculation and use of the geometric mean in identifying teratogenic activity. However, the arithmetic mean was used throughout FETAX publications. The effects of this difference on the interpretation of FETAX data is not known. Also, the use of a two-point graphical method for determining the EC_{50} and LC_{50} values may be difficult to interpret.

In the FETAX validation studies, the validation study management team determined the average of the calculated LC_{50}, EC_{50}, TI, and MCIG values across all replicate definitive tests (generally three replicate definitive tests per compound per participating laboratory). The conclusion as to

the potential teratogenicity of a test substance was then based on the average TI value and the average ratio of the MCIG to the LC_{50}. This method for achieving a consensus conclusion does not take into account the variability among laboratories in reaching their own conclusion as to the potential teratogenicity of the test substance. In contrast, NICEATM used a weight-of-evidence approach based on the results obtained for each laboratory. In this approach, a test substance was classified as positive in FETAX if a majority of laboratories obtained a positive result. Similarly, a test substance was classified as negative in FETAX if a majority of laboratories obtained a negative result. In situations where an equal number of positive and negative studies were available for consideration, the test substance was classified as equivocal and excluded from any analysis. The relative merit of each approach should be assessed.

In a number of FETAX studies, less than three definitive replicates were used to define a FETAX response. The effect of this reduction in replicates on the performance characteristics of FETAX is not known.

In the validation studies, there was excessive variability within and across laboratories in FETAX data, especially in regard to the calculation of the MCIG. This variability may indicate inherent technical difficulties with the FETAX protocol as currently conducted and adversely impacts on the credibility and usefulness of the data for hazard identification.

In addition, where the same substances was tested in multiple laboratories, there was generally poor concordance in regard to the classification of test substances as potential teratogens, even when highly experienced laboratories were involved. This may indicate difficulty with the criteria used to judge a test substance as a FETAX teratogen. This perceived problem also adversely impacts on the credibility and usefulness of the data for hazard identification. In more recent publications, both a TI value greater than 1.5 and an $MCIG/LC_{50}$ less than 0.3 have been used singly and in combination (along with malformation data) to identify teratogens and non-teratogens. A justification for either criteria was not provided. An evaluation of corresponding TI and $MCIG/LC_{50}$ data for each substance tested within each validation study did not reveal a direct correlation between the two indices of teratogenicity and emphasizes the extent of inter-laboratory variability. The relative concordance between a TI value greater than 1.5 and an

Table 36. Concordance between TI >1.5 and MCIG/LC_{50} <0.3 for All Validation Studies

FETAX	**Phase I** (Bantle et al., 1994a)	**Phase II (**Bantle et al., 1994b)	**Phase III.1** (Bantle et al., 1996)	**Phase III.2** (Fort et al., 1998)	**Phase III.3** (Bantle et al., 1999)
Without Metabolic Activation					
Number of substances	3	4	6	2	12
Number of Participating Labs	7[a]	7[a]	7[a,b]	7[a]	3
Number (%) of Concordant Data	11 of 20 trials (55%)	26 of 28 trials (93%)	24 of 39 trials (62%)	8 of 13 Trials (62%)	20 of 35 trials (57%)
With Metabolic Activation					
Number of substances	None	None	None	2	12
Number of Participating Labs				7[a]	3
Number (%) of Concordant Data				3 of 13 trials (23%)	16 of 34 trials (47%)

[a] Six laboratories participated with one laboratory conducting each study twice using different technicians.
[b] Six laboratories instead of seven carried out evaluations for three of the six substances tested.

MCIG/LC_{50} ratio less than 0.3 are tabulated, by validation study, in **Table 36**. For the 12 substances tested in the Phase III.3 Validation Study, the most recent validation study, the extent of concordance for the two indices of teratogenic activity without and with metabolic activation,

was only 57% and 47%, respectively. This lack of concordance adversely impacts on the usefulness of using both decision criteria for hazard identification.

Another issue affecting data interpretation is the utility of an exogenous MAS in FETAX. As indicated, the database for substances tested with metabolic activation is limited to only 35 substances. In the validation studies where the same substances are tested without and with metabolic activation, there is no increase in assay performance. Instead, there is an increase in inter-laboratory variability and an associated decrease in concordance. The rationale for the selection of substances to test without and with metabolic activation during the validation process is not clear, as most of the substances tested are not known to be activated to teratogens by metabolic activation. The utility of an exogenous MAS and the appropriateness of the MAS ingredients used requires further assessment.

7.7 Section 7 Conclusions

In the FETAX validation studies, the assessment of FETAX inter-laboratory reproducibility was adequate, and indicated excessive variability in most validation studies. The corresponding assessment of FETAX intra-laboratory repeatability was limited to an analysis of the three definitive replicates used to define a FETAX study. NICEATM concluded that this analysis may not have been completely appropriate and conducted an independent analysis based on the results for the same substance tested more than once in the same laboratory. In either case, excessive variability was noted within laboratories.

Excessive inter-laboratory variability occurred in some of the FETAX validation studies and the investigators speculated that the variability may have resulted from differences in expertise for scoring malformations in *Xenopus*. However, an analysis by NICEATM determined that, with very few exceptions, performance for FETAX, with and without metabolic activation, against either laboratory mammal or human teratogenicity results were not altered significantly when the analysis was limited to the laboratories of the two most experienced investigators. These results suggest that expertise plays, at best, only a minor role in the variability of the assay and that other factors should be investigated.

In the validation studies, there was excessive variability in FETAX data within and across laboratories, especially in regard to the calculation of the MCIG. This variability may indicate inherent technical difficulties with the FETAX protocol as currently conducted and adversely impacts on the usefulness of the data for hazard identification. In addition, where the same substance was tested in multiple laboratories, there was generally poor concordance in regard to the classification of test substances as potential teratogens, even when highly experienced laboratories were involved. This may indicate difficulty with the criteria used to judge a test substance as a FETAX teratogen. This perceived problem also adversely impacts on the credibility and usefulness of the data for hazard identification.

In more recent publications, both a TI value greater than 1.5 and a MCIG/LC_{50} ratio less than 0.3 have been used singly and in combination (along with malformation data) to identify teratogens and non-teratogens. An evaluation of corresponding TI and MCIG/LC_{50} data for each substance tested within each validation study did not reveal a direct correlation between the two indices of teratogenicity and emphasizes the extent of inter-laboratory variability. For the 12 substances evaluated in FETAX Phase III.3 Validation Study, the extent of concordance for the two indices of teratogenic activity without and with metabolic activation was only 57% and 47%, respectively. This lack of concordance adversely impacts on the credibility and usefulness of the data for hazard identification.

In the published FETAX literature, quantitative negative/solvent control data were included only sporadically. In almost all cases, general statements were made that suitable negative control data were obtained but no supporting data were provided. Similarly, quantitative data for 6-AN, the reference substance for studies conducted without metabolic activation, or CP, the concurrent positive control for studies conducted with metabolic activation, were seldom published. It is worth noting that in the FETAX Phase I Validation Study (Bantle et al., 1994a), the investigators concluded that 6-AN may not be suitable as a reference control. A replacement reference control has not yet been designated (ASTM, 1998). The appropriateness of using CP as a concurrent positive control at a concentration that results in 100% mortality should be evaluated. A response of this magnitude limits a statistical consideration of historical data. Also, as the TI is

considered a primary measure of teratogenic potential, it may be more informative if a concentration of CP is used that allows for an assessment of malformations, as well as mortality. The lack of quantitative negative and positive control data eliminates an evaluation of historical control data. To conduct such an evaluation, appropriate historical control data would need to be obtained from multiple laboratories.

Another issue affecting data interpretation is the utility of an exogenous MAS in the FETAX assay. As indicated, the database for substances tested with metabolic activation is very limited. Furthermore, in the validation studies where the same substances are tested without and with metabolic activation, there is no increase in assay performance. Instead, there is an increase in inter-laboratory variability and an associated decrease in concordance. The rationale for the selection of substances tested without and with metabolic activation during the validation process is not clear, as most of them are not known to be activated to teratogens by metabolic activation. The utility of an exogenous MAS and the appropriateness of the MAS ingredients used requires further assessment.

Limitations associated with FETAX in regard to test method reliability included excessive variability in LC_{50}, EC_{50}, TI, and MCIG values, the lack of concordance among laboratories in FETAX study results, the lack of readily available historical negative and positive control data for FETAX, and the limited database for studies with metabolic activation. Other possible limitations include the use of the arithmetic mean in FETAX studies rather than the geometric mean, as is specified by the ASTM FETAX Guideline (1991, 1998); the use of a two-point graphical method for determining the EC_{50} and LC_{50} values; a consensus call in the FETAX validation studies based on averaging data rather than using independent conclusions across multiple participating laboratories; and the use of less than three definitive replicates to define a FETAX response. The effects of these perceived limitations on the performance characteristics of FETAX are not known.