View Table of Contents
HTML Format

Background Review Document

Frog Embryo Teratogenesis Assay – *Xenopus* (FETAX)

March 10, 2000

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

National Institute of Environmental Health Sciences
(NIEHS)
PO Box 12233
Mail Drop: EC-17
Research Triangle Park, NC 27709

TABLE OF CONTENTS

FETAX FOR HUMAN DEVELOPMENTAL HAZARD IDENTIFICATION

FETAX FOR ECOTOXOLOGICAL HAZARD ASSESSMENT USING WATER/SOIL/SEDIMENT SAMPLES

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS AND ACRONYMS

ASTM	American Society for Testing and Materials
BRD	Background Review Document
°C	degrees centigrade
CAA	Clean Air Act
CASRN	Chemical Abstract Service Registry Number
CFR	Code of Federal Regulations
cm	centimeter
CP	cyclophosphamide
CRM	comparative research material
CV	coefficient of variation
DART	Developmental and Reproductive Toxicology
DNA	deoxyribonucleic acid
EC_{50}	effective concentration (i.e., concentration inducing malformation in 50% of exposed embryos)
ECVAM	European Centre for the Validation of Alternative Methods
EGME	ethylene glycol monomethyl ether
EPA	U.S. Environmental Protection Agency
EST	embryonic stem cell
FETAX	Frog Embryo Teratogenesis Assay—*Xenopus*
FIFRA	Federal Insecticide, Fungicide and Rodenticide Act
FR	Federal Register
g	gram
GAP	Genetic Activity Profile
GLP	Good Laboratory Practice
GST	glutathione-S-transferase
h	measure of inter-laboratory variability
HCl	hydrochloric acid
HSDB	Hazardous Substances Data Bank

ICCVAM	Interagency Coordinating Committee for the Validation of Alternative Methods
k	measure of intra-laboratory variability
kg	kilogram
L	liter
LC_{50}	lethal concentration (i.e., concentration inducing death in 50% of exposed embryos)
MAS	metabolic activation system
MCIG	minimum concentration to inhibit growth
MM	micromass
mg	milligrams
mL	milliliter
MMI	mortality/malformation index
mRNA	messenger ribonucleic acid
NADPH	nicotinamide adenine dinucleotide phosphate (reduced form)
NICEATM	NTP Interagency Center for the Evaluation of Alternative Toxicological Methods
NIEHS	National Institute of Environmental Health Sciences
NIOSH	National Institute for Occupational Safety and Health
NLM	National Library of Medicine
NTP	National Toxicology Program
OECD	Organization for Economic Cooperation and Development
PAH	polycyclic aromatic hydrocarbon
RNA	ribonucleic acid
RTECS®	Registry of Toxic Effects of Chemical Substances
SAR	structure-activity relationships
T_3	triiodothyronine
TI	Teratogenic Index
TOC	total organic carbon
TU	Toxic Unit
TERIS	Teratogenic Effects of Drugs. A Resource for Clinicians

TSCA	Toxic Substances Control Act
µg	microgram
VPA	valproic acid
WEC	whole rat embryo culture

EXECUTIVE SUMMARY

General Introduction:

In 1998, the U.S. Environmental Protection Agency (EPA) requested that the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) evaluate the validation status of the Frog Embryo Teratogenesis Assay—*Xenopus* (FETAX). ICCVAM agreed to coordinate a review of the method, and the National Toxicology Program (NTP) Interagency Center for the Validation of Alternative Toxicological Methods (NICEATM) agreed to prepare a Background Review Document (BRD) summarizing the available data and the extent to which each of the ICCVAM validation and acceptance criteria have been met. NICEATM assessed the validation status of FETAX as a screening assay for detecting potential human teratogens, and for its use in the ecotoxicological assessment of water/soil/sediment samples.

Historical Background: Dr. James Dumont introduced FETAX, which uses the embryos of the South African clawed frog (Xenopus laevis), in 1983. Since its introduction, a number of inter-laboratory studies, largely directed by Drs. John Bantle (Oklahoma State University, Stillwater, OK, U.S.) and Douglas Fort (Stover Biometrics Laboratory, Stillwater, OK, U.S.), have been conducted to validate the utility of the assay for developmental hazard assessment. These validation studies were conducted in collaboration with the U.S. Army and the National Institute of Environmental Health Sciences (NIEHS). In 1991, the American Society for Testing and Materials (ASTM) developed a test guideline for FETAX, which was subsequently revised and republished in 1998. With regard to human developmental hazard assessment, this document reviews the information provided by 276 studies involving 137 substances. For ecotoxicological hazard assessment, test data from ten publications involving 124 water/soil/sediment samples were considered.

FETAX for Human Developmental Hazard Identification

Rationale: FETAX is a 96-hour *in vitro* whole-embryo test developed to determine the teratogenic and developmental toxicity potential of chemicals and complex mixtures. The primary endpoints include mortality, malformations, and growth inhibition. Based on mortality and malformation data obtained over a range of dose levels, the 50% lethal concentration (LC_{50}) (i.e., the concentration estimated to induce lethality in 50% of exposed embryos) and the 50% effective concentration for malformations (EC_{50}) (i.e., the concentration estimated to induce malformations in 50% of exposed embryos) are calculated. These two point estimates are used to calculate the teratogenic index (TI), which is equal to the LC_{50} divided by the EC_{50}. Growth is ascertained by measuring the head to tail length of the embryos. The minimum concentration to inhibit growth (MCIG) (i.e., the lowest effective concentration for growth inhibition) is determined by statistically comparing the mean 96-hour head to tail length of the treated embryos at each treatment concentration to that of the control embryos. The statistical comparison is based on using student's t-test for grouped observations at the p=0.05 level. Any one of three criteria (TI, growth inhibition, or severity of induced malformations) is used to identify a teratogen. TI values greater than 1.5 signify a greater potential for embryos to be malformed in the absence of significant embryo mortality. Growth inhibition is stated to be correlated with teratogenicity in FETAX, and teratogenic hazard is considered to be present when growth is significantly inhibited at concentrations below 30% of the 96-hour LC_{50} (i.e., when the $MCIG/LC_{50}$ ratio is less than 0.30). Teratogens generally cause moderate to severe malformations at concentrations near the 96-hour LC_{50}.

Mechanistic Basis: FETAX is essentially an organogenesis test, and organogenesis is highly conserved across amphibians and mammals. The first 96 hours of embryonic development in *Xenopus* parallel many of the major processes of human organogenesis. Thus, FETAX should be useful in predicting potential human developmental toxicants and teratogens. Due to the nature of the endpoints assessed, FETAX does not provide information on substances that may induce functional developmental deficits in mammals. As *Xenopus* embryos are deficient in mixed function oxidase-dependent metabolic activation processes, the addition of an exogenous metabolic activation system (MAS) to the assay allows for an assessment of the developmental toxicity of metabolites in addition to the parent substance.

Regulatory Rationale: Current Federal regulations require determination of the developmental toxicity potential of many chemicals and products. EPA regulations specify the use of at least one, but usually two mammalian species (e.g., rats, mice, rabbits, hamsters) for the testing of fuels and fuel additives, pesticides, and other materials. The Organization for Economic Cooperation and Development (OECD) guidelines do not explicitly restrict developmental toxicity testing to mammals, although the use of FETAX has not been specifically addressed. Because FETAX is relatively easy, rapid, and inexpensive, the test has been proposed as a screening assay to identify potential human teratogens and developmental toxicants. As a screening test, a positive FETAX response would indicate a potential human hazard and, in the absence of other data, would be considered a presumptive teratogen/developmental toxicant. A negative FETAX response would not necessarily indicate the absence of a hazard, and negative responses would be followed by definitive *in vivo* mammalian testing A positive response would require no further testing unless there is concern about a potential false positive response (e.g., the positive FETAX response occurs at doses not applicable to the *in vivo* situation). For public agencies, such information could also be used to prioritize chemicals for more definitive testing. Regardless of the result obtained, an investigator may conclude that confirmatory testing is merited based on consideration of supplemental information, such as structure-activity relationships (SAR) and other chemical and/or testing data.

FETAX is considered to be applicable to all chemicals individually or in formulations, and to commercial products or mixtures that can be measured accurately at the necessary concentrations in water. This assay has not yet been considered for acceptance by U.S. Federal agencies for human health hazard assessment. The most commonly used protocol for identifying a potential human developmental hazard involves the administration of a test substance at three dose levels to pregnant laboratory mammals (usually mice, rats, or rabbits) during the period of major organogenesis. Treatment is followed by evaluation of maternal responses throughout pregnancy, and then examination of the dam and the uterine contents just prior to term. The developmental toxicity endpoints assessed include mortality (e.g., incidence of total, early, and late fetal deaths), malformations (external, visceral, skeletal), variations (external, visceral, skeletal), growth (body weight), clinical signs (type, incidence, duration, and degree), and gross necropsy and histopathology. Mortality, malformations, and growth are endpoints assessed in FETAX.

A successfully validated FETAX could serve as a screening assay within a tiered scheme (e.g., a negative FETAX study would be followed by an *in vivo* mammalian assay, a positive FETAX study would not require further testing) to identify potential human teratogens and developmental toxicants. In this role, the assay has potential benefits with regard to reducing animal use and the cost and time associated with testing for developmental toxicants.

Test Method Protocol: A comprehensive guideline for conducting FETAX was published in 1991 under the auspices of the American Society for Testing and Materials (ASTM), as a "Standard Guide for Conducting the Frog Embryo Teratogenesis Assay—*Xenopus* (FETAX)," Annual Book of ASTM Standards, Designation E1439-91. In 1998, a revised ASTM FETAX Guideline (Designation E 1439-98) was produced. This guideline appears to be adequate to properly guide an investigator through the necessary test components and to ensure consistency in the testing methodology. One aspect of the protocol that may merit further investigation is the decision criterion used to identify a teratogenic response in FETAX. Several approaches have been suggested for improving the performance characteristics of FETAX compared to mammalian teratogenicity. One potentially significant improvement would be to base the EC_{50} on characteristic malformations only, rather than on all malformations detected as is done currently. Characteristic malformations would be those that increase in frequency and possibly severity with increasing concentrations of the test substance.

Characterization of the Materials Tested: FETAX test data from 276 studies involving 137 substances, excluding environmental samples, were located, reviewed, extracted, and entered into the NICEATM FETAX database. The five most numerically prevalent chemical classes, in descending order, were nitrogen heterocyclic compounds (40 substances), amides and hydrazides (29 substances), organic (phenolic and carboxylic) acids (24 substances), alcohols (including glycols) (22 substances), and salts (20 substances). The five major product classes, in descending order, were pharmaceuticals (45 substances), chemical synthesis components (17 substances), pesticides (13 substances), food additives (11 substances), and dyes (7 substances). In a number of cases, the same substance was placed in more than one chemical or product class.

Reference Data Used for Performance Assessment: Laboratory mammal (rat, mouse, and/or rabbit) reference data were located for 90 of the 137 substances and one environmental sample tested in FETAX. Human data (epidemiological and case-report information) were obtained for 31 of the 137 substances tested in FETAX and mammalian data were located for 30 of these. The quality of the data in terms of accuracy and whether the studies were conducted in compliance with national/international Good Laboratory Practice (GLP) guidelines was not determined.

Test Method Data and Results: The 1991 ASTM FETAX Guideline, with minor exceptions, was followed in the FETAX studies considered by NICEATM. All 137 substances in the FETAX database had been tested using without metabolic activation; 35 had also been tested with metabolic activation. Except for the most recent four of the five FETAX validation studies, blind coding was not used in any study to eliminate potential bias. Also, FETAX studies were not conducted in compliance with national or international GLP guidelines. The effect of these two issues on FETAX data quality is difficult to ascertain.

Test Method Performance Assessment: The performance characteristics (i.e., accuracy[1], sensitivity[2], specificity[3], positive predictivity[4], negative predictivity[5], false positive rate[6], and false negative rate[7]) of FETAX against rat, mice, and/or rabbit teratogenicity test results or human teratogenicity study results were determined by NICEATM. The decision criteria used in determining the performance characteristics of FETAX included single decision criteria (TI >1.5; TI >3.0; MCIG/LC_{50} <0.30) and multiple decision criteria (TI >1.5 plus MCIG/LC_{50} <0.30; TI >3.0 plus MCIG/LC_{50} <0.30). When a multiple decision criterion was used, test substances were

[1] Accuracy: The proportion of correct outcomes of a method. Often used interchangeably with concordance.
[2] Sensitivity: The proportion of all positive chemicals that are correctly classified as positive in a test.
[3] Specificity: The proportion of all negative chemicals that are correctly classified as negative in a test.
[4] Positive Predictivity: The proportion of correct positive responses among materials testing positive.
[5] Negative Predictivity: The proportion of correct negative responses among materials testing negative.
[6] False Positive Rate: The proportion of all negative substances that are falsely identified as positive

classified as positive when both the TI value was greater than the decision point (1.5 or 3.0) and the MCIG/LC_{50} ratio was less than 0.3; equivocal when one, but not both, criterion were positive; and negative when neither criterion was positive.

The performance characteristics of FETAX (with and/or without metabolic activation) was determined against all three laboratory mammal species (rat, mouse, and rabbit) combined or against each species alone. Using a single decision criterion, optimal performance for FETAX, with and without metabolic activation, compared against combined laboratory mammal data was based on a TI value greater than 1.5 (**Table A**). Using a multiple decision criterion did not enhance the performance characteristics of FETAX. Similar performance characteristics were obtained against rat, mouse, or rabbit, when considered individually.

Table A. Performance Characteristics of FETAX

Performance Characteristics	FETAX, with and without metabolic activation, compared to Combined Laboratory Mammal (using TI >1.5)	FETAX, with and without metabolic activation, compared to Human (using MCIG/LC_{50} <0.30)	Combined Laboratory Mammal compared to Human
Accuracy	61% (55/90)*	70% (19/27)	63% (19/30)
Sensitivity	82% (41/50)	67% (8/12)	71% (10/14)
Specificity	35% (14/40)	73% (11/15)	56% (9/16)
Positive Predictivity	61% (41/67)	67% (8/12)	59% (10/17)
Negative Predictivity	61% (14/23)	73% (11/15)	69% (9/13)
False Positive Rate	65% (26/40)	27% (4/15)	44% (7/16)
False Negative Rate	18% (9/50)	33% (4/12)	29% (4/14)

*Numbers in parenthesis indicate the number of accurate results/total number of substances compared.

[7] False Negative Rate: The proportion of all positive substances falsely identified as negative.

The performance of FETAX (with and/or without metabolic activation) was compared against human teratogenic data. Again, both single and multiple decision criteria were evaluated. Optimal performance was based on using a single decision criterion of an MCIG/LC_{50} ratio less than 0.30. The resulting performance characteristics are presented in **Table A**. Using a multiple decision criterion did not significantly increase the performance characteristics of FETAX compared to human teratogenicity study results.

Maximal performance characteristics for laboratory mammal data compared to human teratogenicity results were obtained using rat, mouse, or combined laboratory mammal teratogenicity data, but not using rabbit data alone. The analysis was limited to substances tested in FETAX. The combined laboratory rat, mouse, and rabbit results are provided for comparative purposes in **Table A**.

The performance characteristics of FETAX, with and/or without metabolic activation, was determined for chemical and product classes that contained at least 15 substances with corresponding laboratory mammal or human teratogenicity results. Compared to laboratory mammal data, chemical and product classes evaluated included amides (15 comparisons), amides plus hydrazides (19 comparisons), amines (16 comparisons), amines plus nitrogen heterocyclic compounds (25 comparisons), nitrogen heterocyclic compounds (29 comparisons), phenolic and carboxylic acids (21 comparisons), and pharmaceuticals (40 comparisons). Compared to human study data, chemical and product classes evaluated included nitrogen heterocyclic compounds (17 comparisons) and pharmaceuticals (22 comparisons). Regardless of the single decision criterion used, performance characteristics were not appreciable different from those determined for the total database.

NICEATM evaluated the optimal TI value or MCIG/LC_{50} ratio to use as a single decision criterion in FETAX for identifying teratogenic activity. Performance characteristics (accuracy, sensitivity, specificity) were determined against combined laboratory mammal (rat, mouse, and rabbit) or human teratogenicity results. Accuracy based on using either a TI value or an MCIG/LC_{50} ratio as the single decision criterion value was never greater than ~60%, while a sensitivity of at least 85% was accompanied by a specificity of 40% or less. Differences in performance characteristics between this analysis and the previous analysis reflect differences in the manner in which FETAX test results for the same substance from multiple studies were considered. In this analysis, the median TI value or MCIG/LC_{50} ratio were used; in the previous analysis, a weight-of-evidence approach was used to classify results as positive or negative. The values obtained suggest that the use of FETAX as a screen, based on current decision criteria, is problematic.

The inclusion of a MAS in FETAX is considered essential for predicting developmental hazard in mammals. However, selection of the substances tested with a MAS do not appear to have been based on whether or not metabolic activation was thought to be required for teratogenic activity *in vitro*. Of the 35 substances tested with metabolic activation, only four are known to require metabolic activation to be reactive *in vitro*. Based on the limited database, additional studies to validate the role of metabolic activation in FETAX appear to be justified.

Several approaches have been suggested for modifying the decision criteria to increase the ability of FETAX to correctly identify developmental toxicants. These include:

- an evaluation of the EC_{50} based on characteristic malformations (i.e., those increasing in incidence and severity with increasing test substance concentration) only,

- the calculation of a point estimate for the dose that inhibits growth by 50% rather than using an MCIG, and

- the use of 95% confidence intervals for statistically identifying TI values (and other point estimates) that are significantly different from the decision criteria value.

The effect of these approaches on the performance characteristics of FETAX has yet to be evaluated.

<u>Test Method Reliability (Reproducibility/Repeatability):</u> Five separate but related inter-laboratory FETAX validation studies in three phases were conducted. A total of 26 substances

were tested without metabolic activation and 14 substances with metabolic activation, with three to six different laboratories participating in each validation study (**Table B**). The Phase I Validation Study was classified as a training and protocol evaluation phase; the 1991 ASTM FETAX Guideline was followed. The subsequent four validation studies followed the same guideline with minor modifications (e.g., different preparation scheme for adding the test substance; 20 and not 25 embryos per dish when plastic rather than glass Petri dishes were used).

Validation was measured using the four different measurements obtained from FETAX—LC_{50}, EC_{50}, TI, and the MCIG. The investigators assessed reliability of each FETAX endpoint by calculating the coefficient of variation (CV) and conclusions about reliability were made from evaluating the range of CVs for each measure across laboratories (**Table B**). Additionally, the ASTM E691-92 (ASTM, 1992) guideline on a statistical approach for assessing intra- and inter-laboratory performance was used to evaluate test method reliability.

In the validation studies, there was excessive variability in the LC_{50}, EC_{50}, TI, and especially the MCIG within and across laboratories. A formal investigation into the factors contributing to this excessive variability has not been conducted. The resulting variation in these endpoints contributed to poor concordance among laboratories in regard to the classification of a test substance as a FETAX positive or negative, even when highly experienced laboratories were involved (**Table B**). A possible factor contributing to the variation in results may be that the types and severity of malformations are not currently included in the decision criteria used to classify substances as teratogenic or not teratogenic. A subsequent revision of the decision criteria emphasizing critical, or characteristic, malformations has been proposed.

<u>Test Method Data Quality:</u> Studies were conducted using routine laboratory practices, including standard record-keeping procedures. Studies were not conducted in accordance with Good Laboratory Practice (GLP) guidelines, nor were they generally conducted at facilities at which GLP studies are normally conducted. A quality assurance (QA) data audit of the FETAX Phase III.3 Validation Study indicated that data trails, study records, and results analysis procedures were not sufficient to support a standard GLP QA audit. An analysis of the accuracy of the data in the published report revealed the presence of occasional transcriptional errors; however, none

Table B. Summary of FETAX Validation Studies

	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 Tested	3	4	6	2	12
Number of Participating Laboratories	7[a]	7[a]	7[a,b]	7[a]	3
Tested Without MAS	Yes	Yes	Yes	Yes	Yes
Tested With MAS[c]	No	No	No	Yes	Yes
Coded Substances Used	No	Yes	Yes	Yes	Yes
Dose Selection Process	Common Doses	Common Doses	Individual Laboratory Selected	Individual Laboratory Selected	Individual Laboratory Selected
Overall CV mean and range (%), without MAS	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)
Overall CV mean and range (%), with MAS	N/A	N/A	N/A	51.0 (18.0-131.0)	51.1 (2.3-166.6)
Proportion of Study Results in Agreement (TI >1.5)[d]	3 of 3 (100%)	4 of 4 (100%)	1 of 6 (17%)	2 of 4 (50%)	12 of 24 (50%)
Proportion of Study Results in Agreement ($MCIG/LC_{50}$ <0.30)[e]	0 of 3 (0%)	3 of 4 (75%)	0 of 6 (0%)	2 of 4 (50%)	14 of 23 (61%)

MAS = metabolic activation system.

[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.

[c] Aroclor 1254-induced rat liver S9.

[d] Proportion of times that the participating laboratories agreed in classifying a FETAX study result as positive or negative, based on using a TI value greater than 1.5 as the single decision criterion.

[e] Proportion of times that the participating laboratories agreed in classifying a FETAX study result as positive or negative, based on using an $MCIG/LC_{50}$ of less than 0.30 as the single decision criterion.

of the discrepancies were considered to have significantly altered the reported general conclusions.

<u>Other Scientific Reports and Reviews:</u> No independent peer reviews of FETAX were located. Teratogenicity studies with *Xenopus* that did not follow the ASTM FETAX Guideline were located but excluded from consideration.

<u>Animal Welfare Considerations:</u> FETAX is proposed as a screen for human hazard identification (i.e., positive results only preclude the need for additional testing), and thus will not totally eliminate the use of mammals in teratogenicity and developmental toxicity testing. However, if accepted as a screen, use of this *in vitro* assay would reduce reliance on laboratory mammal tests, and thereby reduce the number of mammals used.

<u>Other Considerations:</u> Sufficient information on facilities and equipment for establishing FETAX is provided in the ASTM FETAX Guideline (1991, 1998). The three to six month estimated technical training time required for conducting the in-life portion of a FETAX study appears to be sufficient. However, based on concerns about differences in expertise in the identification of some of the more subtle malformations induced in *Xenopus* embryos, a more extensive training period may be required for the classification of malformations. The projected cost (<$25,000) and study duration (less than two months) for a Good Laboratory Practice (GLP) compliant FETAX study, with and without metabolic activation, appears to be reasonable. In comparison, a complete rat Prenatal Developmental Toxicity Study would cost about $120,000.

The potential impact of tetraploidy on the extrapolation of teratogenic changes in *X. laevis* to laboratory mammals and humans needs to be considered. Furthermore, the possible advantages of a diploid species of *Xenopus*, such as *X. tropicalis*, in FETAX, should be evaluated.

One recent development, which may greatly increase the utility of FETAX for identifying and prioritizing developmental hazards, is cDNA microarray technology. In this approach, developmental toxicity would be monitored at the level of the gene in terms of either up- or down-regulation. Given that exposures to different classes of developmental toxicants would be expected to result in distinct patterns of altered gene expression, microarray technology could be utilized to categorize and classify these effects. In FETAX, treatment with a known developmental toxicant may provide a gene expression "signature" on a microarray, which represents the cellular response to these agents. When an unknown substance is tested, the microarray response could then be evaluated to see if one or more of these standard signatures is elicited. This approach might be used to elucidate an agent's mechanism of action, assess interactions between combinations of agents, or allow for a comparison between altered gene function in *Xenopus* with changes in analogous genes in mammalian systems. Currently, NIEHS is developing a custom "DNA chip" for *Xenopus* that is oriented toward the expression of genes involved in responses to toxic insults.

A number of *in vitro* systems have been proposed as alternatives or screens to *in vivo* mammalian developmental toxicity assays. A European Centre for the Validation of Alternative Methods (ECVAM)-sponsored validation of three *in vitro* assays considered suitable for the detection of substances posing a mammalian developmental hazard is in progress. The relative performance, cost-effectiveness, and flexibility of FETAX against other *in vitro* assays in identifying substances with mammalian developmental toxicity was not evaluated.

FETAX For Ecotoxicological Hazard Assessment Using Water/Soil/Sediment Samples

Rationale: Due to varying susceptibilities among animals, testing in multiple species is considered necessary to protect the environment. For each species, it is a combination of toxicants, water quality, and the organism itself that defines the hazard for a specific concentration of a toxicant within defined water quality conditions. Ecotoxicological standards are generally based on the susceptibility of the adult animal, which may not provide adequate protection for embryonic development and reproduction in many species. It is inherently impossible to evaluate developmental toxicity without exposing animals throughout development and assessing for adverse effects in multiple life stages. and for Early embryonic and juvenile stages are often the most susceptible periods for the toxic effects of many environmental contaminants. Embryonic development in amphibians is sensitive to water quality. Because of this, FETAX has been used in ecotoxicological studies to evaluate the potential developmental hazard of contaminated surface waters, sediments, waste site soils, and industrial wastewater and to evaluate the efficacy of wastewater treatment procedures. In this context, the resulting data can be used to identify and prioritize sites with increased developmental toxicity risks.

<u>Test Method Protocol:</u> The 1991 and the revised and expanded 1998 FETAX Guideline published by ASTM is detailed, comprehensive, and well structured. Known limits of use for FETAX with water/soil/sediment samples were not described, except it was stated that the test method is incompatible with environmental samples that alter the pH, hardness, alkalinity, and conductivity of the FETAX Solution beyond the acceptable range. Testing of solids is generally limited by the water solubility of the constituents. The effects of other physico-chemical properties (e.g., nitrate levels) on *Xenopus* embryonic development need to be evaluated.

<u>Characterization of Water/Soil/Sediment Samples Tested in FETAX:</u> FETAX test data from ten publications involving 124 water/soil/sediment samples were located, reviewed, extracted, and entered into the NICEATM FETAX Environmental Sample Database.

<u>Reference Data Used for an Assessment of FETAX Performance Characteristics:</u> With one exception, laboratory mammal teratogenicity data for water/soil/sediment samples were not available, while relevant data for humans was nonexistent. Appropriate reference data for non-mammalian aquatic species was limited to a direct comparison in one sediment study and two-related soil extract studies between FETAX and *Pimephales promelas* (fathead minnow). Future ecotoxicological studies with FETAX should include tests on at least one reference species.

<u>FETAX Test Method Data and Results:</u> No attempt was made to obtain original data for any ecotoxicological study considered in this BRD. Generally, coded water/soil/samples were used for ease of identification and chain of custody. These studies were not conducted in compliance with GLP guidelines, nor were they generally conducted at facilities at which GLP-compliant studies are normally conducted. All 124 environmental samples in the NICEATM Environmental Sample Database had been tested using FETAX without metabolic activation; no environmental sample was tested also with metabolic activation.

FETAX has been used to evaluate the developmental toxicity of discharges from abandoned lead and zinc mines, contaminated ground and surface water samples collected near a closed municipal landfill, and direct discharges from industries and municipal wastewater treatment plants. This assay has also been used to assess the potential cause(s) of malformations and abnormalities observed in various species of frogs inhabiting bodies of water throughout the United States. FETAX has been used to assess the comparative hazard of soil samples from multiple waste sites contaminated with metals, PAHs, petroleum products, and organochlorine pesticides. The assay has also been used to test a series of five related fossil fuel mixtures as potential environmental pollutants.

Based on the studies evaluated, FETAX appears to be useful in ecotoxicological studies, and as a means for detecting and prioritizing sites with increased developmental hazard. Studies including other bioassays as part of a battery indicated that FETAX was sensitive enough to detect low levels of developmental abnormalities, but robust enough to be suitable for testing aqueous soil extracts. To increase the validity of the interpretation of such data, it may be useful to further evaluate the influence of the physico-chemical properties of environmental samples on the frequency of malformations in FETAX. Additionally, further research on the performance of the current FETAX protocol as an effective assay for assessing water and sediment quality and detecting changes that can have adverse effects on the ecosystem may provide further insight that could optimize ecotoxicological assessments. It would also be helpful to further evaluate how FETAX could best fit into a test battery for prioritizing of sites for further testing and remediation.

Performance Characteristics of FETAX with Water/Soil/Sediment Samples: Given the lack of sufficient reference data for comparison, the performance characteristics of FETAX, based on tests conducted using water/soil/sediment samples, could not be determined. However, there may be ecotoxicological testing applications where reference data for other species may not be needed or appropriate.

Test Method Reliability (Repeatability/Reproducibility): Due to the lack of appropriate inter-laboratory validation studies, an assessment of test method reliability with environmental samples could not be conducted. One potential issue affecting data interpretation connected with water/soil/sediment samples is the lack of an exogenous MAS incorporated into the FETAX assay. An MAS would be useful where results are being used to predict effects on mammalian species. A FETAX validation study designed to evaluate test method reliability for ecotoxicological applications would be helpful. Such a study should include assessments by several laboratories, and should include the testing of both common samples and environmental samples collected independently. Studies focusing on data interpretation issues could also be helpful in further optimizing the assay. Potential issues to address include the decision criteria used for ranking samples in regard to developmental hazard, and the appropriateness of sample handling and processing techniques. ICCVAM Submission Guidelines should be followed in the design, conduct, and reporting of such studies.

Test Method Data Quality: Studies were not conducted in compliance with national or international GLP guidelines, nor were they generally conducted at facilities at which GLP studies are normally conducted. No data audits were conducted on studies testing environmental samples.

Other Scientific Reports and Reviews: No independent peer reviews of FETAX were located. Other data may exist that might be considered in an evaluation of the performance characteristics of FETAX for identifying developmental hazards in environmental samples.

Animal Welfare Considerations: Multiple species are generally used for ecotoxicological studies. Use of this *in vitro* assay could reduce reliance on tests involving adult organisms.

Other Considerations: Sufficient information on facilities and equipment for establishing FETAX is provided in the ASTM FETAX Guideline (1991, 1998). The estimated three to six month technical training time required for conducting the in-life portion of a FETAX study appears to be sufficient. However, based on concerns regarding the level of expertise needed for the proper identification of malformations induced in *Xenopus* embryos, more intensive training may be needed for this aspect of the assay. The projected cost (<$12,500) and study duration (<two

months) for a GLP compliant complete FETAX study, without metabolic activation, following the ASTM FETAX Guideline (1998), appears to be reasonable.

<u>Other Applications for *Xenopus*:</u> Other tests using *Xenopus* are being evaluated for their ability to identify substances or environmental samples that may disrupt endocrine function (the *Xenopus* Tail Resorption Assay, Vitellogenin Assay), for assessing reproductive toxicity, and for exploring limb mal-development, including possible mechanisms of action (*Xenopus* Limb Bud Assay). These developing test methods require appropriate validation.

GENERAL INTRODUCTION

In 1998, the U.S. Environmental Protection Agency (EPA) requested that the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) evaluate the validation status of the Frog Embryo Teratogenesis Assay—*Xenopus* (FETAX) (**Appendix 9**). The EPA stated that this assay, developed to assess developmental toxicity, appeared to meet many of the ICCVAM validation criteria, and that it had been used in human health and water quality assessments. Possible regulatory applications for developmental toxicity identified by EPA included screening and prioritizing compounds for further testing, evaluating complex mixtures and environmental samples, and providing supplemental information in a weight-of-evidence evaluation of human developmental toxicity hazards. Stated advantages of FETAX included:

- a standardized test procedure;
- a published atlas of abnormalities;
- a database of over 100 compounds suggesting an overall accuracy for predicting mammalian teratogens of approximately 90%;
- the availability of mechanistic data indicating similarities between developmental toxicity in FETAX, laboratory mammals, and humans;
- the ability to test chemicals with and without a metabolic activation system (MAS);
- the ability to use the assay either in the laboratory or *in situ*; and
- an ability to evaluate single chemicals or complex mixtures.

In addition, based on multiple validation studies, test developers stated that FETAX appeared to be reproducible within and between laboratories. Stated possible limitations of the assay and areas requiring further discussion included:

- the appropriateness of the calculated TI for identifying negative and positive responses in the assay;

- the influence of the physico-chemical properties of environmental samples or exposures on the frequency of malformations in FETAX; and

- identification of appropriate applications for regulatory purposes and interpretation of data for human health purposes.

ICCVAM agreed to coordinate a review of the method. Subsequently, NICEATM was charged with preparing a BRD summarizing the available data and the extent to which each of the ICCVAM validation and acceptance criteria have been met (**Appendix 15**).

FETAX, which uses the embryos of the South African clawed frog (*Xenopus laevis*), was introduced in 1983 by Dr. James Dumont (Dumont et al., 1983). The assay was developed to evaluate the teratogenic and developmental toxicity potential of chemicals, metals, and complex mixtures (Dumont et al. 1983; Kamimura and Tanimura, 1986; ASTM, 1991; 1998; Sakamoto et al., 1992; Finch, 1994; Bantle, 1995). A number of inter-laboratory validation studies, largely directed by Drs. John Bantle and Douglas Fort, have been conducted to validate the utility of this assay for developmental hazard assessment. In this short-term *in vitro* assay, carefully selected, prepared (dejellied), and staged *X. laevis* embryos are exposed continuously to a test substance for the first 96 hours of embryonic development (ASTM, 1991; 1998). The primary endpoints assessed include mortality, malformations, and growth inhibition (ASTM, 1991; 1998).

The developers of this assay have proposed that data obtained using FETAX may be extrapolated to other species including mammals, may be used to prioritize chemicals and complex mixtures for further tests that use mammals, and may be used in ecotoxicological (e.g., water/soil/sediment) hazard assessment (ASTM, 1991; 1998; Bantle, 1995; Fort et al., 1995; 1996b; Fort et al., 1997). Initial studies conducted using substances with known laboratory mammal and/or human developmental toxicity suggested that the predictive accuracy of FETAX, in the absence of an MAS, exceeded 85% (Sabourin and Faulk, 1987; ASTM, 1991). Furthermore, it has been proposed that inclusion of an MAS should increase the predictive accuracy of the assay for detecting substances with mammalian (including human) developmental toxicity to approximately 95% (ASTM, 1991; 1998).

This BRD presents an evaluation by NICEATM of the utility of FETAX for detecting potential human teratogens, and its use in water/soil/sediment developmental hazard assessment. The structure of the BRD follows the evaluation criteria guidelines found in the *Evaluation of the Validation Status of Toxicological Methods: General Guidelines for Submissions to the Interagency Coordinating Committee on the Validation of Alternative Methods* (**Appendix 15**).

FETAX FOR HUMAN DEVELOPMENTAL HAZARD IDENTIFICATION

1.0 INTRODUCTION AND RATIONALE OF FETAX

1.1 Scientific Basis for the Use of FETAX

FETAX is essentially an organogenesis test, and organogenesis is highly conserved across amphibians and mammals. The first 96 hours of embryonic development in *Xenopus* parallel many of the major processes of human organogenesis (ASTM, 1991; 1998). Thus, FETAX should be useful in predicting potential human developmental toxicants and teratogens (ASTM, 1991; 1998). Because *Xenopus* embryos are deficient in mixed function oxidase-dependent metabolic activation processes, the addition of an exogenous MAS to the assay allows for an assessment of the need for bioactivation for a substance or complex mixture to induce teratogenic activity. The assay developers have stated that the inclusion of an exogenous MAS in FETAX should increase the accuracy of the test method for determining if substances are likely to be human developmental toxicants (Bantle et al., 1989; ASTM, 1991; 1998).

1.2 Intended Uses of FETAX

1.2.1 Intended Regulatory Uses and Rationale

Because FETAX has been concluded by the developers to be easy, rapid, reliable, and inexpensive, the test (with and without metabolic activation) has been proposed as a screening assay for potential human teratogens and developmental toxicants (i.e., for use in hazard identification but not in risk assessment) (ASTM, 1991; 1998). As a screening test, a positive FETAX response would indicate a potential human hazard while a negative FETAX response would not indicate the absence of a hazard. In the role of a screening assay, a negative response would be followed by *in vivo* mammalian testing, while a positive response would require no further testing unless the investigator is concerned about a potential false positive response (i.e., the positive FETAX response occurs at doses not applicable to the *in vivo* situation). However, regardless of the result obtained, an investigator may conclude that confirmatory testing is merited based on consideration of supplemental information, such as SAR and other chemical and/or testing information.

1.2.2 Currently Accepted Teratogenicity/Developmental Toxicity Test Methods

FETAX is not currently accepted by U.S. Federal agencies as a test for identifying teratogenic or developmental toxicants. U.S. Federal and international regulations pertinent to the potential use of FETAX include the following:

- Under the Clean Air Act (CAA), the EPA requires the registration of fuels and fuel additives. As part of the registration process, there are specific toxicity testing requirements. For *in vivo* fertility assessment/teratology testing (40 CFR 79.63), the rat is the preferred species. If other rodent species are used, justification must be provided.

- Under the Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA), teratogenicity and reproduction studies require two-generation testing in two mammalian species (e.g., rat, mouse, rabbit, hamster) for pesticides registered for use on food crops (40 CFR 158.202, 40 CFR 158.340).

- Currently accepted EPA test methods for inhalation developmental toxicity studies require the use of at least two mammalian species (e.g., rat, mouse, rabbit, hamster). If other mammalian species are used, justifications/reasoning for their selection shall be provided (40 CFR 798.4350). Similarly, EPA guidelines regarding test methods for reproduction and fertility toxicants make use of the rat, though in some cases and with justification, other mammalian species can be used (40 CFR 798.4700; 40 CFR 798.4900). For this purpose, pregnant females are exposed to the test agent during most of organogenesis. Shortly before delivery, the pregnant females are sacrificed, the uteri removed, and the contents examined for signs of developmental toxicity. The fetal remains are observed for soft tissue and skeletal defects as well as for resorption.

- Toxic Substances Control Act (TSCA) testing requirements for reproduction and fertility effects call for the use of rats, although other mammalian species are acceptable with justification (40 CFR 799.9380). TSCA prenatal developmental toxicity testing requirements suggest the use of the "most relevant" species, with the rat being the preferred rodent species and the rabbit the preferred non-rodent species (40 CFR 799.9370). EPA provisional guidelines for developmental neurotoxicity recommend the use of rats (40 CFR 795.250).

- Under the Federal Hazardous Substances Act (FHSA), the Consumer Products Safety Commission (CPSC) will evaluate all available evidence from animal and human studies in order to determine whether classification based on developmental toxicity is warranted (16 CFR 150.135). No specific testing is required under FHSA.

- Specific guidelines for evaluation of developmental toxicity under the Federal Food, Drug, and Cosmetic Act (FFDCA) may vary depending on the Center but typically require testing of rats and/or rabbits. The Center for Food Safety and Applied Nutrition identifies testing for reproductive and developmental toxicity under "Toxicological Principles for the Safety Assessment of Direct Food Additives and Color Additives Used in Food," Rev. 1993 (Redbook II). The Center for Drug Evaluation and the Center for Biologics Evaluation and Research reference to International Conference on Harmonization (ICH) Harmonized Tripartite Guidelines, which again indicate the use of rats and/or rabbits (FR 59 (140): 48749).

- Organization for Economic Cooperation and Development (OECD) guidelines do not explicitly restrict developmental toxicity testing to mammals, although the use of FETAX has not been addressed (OECD 414; OECD 415; OECD 416; OECD 421; OECD 422).

A copy of the current EPA guideline for developmental toxicity risk assessment is provided in **Appendix 13**. The four major manifestations of developmental toxicity are death, structural abnormality, altered growth, and functional deficit. Only the first three are traditionally measured in laboratory animals using the conventional developmental toxicity (also called teratogenicity or Segment II) testing protocol as well as in other study protocols (e.g., multigenerational). As described in this document, the most commonly used protocol for assessing developmental toxicity in laboratory mammals involves the administration of a test substance at three dose levels to pregnant animals (usually mice, rats, or rabbits) during the period of major organogenesis. Treatment is followed by evaluation of maternal responses throughout pregnancy, and then examination of the dam and the uterine contents just prior to term. The high dose is selected to produce some minimal maternal or adult toxicity (i.e., a level that at the least produces marginal but significantly reduced body weight, reduced weight gain, or specific organ toxicity, and at the most produces no more than 10% mortality). The low dose is generally a no observable effect level for adult and offspring effects. The route of exposure in these studies is usually oral, unless the chemical or physical characteristics of the test substance or pattern of human exposure suggest a more appropriate route of administration. The developmental toxicity endpoints assessed include mortality (e.g., incidence of total, early, and late fetal deaths/litter), malformations (external, visceral, skeletal), variations (external, visceral, skeletal), growth (body weight), clinical signs (type, incidence, duration, and degree), and gross necropsy and histopathology. Many of the endpoints evaluated in FETAX are qualitatively similar to these endpoints. However, no information on functional deficits, which is required in certain regulatory situations (U.S. EPA, 1991), is provided by the current FETAX protocol.

In terms of these regulations and guidelines, a successfully validated FETAX could serve as a screening assay within a tiered scheme to identify potential human teratogens and developmental toxicants (ASTM, 1991; 1998). The use of FETAX as a screening assay in this scheme is described in **Section 1.2.1**.

1.2.3 The Use of FETAX to Assess Potential Human Teratogenic Hazards

Based on initial studies, the accuracy of FETAX without metabolic activation for predicting human developmental toxicants was concluded to be greater than 85% (Courchesne, 1985; Sabourin, 1987; ASTM, 1991; 1998; Finch, 1994). Furthermore, it has been proposed that the incorporation of metabolic activation into the assay should increase this accuracy to at least 95%

(ASTM, 1991; 1998). However, analyses conducted by NICEATM of the current FETAX database did not verify these accuracy values (see **Section 6.0**).

1.2.4 Intended Range of Chemicals Amenable to Test and Limits According to Physico-Chemical Factors

In the ASTM FETAX Guideline (1991, 1998), FETAX is considered to be applicable to all chemicals individually or in formulations, and to commercial products or mixtures that can be measured accurately at the necessary concentrations in water. With appropriate modifications, FETAX can be used to conduct tests on aqueous effluents; surface and ground waters; leachates; aqueous extracts of water-insoluble materials; and solid-phase samples, such as soils and sediments, particulate matter, sediment, and whole bulk soils and sediments. The preferred solvent is FETAX Solution, which a prepared water-based solution with a standard pH, alkalinity, and hardness suitable for the growth and survival of *Xenopus* embryos (ASTM, 1991; 1998). If a solvent other than FETAX Solution is used, it must be compatible with *Xenopus* embryonic growth and survival. Testing of water-insoluble materials would be limited by the highest concentration that can be achieved using an appropriate organic solvent. The test method is incompatible with materials (or concentrations of materials) that alter the pH, hardness, alkalinity, and conductivity of the FETAX Solution beyond the acceptable ranges indicated in the ASTM FETAX Guidelines (1991, 1998).

1.3 Section 1 Conclusions

The scientific basis for FETAX and its intended use(s) as a screening assay for the identification of potential human teratogens are adequately described. Test limits are defined, but only limited information is available on the complete range of materials amenable to test (see **Section 4**).

2.0 FETAX TEST METHOD PROTOCOL

2.1 Standard Detailed Protocol

Under the auspices of the ASTM, a comprehensive guideline for FETAX was published in 1991. The guideline was subsequently revised, and the updated version was published in 1998. The two versions of this guideline are designated as a "Standard Guide for Conducting the Frog Embryo Teratogenesis Assay—*Xenopus* (FETAX)," Annual Book of ASTM Standards, Designation E1439–91 and E1439–98, respectively. The most recent guideline expands the information procedures outlined in the 1991 ASTM Guideline. The two versions are provided in **Appendix 10** and **11**, respectively. The ASTM FETAX Guideline includes information on the following topics:

- terminology,
- summary of the guideline,
- significance and use of the assay,
- safety precautions,
- apparatus,
- water for culturing *Xenopus* adults,
- preparation of FETAX solution water,
- test material,
- test organisms,
- procedure,
- analytical methodology,
- test acceptability,
- documentation,
- key words, and
- references.

The appendices to the ASTM FETAX Guideline include a list of alternative species, additional endpoints, and alternative exposure scenarios. The 1998 ASTM FETAX Guideline also includes appendices on concentration steps for range-finding tests, microsome isolation reagents, and nicotinamide adenine dinucleotide phosphate (NADPH)-generating system components. The procedures presented in the ASTM FETAX Guideline (1991, 1998) are considered to be applicable to all chemicals individually or in formulations, commercial products, or mixtures. In addition, the 1998 ASTM FETAX Guideline allows, with appropriate modification, the use of FETAX for conducting tests on surface and ground waters, solid phase samples such as soils and sediments, and whole bulk soils and sediments.

A brief description of the ASTM FETAX Guideline (1991, 1998) follows.

2.1.1 Materials, Equipment, and Supplies

Adults should be kept in an animal room isolated from extraneous light that might interfere with a consistent 12-hour photoperiod. Adults can be maintained in large aquaria or in fiberglass or stainless steel raceways at densities of four to six animals per 1800 cm^2 of water surface area. The sides of the tanks should be opaque and at least 30 cm high. The water depth should be between 7 and 14 cm. Water temperature for adults should be 23 ± 3°C. Two types of breeding aquaria are described in detail (ASTM, 1991; 1998). For conducting FETAX, a constant temperature room or a suitable incubator for embryos is required, although a fixed photoperiod is unnecessary. The incubator must be capable of maintaining a temperature of 24 ± 2°C. Covered 60-mm glass Petri dishes should be used as test chambers, except that disposable 55-mm polystyrene Petri dishes should be used if a substantial amount of the test substance binds to glass, but not to polystyrene, or when metabolic activation is incorporated.

Equipment and facilities that contact stock solutions, test solutions, or water in which embryos will be placed should not contain substances that can be leached or dissolved by aqueous solutions in amounts that would adversely affect embryonic growth or development. Additionally, items that contact stock solutions or test solutions should be chosen to minimize sorption of most test materials from water. Glass, Type 316 stainless steel, nylon, and fluorocarbon plastic should be used whenever possible to minimize dissolution, leaching, and sorption. Rigid plastics may be used for holding, acclimation, and in the water supply system, but they should be soaked for a week before use.

FETAX Solution, stock solutions, or test solutions should not contact brass, copper, lead, galvanized metal, or natural rubber before or during the test. Items made of neoprene rubber or other materials not mentioned above should not be used unless it has been shown that their use will not adversely affect either survival or growth of the embryos and larvae of the test species.

A binocular dissection microscope capable of magnifications up to 30x is required to count and evaluate embryos for malformations. A simple darkroom enlarger is used to enlarge embryo images two to three times for head to tail length measurements. It is also possible to measure embryo length through the use of a map measurer or an ocular micrometer. However, the process is greatly facilitated by using a digitizer interfaced to a microcomputer. The microcomputer is also used in data analysis.

Before FETAX is conducted in new test facilities, it is recommended that a "non-toxicant" test be conducted, in which all test chambers contain FETAX solution with no added test material. The embryos should grow, develop, and survive in numbers consistent with an acceptable test. The magnitude of the chamber-to-chamber variation should be evaluated.

2.1.2 Detailed Procedures for FETAX

As recommended in the ASTM FETAX Guideline (1991, 1998), the following information should be known about the test material before a test is conducted:

- identities and concentrations of major ingredients and major impurities,
- solubility and stability in water,
- estimate of toxicity to humans, and
- recommended safe-handling procedures.

An acceptable clutch of eggs has the capability of developing into Developmental Stage 46 tadpoles with less than 10% gross abnormalities and less than 10% mortality. In practice, 95% normal, live embryos should be obtained routinely. Recognition of high quality eggs is based on the following (J. Bantle, personal communication):

- The eggs must be normally pigmented on the top surface;
- The pigment must be even in coloration and not mottled;
- They cannot have been laid in strings (see Bantle et al., 1998);
- Less than 30% of the eggs should exhibit abnormal pigmentation when first laid;
- Greater than 70% should rotate such that the animal (dark) pole is facing up in the dish;
- Fertilization and normal cleavage rates must be in excess of 70%;

The process of dejellying with 2% cysteine is critical if the developing embryos are not to be damaged by the treatment. Damaged embryos often look normal but soon undergo abnormal cleavage. Treatment with cysteine must only progress until the embryos roll with just a slight amount of stickiness. A method of quantifying this step has not yet been developed. Additionally, excess treatment does not also show immediately as a change in morphology. To learn this process, embryos must be dejellied, the normal-looking ones chosen, and then allowed to grow to Developmental Stage 46 to see if mortality and malformation rates are acceptable. Eggs from the same batch may be subjected to different lengths of dejelly time in order to assess the effects of time on the process.

Test substance exposure is continuous throughout the test. For each dose group, two dishes each containing 25 embryos and 10 mL of test solution are used. For the control group, four dishes of 25 embryos each are used. However, studies that employ 55-mm polystyrene Petri dishes rather than 60-mm glass Petri dishes use 20 embryos per dish (Bantle et al., 1998). Both versions of the ASTM FETAX Guideline (1991, 1998) state that embryos must be randomly assigned to test dishes, but the 1998 version includes a revision to make an exception to random assignment when a forced air incubator is used to eliminate the occurrence of hot or cold locations. A temperature of 24 ± 2°C must be maintained throughout the 96-hour test duration. Temperatures higher than 26°C cause malformation, whereas low temperatures prevent the controls from reaching Developmental Stage 46 within 96 hours. If 90% of the embryos in the control dishes have not reached Developmental Stage 46 by 96 hours, the test may be extended by three hours. Deviations from this standard exposure time must be reported as deviating from standard FETAX conditions. The pH of the stock and test solutions must be between 6.5 and 9.0, with 7.7 considered optimal. The pH of a control dish and the pH of the highest test concentration should be measured at the beginning of the test, and subsequently at 24-hour intervals.

Since early *Xenopus* embryos have limited ability to metabolize xenobiotics, particularly in regard to cytochrome P-450 activity, the incorporation of metabolic activation into the standard protocol is necessary when FETAX is used to evaluate developmental toxicity/teratogenicity for human health hazard assessment. The MAS is composed of rat liver microsomes and a NADPH-generating system. The rat liver microsomes may be obtained from an Aroclor 1254-treated male rat. Aroclor 1254 is a broad-spectrum cytochrome P-450-inducing agent and liver microsomes from such rats are appropriate in the majority of experimental studies. Rats exposed to isoniazid or uninduced microsomes may be used in those cases where Aroclor 1254-induction is known to repress specific P-450 isozymes. The nature of the test material may suggest the most appropriate inducing system to use. In cases where limited data are available concerning test substance biotransformation, the ASTM FETAX Guideline (1991, 1998) proposes that a set of Aroclor 1254- and isoniazid-induced rat liver microsomes mixed in equivalent activity ratios be used. However, D. Fort (personal communication) has concluded that a mixture of β-napthoflavone- (or 3-methylcholanthrene-), phenobarbital-, and isoniazid-induced microsomes is the most effective source for an MAS. The P-450 activities of each lot of prepared microsomes will vary. Therefore, the P-450 activity of each lot must be determined and a standard amount added to each dish. It is important to include an MAS-only (microsomes and generator system without test material) negative control. Microsomal protein can slow growth and development at concentrations greater than 60 µg/mL. Reduced nicotinamide adenine dinucleotide, which is required for microsomal activity, can also cause abnormal development and its concentration must be kept low. Additional research may be needed to establish the most appropriate criteria for using the different MAS proposed and the optimal conditions for each. The use of an exogenous MAS in FETAX may not result in the same effects that would be expected to occur if *Xenopus* embryos were P-450 metabolically competent.

Following range-finding tests to identify the appropriate doses to test (see **Section 2.1.3**), three replicate definitive tests are performed. Each of the three definitive tests is conducted using embryos from a different male/female pair of *X. laevis*. If FETAX is being used for human health developmental hazard assessment, definitive tests should be conducted with and without metabolic activation. At a minimum, five concentrations for each endpoint are used. However, additional concentrations between the EC16 and EC84 are highly recommended to ensure obtaining accurate 96-hour LC_{50} and EC_{50} values. The same test material concentrations must be used for each replicate definitive test. The experiments, with and without metabolic activation, should yield acceptable 96-hour LC_{50} and EC_{50} values. If they do not, the tests should be repeated. Prior testing suggests that intra-test variability should yield a coefficient of variation that is less than 100%. In some cases where test variability is extremely high, it may be necessary to determine whether the test material is rapidly degrading, salting out, or volatilizing out of solution.

As defined by the ASTM FETAX Guideline (1991, 1998), a FETAX study should be considered unacceptable if one or more of the following occurs:

- Embryos from more than one mating pair were used in the same test or in replicate tests;
- Hardware cloth or metal mesh was used as a support in the breeding aquarium;
- In the negative controls, either the mean survival is less than 90% or the mean malformation in embryos is greater than 10%, or both;
- Ninety percent of the FETAX-solution-only controls do not reach Developmental Stage 46 by the end of 96 to 99 hours;

- Dilution water was used in the test, and it did not allow embryonic growth at the same rate as FETAX solution;

- The deionized or distilled water does not conform to the Type I ASTM standard;

- A water, FETAX Solution, an MAS control (where an MAS is used), or solvent control was not included in the test;

- The concentration of solvent was not the same in all treatments, except for a dilution-water or FETAX-solution control;

- Identification of the Developmental Stage of the embryos was performed using a reference other than Nieuwkoop and Faber (1975);

- The test was started either with less than Developmental Stage 8 blastulae or with greater than Developmental Stage 11 gastrulae;

- All Petri dishes (or other containers) were not physically identical throughout the test.

- Petri dishes were not randomly assigned to their positions in a non-forced air incubator.

- The embryos were not randomly assigned to the Petri dishes;

- Required data concerning mortality, malformation, and growth were not collected;

- The pH of the test solution was less than 6.5 or greater than 9.0 in the control or highest test concentration;

- Dead embryos were not removed after each 24-hour (±2 hour) interval;

- There was consistent deviation from the temperature limits (a short-term deviation of more than ±2°C might be inconsequential); or

- The reference toxicant produced significant variability (± 2 standard deviation units from the historical mean values) compared to historical data plotted on a control chart.

2.1.3 Dose-Selection Procedures—Range-Finding Test

The ASTM FETAX Guideline (1991, 1998) describes the range-finding tests and the concentration selection procedure. Range-finding tests should be used whenever possible to identify the best approximation of the 96-hour LC_{50} and EC_{50} for definitive testing. Concentration selection is a multistep process that depends on the nature of the test material and the results of the first range-finding test. The first range-finding test consists of a series of at least seven concentrations that differ by a factor of ten. If FETAX is being used for human developmental hazard assessment, range-finding tests should be conducted with and without metabolic activation.

A second range-finding test series is performed using a sliding scale of concentrations provided in the ASTM FETAX Guideline (1991, 1998). The concentration values range from 0.001 to 100; in steps of 0.0005 between 0.001 and 0.1, in steps of 0.05 between 0.1 and 1, in steps of 0.5 between 1 and 10, and in steps of 5 between 10 and 100. Using the sliding scale, the value closest to the 96-hour LC_{50} (for tests conducted with and without metabolic activation) should be identified and then three values immediately below and three values immediately above the estimated LC_{50} should be chosen. The same method should be used to estimate concentrations surrounding the 96-hour EC_{50}. In addition, the 96-hour LC_5, LC_{16}, LC_{84}, and LC_{95} and the EC_5, EC_{16}, EC_{84}, and EC_{95} may be calculated. By determining these values, the concentrations to be tested in the definitive tests are established and the slopes of the concentration-response curves are taken into consideration. Growth inhibition data are not collected from range-finding tests. For some test materials, it may be necessary to use the results of the first definitive experiment as another range-finding test and to adjust the test concentrations accordingly.

2.1.4 Endpoints Measured

The three endpoints measured are mortality, malformations, and embryonic growth.

Mortality: Dead embryos must be removed when solutions are changed at the end of each 24-hour period during the 96-hour test. If dead embryos are not removed, microbial growth can occur that might kill live embryos. Death at 24 hours (Developmental Stage 27) is ascertained by the extent of skin pigmentation, structural integrity, and irritability of the embryo. At 48 hours (Developmental Stage 35), 72 hours (Developmental Stage 42), and 96 hours (Developmental Stage 46), the lack of a heartbeat is an unambiguous sign of death. Based on the mortality data obtained over a range of dose levels, the LC_{50} value is calculated (ASTM, 1991; 1998).

Malformations: Malformations must be recorded at the end of the 96-hour treatment period. The Atlas of Abnormalities (Bantle et al., 1998) should be used in scoring malformations. The number of malformations in each category should be reported in standard format for ease of comparison. Based on malformation data obtained over a range of dose levels, the EC_{50} value is calculated (ASTM, 1991; 1998).

Generally, the two point estimates for mortality and malformations are then used to calculate a TI, which is equal to the LC_{50} divided by the EC_{50} (Bantle et al., 1989; ASTM, 1991; 1998).

Embryonic Growth: The ability of a material to inhibit embryonic growth is often the most sensitive indicator of developmental toxicity (ASTM, 1991; 1998). Head to tail length data (growth) must be collected at the end of each test. If the embryo is curved or kinked, then the measurement follows the contour of the embryo. Measurement should be made after the embryos are fixed in 3% formalin. Using length data, the MCIG is determined by statistically comparing the mean head to tail length of the treated embryos at each dose group to that of the embryos in the control group (ASTM, 1991; 1998).

2.1.5 Duration of Exposure

The ASTM FETAX Guideline (1991, 1998) specifies that *X. laevis* embryos are exposed for 96 hours to the test material. However, if 90% of the embryos in the control dishes have not reached Developmental Stage 46 by this time, the test may be extended by three hours to attain this developmental stage.

2.1.6 Known Limits of Use

As presented in **Section 1.2.4**, FETAX is considered to be applicable to most chemicals and mixtures. Testing of water-insoluble materials would be limited by the highest concentration that can be achieved using an appropriate organic solvent. The test method is incompatible with substances (or concentrations of substances) that alter the pH, hardness, alkalinity, and conductivity of the FETAX solution beyond the acceptable range indicated by the ASTM FETAX Guideline (1991, 1998).

2.1.7 Nature of the Responses Assessed

In FETAX, the primary endpoints assessed are mortality, malformations, and growth inhibition (ASTM, 1991; 1998; Finch, 1994) (see **Section 2.1.4**). Mortality is an easily observable endpoint. Growth inhibition, as measured by a significant decrease in the head to tail length, is also easily measured. Malformations in *Xenopus* can be difficult to identify (see **Section 6.6.2**)

2.1.8 Appropriate Vehicle, Negative, and Positive Controls

As specified by the most recent ASTM FETAX Guideline (1998), a stock solution should be prepared anytime the test substance can not be directly added to the test vessel. Test substances administered using a stock solution should be prepared in such a manner as to ensure that the embryos are exposed to a homogeneous mixture. The concentration and stability of the test substance in a stock solution should be determined before testing. Stock solutions should be prepared daily unless analytical data indicate the solution is stable with time. If the test material is subject to photolysis, the stock solution should be shielded from light.

The preferred solvent for this assay is FETAX Solution; ingredients are provided in the ASTM FETAX Guideline (1991, 1998). The minimum necessary amount of a strong acid or base may be used in the preparation of an aqueous stock solution, but this might appreciably affect the pH of test solutions. Use of a more soluble form of the test material, such as chloride or sulfate salts of organic amines, sodium or potassium salts of phenols or organic acids, and chloride or nitrate salts of metals, might affect the pH more than the use of a minimum necessary amount of a strong acid or base. Prior to testing, all available chemical and physical data on the test substance should be obtained and considered prior to making decisions on pH adjustments.

If a solvent other than FETAX Solution is used, its concentration in test solutions 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 chemicals, 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. Ethanol is not recommended because of its potential teratogenicity. Methanol has high toxicity in FETAX. Acetone might stimulate the growth of microorganisms and is quite volatile. Organic solvents should be reagent-grade six or better. A surfactant should not be used in the preparation of a stock solution because it might affect the form and toxicity of the test material in the test solutions.

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 concentration of solvent must be the same in the solvent control and in all test solutions.

For studies conducted without metabolic activation, 6-aminonicotinamide (6-AN; purity > 99%) is proposed in the ASTM FETAX Guideline (1991; 1998) as the positive or reference toxicant, as this substance presents a mortality and malformation database convenient for reference purposes. The 1998 ASTM FETAX Guideline provides reference values for 6-AN for the 96-hour LC_{50} (2.23 mg/mL) and the 96-hour EC_{50} (0.005 mg/mL), which yield a TI of 446. The MCIG should be ~1.15 mg/mL. However, based on the excessive variability in results obtained for 6-AN in the Phase I Validation Study, the investigators concluded that a replacement reference substance should be identified (Bantle et al., 1994a). A replacement for 6-AN has not yet been identified. A concurrent positive control for studies conducted without metabolic activation is not recommended in the ASTM FETAX Guideline (1991, 1998). Rather, at least quarterly, concentration-response experiments must be performed and the results of these tests compared with historical tests to judge the laboratory quality of FETAX data. Only those biological responses related to mortality and malformations are considered in this analysis; growth inhibition need not be evaluated for 6-AN. NICEATM suggests that the appropriateness of a reference positive control, as opposed to a concurrent positive control, for FETAX studies conducted without metabolic activation should be critically evaluated.

The recommended concurrent bioactivation positive control for studies conducted with metabolic activation is cyclophosphamide (CP) at a concentration of 4 mg/mL. The MAS-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 ASTM FETAX Guideline (1991, 1998) states that a control is needed also to demonstrate that the cytochrome P-450 system is responsible for the observed bioactivation. For this control, a small amount of dithionite may be added directly to the microsomes followed by bubbling carbon monoxide through the microsomal protein at a steady rate for three minutes to inactivate the cytochrome P-450.

NICEATM concluded that the appropriateness of using CP at a concentration that results in 100% mortality should be critically 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.

2.1.9 Acceptable Range of Negative and Positive Control Responses

For negative or solvent controls, the percentage of malformed embryos must not exceed 10%, while mean survival must be greater than 90% (ASTM, 1998).

For 6-AN and CP, the two positive control chemicals recommended in the ASTM FETAX Guideline (1991, 1998), no specific acceptable range of values was provided. However, the ASTM FETAX Guideline (1991, 1998) states that the reference toxicant test must produce data within two standard deviations of the historical mean values. No information is provided in the ASTM FETAX Guideline (1991; 1998) on the number of studies required to generate appropriate historical data or the time period over which such data should be retrospectively assessed. When conducting studies with metabolic activation, the MAS-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. No other information is provided in the ASTM FETAX Guideline (1991, 1998).

2.1.10 Data Collection

As described in the ASTM FETAX Guideline (1991, 1998), data are collected on the incidence of embryos that have died during the 96-hour culture period; the head to tail length, a measure of growth, among the surviving embryos at the end of the 96-hour culture period (see **Section 2.1.4**), and on the incidence and type of malformations present among the surviving embryos at the end of the 96-hour culture period. Malformations are scored using a binocular dissection microscope capable of magnifications up to 30x. The standard FETAX scoring form (ASTM, 1991; 1998) includes the following categories to be scored during an assessment of malformations:

- severe,
- stunted,
- gut,
- edema (multiple, cardiac, abdominal, facial, cephalic, optic),
- axial malformations (tail, notochord, fin),
- face,
- eye,
- brain,

- hemorrhage,
- cardiac,
- blisters,
- other (specify)

2.1.11 Data Storage Media

Original data are collected on FETAX-specific forms and maintained in study books. Example forms are provided in the ASTM FETAX Guideline (1991, 1998). Data are then generally entered into computerized spreadsheets for manipulation and analysis.

2.1.12 Measures of Variability

In FETAX, as described in the ASTM FETAX Guideline (1991, 1998), each test substance concentration involves the use of two replicate dishes, while each control treatment group involves the use of four replicate dishes. Each plastic or glass Petri dish contains 20 or 25 embryos, respectively. Sterile plastic Petri dishes are used with MAS to reduce the possibility of bacterial contamination. To evaluate the teratogenicity of a test material, three replicate definitive tests are performed. Each of the three definitive tests is conducted using embryos from a different male/female pair of *X. laevis*. If FETAX is being used for human health hazard assessment, definitive tests are conducted with and without metabolic activation. The ASTM FETAX Guideline (1991, 1998) specifies that the geometric mean for the 96-hour LC_{50}, the 96-hour $EC_{50,}$ the TI, and the MCIG, as well as their 95% confidence limits be calculated using the data from the three replicate definitive tests and provided in the study report. **Section 2.1.13** describes the statistical methods used to calculate the 96-hour LC_{50}, the 96-hour $EC_{50,}$ and the MCIG.

Intra- and inter-laboratory variation in FETAX has been evaluated using the four different measurements—LC_{50}, EC_{50}, TI, and the MCIG—that can be obtained from each experiment. In some studies, the types of malformations present in the embryos were considered also. Reproducibility and reliability of each FETAX endpoint were evaluated by calculating coefficients of variation (CV [%]), and comparing the CVs for each measure across laboratories. Additionally, a statistical approach described in ASTM E691—92 (ASTM, 1992) (**Appendix 12**), a guide for evaluating inter- and intra-laboratory variability, was used.

Historical negative and positive control data can be used to evaluate variability in performance within a laboratory across time. The ASTM FETAX Guideline (1991, 1998) states that, at least quarterly, concentration-response experiments must be performed for the positive control (without metabolic activation) 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.

2.1.13 Statistical and Non-Statistical Methods

As described in the ASTM FETAX Guideline (1991, 1998), if the test contains a dilution-water or a FETAX-Solution control and a solvent control, the mortality, malformation, and growth inhibition of these treatment groups should be compared using a two-tailed student’s t-test. If a statistically significant difference in mortality, malformation, or growth inhibition is detected between the two controls, only the solvent control may be used as the basis for comparison in the calculation of results.

For the range-finding and definitive tests, probit analysis, trimmed Spearman-Karber analysis, or the two-point graphical method are used to estimate the LC_{50} and EC_{50} values. The graphical method is used only when regular statistical analyses fail to generate useful data. Generally, probit analysis is used when the data meet normal distribution and homogeneity of variance assumptions, and the trimmed Spearman-Karber test is used when the data fail to meet these assumptions. However, range-finding tests may bypass the homogeneity of variance requirements. Data sets that are marginal in terms of concentration-response information should not be analyzed by probit analysis as it may skew the data (D. Fort, personal communication). Spearman-Karber should be used when in doubt or to confirm the results of probit analysis.

The TI, the ratio of the LC_{50} to the EC_{50}, is calculated for each test, and then the mean of the three tests determined. The MCIG is determined using a student’s t-test for grouped observations, with significance at the p = 0.05 level.

The decision criteria described in the ASTM FETAX Guideline (1991, 1998) are based on non-statistical methods (see **Section 2.1.14**).

2.1.14 Decision Criteria

The decision criteria for FETAX are described in the ASTM FETAX Guideline (1991, 1998). The assay provides concentration-response data for mortality, malformations, and growth inhibition. These data can be compared with similar data on a molar basis using other pure test materials or using standard amounts of environmental samples to yield a relative ranking of toxicity. A test substance is considered to be a developmental toxicant when it causes any deficit in an embryo, especially at concentrations lower than those required to induce adult toxicity. In comparison, a teratogen causes some observable abnormality in embryonic development. Three separate FETAX decision criteria (i.e., TI, growth inhibition, and severity of malformations) are used to identify teratogens. Any single decision criterion is considered sufficient to identify a potential teratogenic hazard (ASTM, 1991; 1998).

The TI presents a relative ranking of hazard from nearly 1 to several thousand. The hazard becomes a concern when the mean TI value of three definitive tests is greater than 1.5 (ASTM, 1991; 1998). More recently, Fort et al. (2000a) used a decision criterion in which TI values greater than 1.5 indicate increasing teratogenic hazard, while TI values greater than 3.0 indicate concern. The mortality and malformation concentration-response curves should have similar slopes with acceptable confidence limits when compared to data from 6-AN reference experiments. The term "acceptable" is not defined in the guideline. The TI values of different test substances can be compared to generate relative potency rankings.

In terms of malformations, non-teratogens cause slight to moderate malformations at concentrations near the 96-hour LC_{50}. Teratogens generally cause moderate to severe malformations at these concentrations. Comparison can be made to the reference control 6-AN to identify what constitutes a severe malformation (ASTM, 1991; 1998). An Atlas of Abnormalities (Bantle et al., 1998) is available for judging the severity of malformation.

Growth inhibition is correlated with teratogenesis in FETAX. Teratogenic hazard becomes apparent when growth is significantly affected at concentrations below 30% of the 96-hour LC_{50}. When using this decision criterion, it is important to ensure that the test concentrations selected are adequate to define the MCIG.

Although the ASTM FETAX Guideline (1991, 1998) states that any single decision criterion is considered sufficient to identify a potential teratogenic hazard, Bantle et al. (1999) and Fort et al. (2000a) have also evaluated FETAX results based on multiple decision criteria. In the most recent multi-laboratory validation study (Bantle et al., 1999), each test chemical was judged to pose a developmental hazard when the TI and the MCIG/LC_{50} ratio both indicated hazard (i.e., the TI was value greater than 1.5 and the MCIG/LC_{50} ratio was less than 0.30), and definitely no hazard when both decision criteria fell into the non-hazard category (i.e., the TI value was less than or equal to 1.5 and the MCIG/LC_{50} ratio was greater than or equal to 0.30) (see **Section 7.2.5**). The teratogenic hazard was considered equivocal when one, but not both, of the two decision criteria were positive. In such cases, the types and severity of malformations were examined for guidance in assessing teratogenic hazard. However, due to the subjectivity of malformation identification, this approach was not made a permanent part of the decision criteria (Bantle et al., 1999). In the comparative FETAX - rat teratogenic study conducted by Fort et al. (2000a), a similar multiple criteria approach was used except that the TI decision criterion was based on a TI value greater than 3.0.

2.1.15 Test Report Information

As stated in the ASTM FETAX Guideline (1991, 1998), the test report for an acceptable FETAX study should include the following information either directly or by reference to existing publications:

- The name of the test substance, the name of the investigator(s), the location of laboratory, and the dates of initiation and termination of test;

- The source of test substance, its lot number, composition (identities and concentrations of major ingredients and major impurities), known chemical and physical properties, and the identity and concentration(s) of any solvent used;

- If a dilution water other than FETAX Solution is used, its chemical characteristics and a description of any pretreatment;

- A recent analyses of FETAX Solution and adult culture water;

- pH measurements of the control and of the highest test concentrations at the end of each 24-hour time period;

- Available data on sample hardness, alkalinity, conductivity, total organic carbon (TOC), concentration of dissolved oxygen, and metal content;

- The mortality, malformation rates, and the mean embryo length at 96 hours in the dilution-water, FETAX Solution, or solvent control;

- The mortality and malformation results obtained for the positive control. If a full concentration-response curve was performed, then the 96-hour LC_{50}, the 96-hour EC_{50}, and their confidence limits should be reported;

- The 96-hour LC_{50}, the 96-hour $EC_{50,}$ the TI, and the MCIG for each test. Also, the geometric means of these values and their 95% confidence limits;

- Concentration-response data for mortality, malformation, and growth inhibition may be provided;

- A table for each test that lists the percent mortality, percent malformation, and the head to tail length at each concentration tested;

- The names of the statistical tests employed, the alpha-levels of the tests, and some measure of the variability of the hypothesis tested;

- The types, frequency, and severity of malformations; and

- Any deviations from a standard FETAX study (e.g., exposure periods exceeding 96 hours or pulse exposures, the use of static exposure techniques, not using FETAX Solution as the diluent).

Although this level of detail was specified in the ASTM FETAX Guideline (1991), such detail was not provided in any FETAX study report evaluated by NICEATM. The following were generally not included in FETAX study reports:

- a table for each test that listed the percent mortality, percent malformation, and the head to tail length at each concentration tested;

- quantitative information on the types, frequency, and severity of malformations detected;

- pH measurements of the control and of the highest test concentrations at the end of each 24-hour time period;

- quantitative information on mortality, malformation rates, and the mean embryo length at 96 hours in the dilution-water, FETAX Solution, or solvent control; and

- the mortality and malformation results obtained for the positive control.

2.2 Commonly Used Variations in the FETAX Standard Protocol and Rationale

The ASTM FETAX Guideline (1991, 1998) discusses other types of data that can be collected in FETAX, which increases its versatility. The types of data listed below have been collected in past studies (ASTM, 1991; 1998).

Collecting data on pigmentation might be useful for measuring neural damage because it is thought that the size of the pigment patches is under control of the nervous system. Agents that affect these nerves cause smaller pigment patches and the overall color of the 96-hour larvae will pale. Comparison to the standard Atlas of Abnormalities (Bantle et al., 1998) and suitable controls must be made to determine abnormal pigmentation. Other causes of depigmentation are possible, including loss of melanin production. A concentration-response curve can be generated and an EC_{50} related to pigmentation can be determined. Scoring for pigmentation is considered to be subjective.

Collecting locomotion data is potentially useful in measuring specific neural or muscle damage since larvae with substantial cellular damage swim poorly, erratically, or not at all. The ability to swim properly should be determined by comparison to appropriate controls. A concentration-response curve can be generated and an EC_{50} related to locomotion can be determined. Scoring for locomotion is considered to be subjective.

The embryos hatch from the fertilization membrane between 18 and 30 hours. The number failing to hatch at 48 hours could be recorded. Delay or failure indicates a slowing of developmental processes. This is analogous to developmental staging the embryos at the end of the 96-hour time period except that it is much easier to score hatching. A concentration-response curve can be generated and an EC_{50} related to hatching can be determined.

In special circumstances, exposure periods exceeding 96 hours or pulse exposures, or both, may be performed. Studies conducted with longer exposures should be reported as deviating from the standard FETAX assay. In the static technique, the test substance is added at the beginning of the test and is not changed. It should be recognized that many test substances degrade in a short time period. The static technique should only be used for test substances that are extremely stable and do not volatilize or sorb to the test dishes. The cost or the available amount of the test substance might also dictate that the static technique be used.

A toxicant-delivery system is used to continuously deliver toxicant and dilution water to the embryos in a flow-through system. Small glass containers with bottom screening are used to contain the embryos in a larger diluter apparatus. The flow-through technique is recommended for test substances that degrade quickly or are volatile. Every attempt should be made to use FETAX solution as the diluent. This variation in procedure must be reported as deviating from the standard FETAX assay.

2.3. Basis for Selection of FETAX

FETAX is proposed as a screen for human developmental hazards based on the conclusion of the developers that the assay is easy, rapid, reliable, inexpensive, and predictive of mammalian developmental hazards. FETAX is essentially an organogenesis test, and organogenesis is highly conserved across amphibians and mammals. The first 96 hours of embryonic development in *Xenopus* parallel many of the major processes of human organogenesis (ASTM, 1991; 1998).

2.4 Confidentiality of Information

Copies were obtained by NICEATM of all original data collected during the five FETAX validation studies (see **Section 7.0**). Original data was not sought by NICEATM for any other publication containing FETAX data or for any publication containing laboratory mammal or human data.

2.5 Basis for FETAX Decision Criteria

As specified by the ASTM FETAX Guideline (1991; 1998), three separate decision criteria (a TI value greater than 1.5, a MCIG/LC_{50} ratio less than 0.3, and severity of malformations) have been used to identify potential human teratogens. The ASTM FETAX Guideline (1991, 1998) concludes that any single decision criterion is sufficient to identify a potential teratogenic hazard, and that these three decision criteria are based on empirical evidence resulting from over 100 materials tested (without metabolic activation) in FETAX.

More recently, Bantle et al. (1999) and Fort et al. (2000a) have evaluated study results based on multiple decision criteria. In the Phase III.3 Validation Study conducted by Bantle et al. (1999; see **Section 7.2.5**), each test substance was judged to have developmental hazard when both the TI value and the MCIG/LC_{50} ratio indicated hazard (i.e., the TI >1.5 and the MCIG/LC_{50} <0.30), and definitely not hazardous when both decision criteria fell into the non-hazard category (i.e., the TI ≤1.5 and the MCIG/LC_{50} ≥0.30). The hazard was considered equivocal when any one of the two decision criteria suggested hazard. In such cases, the types and severity of malformations were examined for guidance in assessing teratogenic hazard. However, due to the subjectivity of malformation identification, this approach was not made a permanent part of the decision criteria. In the comparative FETAX - rat teratogenic study conducted by Fort et al. (2000a), a similar combined decision criteria approach was used except that the TI decision criterion was a TI value greater than 3.0. No further information for the basis of these criteria were provided.

2.6 Basis for Numbers of Replicates and Repeat Tests in FETAX

In FETAX, as defined by the ASTM FETAX Guideline (1991, 1998), one or more range-finding tests and three replicate definitive tests are performed on each test substance. Each of the three definitive tests is conducted using embryos from a different male/female pair of *X. laevis.* Each test consists of several different concentrations of the test substance with two replicate dishes at each test concentration and four replicate dishes for each control. Each plastic or glass Petri dish contains 20 or 25 embryos, respectively. The number of embryos per dish, the number of replicate dishes per test substance concentration, and the number of replicate tests per study were not based on a formal scientific analysis. Rather, selection was based on the best scientific judgement of the developers/users of the assay at the time the ASTM FETAX Guideline (1991, 1998) was prepared (J. Bantle and D. Fort, personal communication).

2.7 Validation Study Based Modifications to the Standard Protocol

The FETAX protocol used in the Phase I Validation Study (Bantle et al., 1994a) followed the 1991 ASTM FETAX Guideline. Based on the results obtained, several changes to the standard FETAX protocol were recommended by the investigators; including:

- increasing the acceptable malformation rate in FETAX Solution controls from 7% to 10%;
- distributing 25-mL volumes of the toxicant solution to 50-mL flasks prior to aliquoting into dishes; and
- potentially eliminating 6-AN as the positive reference control.

The first recommendation was based on the larger than anticipated range in the incidence of malformations among control cultures in several laboratories. The purpose of the second recommendation was to potentially reduce intra-dish variability within a treatment group. The recommendation for potentially eliminating 6-AN as the positive control for studies conducted without metabolic activation was based on the extensive variability seen within and across laboratories for this test material. However, a possible replacement positive control for 6-AN has not yet been identified (J. Bantle, personal communication) and this chemical was still recommended as reference control in studies conducted without metabolic activation in the revised 1998 ASTM FETAX Guideline.

The FETAX protocol used in the subsequent validation studies (Phase II, Phase III.1, Phase III.2, and Phase III.3) incorporated the first two protocol changes recommended in Phase I. In Phase III.2 and Phase III.3, the validation protocol was modified to include an exogenous MAS and CP as the appropriate concurrent positive to demonstrate the suitability of the MAS for bioactivation. Also, in Phase III.2 and III.3, 20 embryos were used per dish rather than the 25 recommended by the ASTM FETAX Guideline (1991, 1998). This modification to the protocol was due to the use of slightly smaller plastic Petri dishes in studies incorporating an MAS. The arithmetic mean rather than the geometric mean recommended by the ASTM FETAX Guideline (1991, 1998) was calculated for the 96-hour LC_{50}, the 96-hour $EC_{50,}$ the TI, and the MCIG in these studies.

2.8 Section 2 Conclusions

The 1991 and the revised and expanded 1998 FETAX Guideline published by ASTM are detailed, comprehensive, and well-structured. Adequate information is provided on the necessary materials, equipment, and supplies; range-finding and definitive tests; endpoint (mortality, malformations, and embryonic growth) assessment; nature of the responses assessed; the duration of exposure; data collection and data storage media; measures of variability; statistical and non-statistical methods; test report information; commonly used protocol variations and rationale; the use of alternative species; and the basis for selection of FETAX.

Known limits of use for FETAX were not described, except that it was stated that the test method is incompatible with materials (or concentrations of materials) that alter the pH, hardness, alkalinity, and conductivity of the FETAX solution beyond the acceptable range indicated by the ASTM FETAX Guideline (1991, 1998). It would also be expected that the testing of water insoluble materials would be limited by the highest concentration that can be achieved using an appropriate organic solvent (and concentration) that does not alter embryonic growth or survival.

Appropriate vehicle, negative, and positive controls were described. The recommended positive controls were 6-AN for studies without metabolic activation and CP for studies with metabolic activation (ASTM, 1991; 1998). However, one conclusion of the Phase I Validation Study (Bantle et al., 1994a) was that 6-AN was not an appropriate positive control for studies without metabolic activation and that another chemical should be identified for this purpose. To date, a replacement for 6-AN has not been identified. The ASTM FETAX Guideline (1991, 1998) recommends that concentration-response experiments for 6-AN be performed at least quarterly and the results of these tests compared with historical tests to judge the laboratory quality of FETAX data. NICEATM concluded that the inclusion of a concurrent positive control in each study without metabolic activation should be considered.

Information on the acceptable range of negative control response for FETAX was provided in the ASTM FETAX Guideline (1998). It was also stated that the reference toxicant 6-AN test must produce data within two standard deviations of the historical mean values. However, no information was provided on the number of experiments required to generate appropriate historical data or the time period over which such data should be retrospectively assessed. For studies conducted with metabolic activation, the bioactivated CP should kill 100% of the embryos within 96 hours. A response of this magnitude limits the utility of historical control data and the use of a test concentration that would enable an analysis of both mortality and malformation data should be considered.

The ASTM FETAX Guideline (1991, 1998) specifies that the geometrical mean for the 96-hour LC_{50}, the 96-hour $EC_{50,}$ the TI, and the MCIG, as well as their 95% confidence limits be calculated using the data from the three replicate definitive tests and provided in the study report. However, in all reports evaluated, the arithmetic mean only has been calculated, and 95% confidence limits were generally not provided.

The three decision criteria used to distinguish between a teratogen and a non-teratogen in FETAX are well described in the ASTM FETAX Guideline (1991, 1998). In the ASTM Guideline, it was stated that these three decision criteria are based on empirical evidence resulting from over 100 materials tested in FETAX, without metabolic activation. Data to support this statement were not provided. Recently, Bantle et al. (1999) and Fort et al. (2000a) have also evaluated study results based on multiple decision criteria. In their analysis, Fort et al. (2000a) increased the TI decision point value from 1.5 to 3.0. In addition, in both studies, the types and severity of malformations were examined for guidance in assessing teratogenic hazard. As judged by NICEATM, the use of multiple decision criteria rather than single decision criterion does not appear to improve the performance characteristics of FETAX against laboratory mammal or human data (see **Section 6.6**).

Selection of the number of embryos per dish (i.e., 20 or 25), the number of replicate dishes per test concentration (i.e., two), and the number of replicate tests per FETAX definitive study (i.e., three) were based on the best scientific judgement of the developers/users of the assay at the time the ASTM FETAX Guideline (1991, 1998) was developed (J. Bantle and D. Fort, personal communication). It may be useful to conduct a formal analysis of the impact of different numbers of embryos per dish, dishes per test concentration, and replicate definitive tests on the performance of FETAX.

3.0. CHARACTERIZATION OF SUBSTANCES TESTED IN FETAX

FETAX test data from 276 studies involving 137 substances, not including environmental samples, were located, reviewed, extracted, and entered into the NICEATM FETAX database (**Appendix 2** contains substances tested without metabolic activation, **Appendix 3** contains substances tested with metabolic activation). Sources for these data included peer-reviewed literature (including studies accepted for publication) and non peer-reviewed book chapters. Excluded from consideration was information provided in abstracts, manuscripts not accepted for publication, publications that did not provide quantitative data, studies conducted where the test substances were not identified, and studies conducted that did not follow the general FETAX protocol described in the ASTM FETAX Guideline (1991, 1998).

3.1 Rationale for Chemicals/Products Selected for FETAX Validation Studies

Only limited information is available on the selection rationale for the chemicals/products tested in the five FETAX validation studies. It does not appear that selection was based on testing substances that represented a range of chemical or product classes. Rather, selection appeared to have been based primarily on the availability of prior FETAX test results and laboratory mammal teratological data. Specific chemical selection rationale for each FETAX validation study is presented by individual validation study.

Validation Study Phase I was classified as a training and protocol evaluation phase (Bantle et al., 1994a). 6-AN, hydroxyurea, and isoniazid were selected for testing without metabolic activation based on their positive performance in previous FETAX studies (Bantle et al., 1994a).

In Phase II (Bantle et al., 1994b), caffeine, 5-fluorouracil, saccharin, and sodium cyclamate were tested without metabolic activation. These test substances were selected for testing without metabolic activation based on their negative (saccharin, sodium cyclamate) and positive (caffeine, 5-fluorouracil) performance in previous FETAX studies (Bantle et al., 1994b).

Validation Study Phase III.1 (Bantle et al., 1996) involved the testing, without metabolic activation, of β-aminopropionitrile, ascorbic acid, copper sulfate, monosodium glutamate, sodium acetate, and sodium arsenate. The rationale for selecting these test substances was not provided in the validation report. Three of the six test substances (ascorbic acid, sodium acetate, copper sulfate) had been tested previously in FETAX. In laboratory mammals, ascorbic acid, monosodium glutamate, and sodium acetate are non-teratogenic, while sodium arsenate and copper sulfate are teratogenic.

The purpose of Validation Study Phase III.2 (Fort et al., 1998) was to conduct an inter-laboratory validation of an exogenous MAS developed for use with FETAX. Caffeine and CP were tested, with and without metabolic activation, and were selected based on their activation profiles. CP is efficiently bioactivated by P-450 to reactive metabolites, while the addition of metabolic activation was not anticipated to significantly alter the response of *X. laevis* to caffeine. CP is a human and laboratory mammal teratogen; caffeine is a teratogen in laboratory mammals but not humans.

Validation Study Phase III.3 involved the testing, with and without metabolic activation, 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 dimethyl ether) (Bantle et al., 1999). The rationale for the selection of the test substances was not provided in the validation report. However, it is likely that selection was based on the availability of relevant laboratory mammal data and the suitability of the test substance for testing in FETAX (e.g., water solubility, lack of volatility). Of the 12 substances tested, Bantle et al. (1999) reported that seven (boric acid, dichloroacetate, sodium arsenite, sodium bromate, sodium iodoacetate, tribromoacetic acid, triethylene glycol dimethylether) were classified as teratogens in laboratory mammals, two (glycerol, phthalic acid) were classified as non-teratogens in laboratory mammals, and three (ascorbic acid, sodium acetate, copper sulfate) were classified as equivocal with respect to laboratory mammal teratogenicity (i.e., were not consistently positive in all laboratory mammal species tested).

3.2 Rationale for the Numbers of Chemicals/Products Tested in FETAX

A rationale for the numbers of chemicals/products tested in each of the five validation studies was not provided. However, the most likely basis was the extent of available funding.

3.3 Description of Chemical and Product Classes Evaluated in FETAX

Information on chemical and product classes for substances tested in FETAX are provided in **Appendix 1**; the most common chemical and product classes are provided in **Tables 1a** and **1b**, respectively. Substances were assigned to chemical classes based on available information from standardized references (e.g., *The Merck Index* [Budavari, 1996]) and from an assessment of chemical structure by an organic chemist. The most numerically prevalent chemical classes were

Table 1a. Major Chemical Classes Evaluated with FETAX

Major Chemical Classes	Number of Chemicals
Alcohols (including glycols)	22
Amides	16
Amides and Hydrazides	29*
Amines	19*
Halogenated Organic Compounds	12
Esters	12
Heavy Metals	14
Hydrazides and Hydrazines	14
Nitrogen Heterocyclic Compounds	40*
Organic (Phenolic and Carboxylic) Acids	24*
Salts	20
Total	260

*Classes indicated had adequate comparative data (i.e., at least 15 chemicals with FETAX and either laboratory mammal or human study results) to warrant an assessment of performance (**Section 6**).

Table 1b. Major Product Classes Evaluated with FETAX

Major Product Classes	Number of Products
Antimicrobials	5
Chemical Synthesis	17
Cosmetics	6
Dyes	7
Food Additives	11
Fossil Fuels	6
Pesticides	13
Pharmaceuticals	45*
Photographic Chemicals	5
Polymers	6
Total	121

*Classes indicated had adequate comparative data (i.e., at least 15 chemicals with both FETAX and either laboratory mammal or human study results) to warrant an assessment of performance (**Section 6**)

alcohols (including glycols); amides; amines; halogenated organic compounds; esters; heavy metals and their salts; hydrazides and hydrazines; nitrogen heterocyclic compounds; organic (phenolic and carboxylic) acids; and salts. Of the 137 substances tested in FETAX, 8 substances were not classified within these chemical classes, 67 substances were included in one chemical class, 41 substances were included in two chemical classes, 15 substances were included in three chemical classes, three substances were included in four chemical classes, two substances were included in five chemical classes, and one substances was included in six chemical classes.

Product classes were assigned based primarily on ChemFinder and *The Merck Index.* The most common product classes tested in FETAX were antimicrobials, chemical synthesis, cosmetics, dyes, food additives, fossil fuels, pesticides, pharmaceuticals, photographic chemicals, and polymers (including monomers). Of the 137 substances tested in FETAX, 63 substances were not classified within these product classes, 50 substances were included in one product class, 14 substances were included in two product classes, seven substances were included in three product classes, and three substances were included in four product classes.

3.4 Coding Used in FETAX Validation Studies

Coded chemicals were not used in the Phase I Validation Study (Bantle et al., 1994a), but were used in the Phase II (Bantle et al., 1994b), Phase III.1 (Bantle et al., 1996), Phase III.2 (Fort et al., 1998), and Phase III.3 (Bantle et al., 1999) Validation Studies.

3.5 FETAX-Tested Substances in the Smith et al. (1983) List of Candidate Substances/Conditions for *In Vitro* Teratogenesis Test Validation

In 1983, Smith et al. published a list of candidate substances/conditions for *in vitro* teratogenesis test validation. NICEATM identified the number of Smith list substances evaluated in FETAX, with or without metabolic activation (**Table 2**). NICEATM also identified those substances listed by Smith et al. (1983) that might be expected to require metabolic activation before a teratogenic response would be induced. This identification was based on whether the substance was positive in one or more *in vitro* genetic toxicological tests (generally the *Salmonella typhimurium* reverse mutation assay) with, but not without, metabolic activation. *In vitro* genetic toxicology data were obtained from the EPA Genetic Activity Profile (GAP) database (www.epa.gov/gapdb/) and the NTP Salmonella test database. This method for identifying substances that may require metabolic activation to be teratogenic *in vitro* assumes a common mechanism between mutagenicity and teratogenicity that may not be valid. Of the 47 substances listed, 26 substances (55%) were tested in FETAX without metabolic activation, while nine of these 26 substances (19% of the total list) were tested also with metabolic activation. Of the nine substances tested with metabolic activation, relevant *in vitro* genetic toxicology data were located for seven. Two of these seven substances potentially require metabolic activation to be teratogenic.

Table 2. Smith et al. (1983) Suggested List of Substances/Conditions for *In Vitro* Teratogenesis Testing

Substance	**Tested in FETAX**	
	Without Activation	**With Activation**
Acetozolamide	Not Tested	Not Tested
Amaranth	**Tested**	Not Tested
6-Aminonicotinamide	**Tested**	Not Tested
Aspirin	Not Tested	Not Tested
Caffeine*	**Tested**	**Tested**
Carbon tetrachloride*	Not Tested	Not Tested
Chlorambucil**	Not Tested	Not Tested
Coumarin*	**Tested**	Not Tested
Cyclophosphamide**	**Tested**	**Tested**
Cytochalasin D*	**Tested**	**Tested**
Dexamethasone	Not Tested	Not Tested
Diazapam*	**Tested**	Not Tested
Diethylstilbestrol*	Not Tested	Not Tested
Dilantin	**Tested**	**Tested**
Diphenylhydramine HCl	**Tested**	Not Tested
Doxylamine succinate*	**Tested**	**Tested**
EM12	Not Tested	Not Tested
Ethyl alcohol*	**Tested**	Not Tested
Ethylenethiourea*	Not Tested	Not Tested
N-Ethyl-N-nitrosourea*	**Tested**	**Tested**
5-Fluorouracil*	**Tested**	Not Tested
Formaldehyde*	Not Tested	Not Tested
Hexahydrophthalimide glutarimide	Not Tested	Not Tested
Hydroxyurea*	**Tested**	Not Tested
Hyperthermia	Not Tested	Not Tested

Table 2. Smith et al. (1983) Suggested List of Substances/Conditions for *In Vitro* Teratogenesis Testing (Continued)

Substance	Tested in FETAX	
	Without Activation	With Activation
Isoniazid*	**Tested**	**Tested**
Meprobamate	Not Tested	Not Tested
Methotrexate*	**Tested**	Not Tested
Methyl mercury chloride*	**Tested**	Not Tested
Mirex	Not Tested	Not Tested
Nitrilotriacetate*	**Tested**	Not Tested
Penicillin G	Not Tested	Not Tested
L-Phenylalanine	Not Tested	Not Tested
Phthalimide	Not Tested	Not Tested
Procarbazine*	**Tested**	Not Tested
Retenoic acid (all trans)*	**Tested**	Not Tested
Retinoic acid –13 cis*	**Tested**	Not Tested
Saccharin*	**Tested**	Not Tested
Sodium arsenate*	**Tested**	Not Tested
Sodium cyclamate*	**Tested**	Not Tested
Testosterone proprionate	Not Tested	Not Tested
Thalidomide	Not Tested	Not Tested
Trichloroethylene*	**Tested**	**Tested**
Trichlorophenoxyacetic acid*	Not Tested	Not Tested
Urethane**	**Tested**	**Tested**
Vincristine sulfate*	Not Tested	Not Tested
Vinyl chloride**	Not Tested	Not Tested

Bolded chemical names indicate substances tested in FETAX without and/or with MAS.
* or ** indicates chemicals that do not or do appear to require metabolic activation, respectively, to induce a positive response in an *in vitro* genetic toxicological test according to the EPA Genetic Activity Profile (GAP) database (www.epa.gov/gapdb/) and the NTP Salmonella test database.

3.6 Section 3 Conclusions

In the five FETAX validation studies, it appears that selection rationale for the substances tested was based primarily on the availability of prior FETAX test results and laboratory mammal teratological data rather than on selecting materials with relevant mammal/human data that represented a range of chemical or product classes. A rationale for the numbers of substances tested in each of the five validation studies was not provided. The most likely explanation is the level of available funding. Coded substances were used in all but the first of five validation studies. However, in the Phase II Validation study, all laboratories used the same preset test substance concentrations. If additional validation studies are considered for FETAX, more substances on the Smith et al. list or an updated list should be considered for inclusion. Also, consideration should be given to the role of metabolic activation in *in vitro* teratogenicity studies, and in the identification of appropriate substances to test with metabolic activation.

4.0 REFERENCE DATA USED FOR AN ASSESSMENT OF FETAX PERFORMANCE CHARACTERISTICS

4.1 Description of Laboratory Mammal and Human Reference Data Sources

Reference teratogenic data were obtained from several general sources listed below. If teratogenesis and developmental toxicity studies were not listed for a particular substance in these general sources, NICEATM staff searched the Developmental and Reproductive Toxicology (DART) database, available through the TOXNET system (http://sis.nlm.nih.gov/sis1/), a product of the National Library of Medicine (NLM), and the ReproTox® System, produced by the Reproductive Toxicology Center and available on the MICROMEDIX' TOMES CPS™ CD-ROM. Keywords included specific chemical names, synonyms, and Chemical Abstract Service Registry Numbers (CASRN). For substances not located in these two databases, NLM's MEDLINE and TOXLINE databases were also searched for teratogenicity information.

- Friedman, J.M., and J.E. Polifka. 1994. Teratogenic Effects of Drugs. A Resource for Clinicians (TERIS). Johns Hopkins University Press, Baltimore, MD.

- National Institute for Occupational Safety and Health (NIOSH). RTECS® (Registry of Toxic Effects of Chemical Substances). On: the TOXNET® system. Internet Resource Internet Resource (http://sis.nlm.nih.gov/sis1/).

- National Library of Medicine (NLM). HSDB® (Hazardous Substances Data Bank). On: the TOXNET® system. Internet Resource Internet Resource (http://sis.nlm.nih.gov/sis1/).

- Schardein, J.L. 1993. Chemically Induced Birth Defects, 2nd Edition, Marcel Dekker, Inc, New York, NY.

- Shepard, T.H. 1995. Catalog of Teratogenic Agents. 8th Edition. John Hopkins University Press, Baltimore, MD.

- Smith, M.K., G.L. Kimmel, D.M. Kochhar, T.H. Shepard, S.P. Spielberg, and J.G. Wilson. 1983. A selection of candidate compounds for *in vitro* teratogenesis test validation. Teratog. Carcinog. Mutagen. 3:461-480

- Szabo, K.T. 1989. Congenital Malformations in Laboratory and Farm Animals. Academic Press, Inc., New York, NY.

In the reference data collection process conducted by NICEATM, there was no intent to collect all laboratory mammal and human teratogenicity data (i.e., the search strategy was limited to substances tested in FETAX), to obtain original data for the reference studies, to evaluate the appropriateness of the study design, or to critically review the scientific merit of the conclusions of the investigator. In considering the reference data, a weight-of-evidence approach was not used in classifying a substance as a teratogen or non-teratogen. Rather, the presence of at least one positive teratogenic study resulted in the substance being classified as a teratogen for the species evaluated. While potentially resulting in some false positive classifications, this approach was considered by NICEATM to be the most conservative.

4.2. Laboratory Mammal Reference Data

The laboratory mammal reference data are provided by substance in **Appendix 4**. Laboratory mammal (mouse, rat, and rabbit) teratogenicity data were obtained for 90 of the 137 substances evaluated in FETAX plus one environmental sample. These data were entered by individual species. Data on the teratogenicity of these substances in other species, both mammalian and non-mammalian, were included as a separate entry, where identified. Where available, descriptive information on the types of malformations observed was included in the database. In using these data to evaluate the performance characteristics of FETAX against combined laboratory mammal (i.e., rat, mouse, and rabbit) results, positive studies were given weight over negative studies within an individual species and, where multiple species had been evaluated, the overall teratogenicity classification was made on the basis of a positive response in any single species. In addition, the performance characteristics of FETAX against each of the three primary laboratory mammal species were calculated. Data from the other non-human species were not considered in an evaluation of the performance characteristics of FETAX.

4.3 Availability of Original Laboratory Mammal Reference Test Data

The availability of original test data for the reference mammalian assays is not known.

4.4 Laboratory Mammal Reference Data Quality

Generally, teratogenicity findings for laboratory mammals (e.g., rat, mouse, and rabbit) were obtained from reviews, compilations of data, or individual published reports. The sources used were considered authoritarian for this purpose. However, the quality of the data in terms of accuracy and whether the studies were conducted in compliance with national/international Good Laboratory Practice (GLP) Guidelines is not known.

4.5 Availability and Use of Human Teratogenicity Data

Human teratogenicity data were obtained for 34 chemicals from the sources listed in **Section 4.1**. These data are summarized in **Appendix 4**.

A single positive human study was considered to be definitive for the purpose of classifying a substance as a human teratogen. While potentially resulting in false positive classifications, this approach was considered by NICEATM to be the most conservative for the purpose of analyzing the performance characteristics of FETAX against the human database.

4.6 Section 4 Conclusions

Reference teratogenic data were obtained from general sources; additional information (e.g., research papers, literature reviews, book chapters) were located by searching the DART database, the ReproTox® System, and the MEDLINE and TOXLINE databases. Studies using humans, rats, mice, rabbits, and other species (both mammalian and non-mammalian) were considered. Sources for human data included case reports, epidemiological studies, case-control studies, literature reviews, and other secondary references. The search was not intended to be comprehensive; only substances tested in FETAX were considered and no effort was made to critically evaluate the conclusions of the investigator. In classifying substances as teratogens or non-teratogens in rats, mice, rabbits, or humans, a single positive study was sufficient to classify the substance as a teratogen. This approach may have resulted in some false-positive classifications within the database. A critical evaluation of the current laboratory mammal (rat, mouse, rabbit) and human teratogenicity databases by appropriate experts would be an important contribution to this field of investigation, and to the development and validation of alternative *in vitro* teratogenicity assays.

5.0 FETAX TEST METHOD DATA AND RESULTS

5.1 Availability of Detailed FETAX Protocol

A comprehensive ASTM guideline for FETAX was published in 1991 and a revised guideline was published in 1998. The 1991 and 1998 versions of the ASTM FETAX Guideline are provided in **Appendix 10** and **11**, respectively. The protocol used in the FETAX Phase I Validation Study followed the 1991 ASTM Guideline (Bantle et al., 1994a). This guideline, with minor modifications, was followed in the Phase II (Bantle et al., 1994b), Phase III.1 (Bantle et al., 1996), Phase III.2 (Fort et al., 1998), and Phase III.3 (Bantle et al., 1999) Validation Studies. **Section 7.2** discusses each validation study and any protocol modifications. Unless noted otherwise, the 1991 ASTM FETAX Guideline was followed in the other FETAX studies.

5.2 Availability of Original and Derived FETAX Data

Original and derived data were obtained for each of the five FETAX validation studies from Dr. Bantle, the lead investigator.

5.3 Statistical Approach used to Evaluate FETAX Data

The statistical and non-statistical methods used by the individual investigator to analyze FETAX data obtained in their laboratory are described in **Section 2.1.13**. To obtain a consensus call for each substance tested in each validation study, the validation study management team determined the average of the calculated LC_{50}, EC_{50}, TI, and MCIG values among 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 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.

5.4 FETAX Test Results for Individual Substances

FETAX test data from 276 separate studies involving 137 individual substances (not including environmental samples) were located, reviewed, extracted, and entered into the NICEATM FETAX database (**Appendix 2** contains substances tested without metabolic activation, **Appendix 3** contains substances tested with metabolic activation). Sources for these data included peer-reviewed literature (including studies accepted for publication) and non peer-reviewed book chapters. Information provided in abstracts and manuscripts not accepted for publication were not considered. All 137 substances had been tested using FETAX without metabolic activation; 35 of these 137 substances had also been tested with metabolic activation.

A summary of the responses for substances tested multiple times, as well as the weight-of-evidence conclusion, are provided in **Appendix 6**.

FETAX test results are classified in the database as positive or negative based on the criteria provided in the ASTM FETAX Guideline (1991, 1998) (i.e., positive if the TI value is greater than 1.5 or if the MCIG/LC_{50} ratio is less than 0.30). Also, in keeping with a recent study (Fort et al., 2000a), FETAX test results are classified as positive if the TI value is greater than 3.0. In addition, consistent with recent studies where both the TI value and the MCIG/LC_{50} ratio were considered together in classifying FETAX results, compounds are classified as positive based on obtaining concordant positive results for both endpoints using a TI value greater than 1.5 and an MCIG/LC_{50} ratio less than 0.30 (Bantle et al., 1999), or using a TI value greater than 3.0 and an MCIG/LC_{50} ratio less than 0.30 (Fort et al., 2000a), negative if neither the TI value or the MCIG/LC_{50} ratio were positive, and equivocal if only one of the two endpoints were positive. Due to the lack of quantitative *X. laevis* malformation data in the majority of publications, this endpoint was not considered in the assessment of performance characteristics by NICEATM. The importance of using agent-specific characteristic abnormalities in classifying materials as positive in FETAX is discussed in **Section 6.6.2**.

5.4.1 FETAX Test Results Without Metabolic Activation

Of the 137 substances tested without metabolic activation, 105 substances were tested only once. The remaining 32 substances were tested in multiple studies. The number of multiple studies ranged from three to 14. TI data were available for all studies, while MCIG data were available for 96 (70%) of the test substances. Qualitative data on malformations observed in *X. laevis* without metabolic activation were available for 35 substances, including three environmental samples. Quantitative malformation data by test substance concentration were not provided in any study.

5.4.2 FETAX Test Results With Metabolic Activation

Of the 35 substances tested with metabolic activation, 21 were tested only once. The remaining 14 substances were tested in multiple studies ranging from three and eight. TI data were available for all studies, while MCIG data were available for 27 (77%) of the test substances. Qualitative data on malformations observed in *X. laevis* without metabolic activation were available for six substances; quantitative malformation data by test substance concentration were not provided in any study.

5.5 FETAX Test Results With Binary Mixtures

FETAX has been used also to assess the teratogenicity and embryotoxicity of binary mixtures, in the absence of metabolic activation only. The rates of malformation by binary mixtures are expected to depend on the mode of teratogenesis for the component substances of the mixture. For those mixtures comprised of substances that follow the same modes of action, concentration-addition rates of malformation are expected. In contrast, response-addition rates are expected for those mixtures containing substances with different modes of action.

Dawson and Wilke (1991a, b) tested a total of 12 defined binary mixtures (**Table 3**) using FETAX. In the first study (Dawson and Wilke, 1991a), three mixtures were tested using ratios of 0:1, 3:1, 1:1, 1:3, and 1:0. Compound selection was based on their different modes of teratogenicity and their mortality/malformation index (MMI). All of the mixtures tested

Table 3. Teratogenicity/Embryolethality of Binary Mixtures Tested in FETAX (Dawson and Wilke, 1991a, b).

Mixture	Mixture Ratio	Toxic Units[1] (mixture)	Displayed Effect	Malformations Observed
Semicarbazide:Isoniazid	3:1	1.09	Concentration addition	Skeletal
	1:1	1.03		
	1:3	1.02		
Valproic acid:Pentanoic acid	3:1	1.02	Concentration addition	Head and osmoregulatory
	1:1	1.03		
	1:3	0.98		
Butyric acid:Pentanoic acid	3:1	0.98	Concentration addition	Head and osmoregulatory
	1:1	1.06		
	1:3	1.06		
Hydroxyurea:Isoniazid	3:1	1.15	Response addition	Skeletal, head, visceral and osmoregulatory
	1:1	1.29		
	1:3	1.29		
Isoniazid:6-Aminonicotinamide	3:1	1.23	Response addition	Skeletal and eye
	1:1	1.27		
	1:3	1.15		
Isoniazid:Retinoic acid	3:1	1.23	Response addition	Skeletal and mouth
	1:1	1.22		
	1:3	1.30		
Hydroxyurea:Retinoic acid	3:1	1.25	Response addition	Skeletal, head, mouth, visceral, and osmoregulatory
	1:1	1.39		
	1:3	1.35		
6-Aminonicotinamide: Retinoic acid	3:1	1.24	Response addition	Eye and mouth
	1:1	1.36		
	1:3	1.27		
Retinoic acid:Nicotine	3:1	1.35	No interaction[2]	Mouth
	1:1	1.76		
	1:3	1.37		
Isoniazid: ß-Aminopropionitrile	3:1	0.97	Response addition	Connective tissue lesions (typical of osteolathyrism), visceral edema, gut mis-coiling, facial malformations
	1:1	1.01		
	1:3	1.00		
Valproic acid:Butyric acid	3:1	1.01	Response addition	Reduced head size, visceral/cranial edema, poor gut coiling, skeletal kinking, occasional mouth/eye defects.
	1:1	0.96		
	1:3	0.98		
Isoniazid:Valproic acid	3:1	1.33	Response addition	Not provided
	1:1	1.53		
	1:3	1.19		
Semicarbazide:Isoniazid (embryolethality)	3:1	1.12	Response addition	Not evaluated
	1:1	1.12		
	1:3	1.11		
Hydroxyuera:Isoniazid (embryolethality)	3:1	1.52	Response addition[3]	Not evaluated
	1:1	1.35		
	1:3	1.15		

[1] Toxic unit = EC_{50} in mixture/EC_{50} alone; [2] Effect was not greater than that observed for each compound individually; [3] One concentration (3:1) produced a TU value indicative of antagonism. This was attributed to an excess concentration of isoniazid, which effected the efficiency of the absorption of hydroxyurea in the mixture.

displayed response addition. In the second study (Dawson and Wilke, 1991b), nine binary mixtures of developmental toxicants were tested for teratogenicity. Each of the mixtures was tested using ratios of 0:1, 3:1, 1:1, 1:3, and 1:0. The mixtures were analyzed using the toxic unit (TU) method, which is based on an individual substance's EC_{50} value being defined as 1.0 TU for malformation induced in *X. laevis* by that substance (Dawson, 1991). Three of these mixtures had calculated TU values near 1.0, which is indicative of concentration addition. The remaining six mixtures displayed TU values greater than 1.0, indicative of response addition. The investigators concluded that the results of these studies indicated that a developmental endpoint could be useful in the assessment of joint toxic action studies (Dawson and Wilke, 1991a, b).

Dawson and Wilke (1991b) also tested two binary mixtures—semicarbazide:isoniazid and hydroxyurea:isoniazid—for lethal effects. The semicarbazide:isoniazid mixture displayed response addition, although both substances are known to have the same mode of action for teratogenic effects (osteolathyrism). This suggested to the investigators that the two substances followed different modes of action for embryolethality. The hydroxyurea:isoniazid mixture was also found to display response addition, although an antagonistic TU value for the 3:1 ratio was observed. This result was attributed to the high relative concentration of isoniazid reducing the efficiency of hydroxyurea absorption and, therefore, its contribution to the mixture's lethality (Dawson and Wilke, 1991b).

A mixture comprised of ten aliphatic carboxylic acids was tested using FETAX malformations as an endpoint (Dawson, 1991). The results of this study are shown in **Table 4**. Based on the TU method, Dawson concluded that the mixtures displayed a concentration additive response.

Dawson and Wilke (1996) conducted an extensive evaluation of malformation dose-response curves for binary mixtures of differently acting teratogenic substances in FETAX. This study was purported to be the first to examine substances in combination where only one of the agents was present at an effective dose in the mixture. The substances tested were 6-AN, β-aminoproprionitrile, benzoic hydrazide, butyric acid, cytarabine, 2-ethylhexanoic acid,

Table 4. Malformation Information for Ten Carboxylic Acid Mixtures Tested in FETAX (Dawson, 1991)

Concentration of Mixture (mL)[1]	Number of Embryos Exposed/ Survivors	Number of Malformations	Malformations Observed[2]
0	75/75	12	Skeletal kinking (3), microcephaly (2), gut coiling (2), eye edema/blister (2), general edema (2), mouth (1)
2	75/75	26	Microcephaly (6), gut coiling (5), skeletal kinking (5), general edema (4), mouth (3), eye edema/blister (3)
4	75/75	41	Microcephaly (18), gut coiling (14), mouth (4), eye edema/blister (4), skeletal kinking (1)
5	75/75	53	Microcephaly (25), gut coiling (20), eye edema/blister (3), mouth (2), skeletal kinking (2), general edema (1)
6	75/75	76	Microcephaly (35), gut coiling (32), skeletal kinking (3), eye edema/blister (3), general edema (2), mouth (1)
8	75/75	127	Microcephaly (56), gut coiling (52), eye edema/blister (8), skeletal kinking (5), mouth (4), general edema (2)

[1] Total solution in exposure dishes = 10 mL

[2] Numbers in parenthesis indicate the number of embryos that exhibit the preceding malformation.

5-fluorouracil, hydroxyurea, isoniazid, penicillamine, pentanoic acid, trans-retinoic acid, thiosemicarbazide, and valproic acid. The binary mixtures were prepared such that in mixtures of the agents was present in almost ineffective concentrations. For 16 pairs of substances, the 1:1 mixture was slightly more effective in inducing malformations than would be expected based on additivity alone. In contrast, the 1:3 and 3:1 mixtures were not more effective than the effective agent in that combination alone.

FETAX tests have also been performed on mixtures of mixed xylenes and toluene (Kononen and Gorski, 1997). The LC_{50} and EC_{50} values in these experiments were less than predicted, thus indicating possible synergism between the two substances. However, confidence levels were not calculated at the various concentration levels and therefore further testing would need to be performed to confirm this interpretation.

Based on the information presented, FETAX appears to be useful for conducting toxicity assessments on substance mixtures. Both embryolethality and malformation are relevant endpoints to be evaluated when assessing mixtures, although modes of action also need to be considered. Embryolethality is best used for non-teratogenic mixtures since the mode of action does not effect the outcome of testing. Teratogens are best evaluated using the malformation endpoint due to the likelihood of separate modes of action for malformation and lethality that would make interpretation of results difficult. The need for a developmental malformation endpoint was stressed as a means of identifying chronic toxicity rendered by developmental abnormalities (Dawson and Wilke, 1991b).

5.6 Use of Coded Chemicals and Compliance with GLP Guidelines

Coded substances were not used in the Phase I Validation Study (Bantle et al., 1994a), but were used in the Phase II (Bantle et al., 1994b), Phase III.1 (Bantle et al., 1996), Phase III.2 (Fort et al., 1998), and Phase III.3 (Bantle et al., 1999) Validation Studies. It does not appear that blind coding was used in any other FETAX study. However, in the Phase II Validation Study, the same preset test concentrations were used by all laboratories for each test substance.

FETAX validation studies were not conducted in compliance with national or international GLP guidelines, nor were they generally conducted at facilities at which GLP studies are normally conducted. It does not appear that any FETAX study was conducted in compliance with GLP guidelines.

5.7 Availability of Non-Audited FETAX Data

None of the FETAX data obtained by NICEATM had been audited by a Quality Assurance Unit. However, copies of all original data collected in the five FETAX validation studies were obtained by NICEATM for a possible independent audit. (see **Section 7.0**). Original data was not sought by NICEATM for any other FETAX study.

5.8 Section 5 Conclusions

A detailed ASTM FETAX protocol was first published in 1991. With minor exceptions, the FETAX validation studies followed this protocol. Original and derived data were obtained for all five FETAX validation studies only; no attempt was made by NICEATM to obtain any other original FETAX data.

The averaging method used in the FETAX validation studies for achieving a consensus call 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. The relative appropriateness and merits of these two approaches should be evaluated.

The FETAX database includes 276 separate studies involving 137 substances. All 137 substances had been tested using FETAX without metabolic activation; 35 of these 137 substances had also been tested with metabolic activation. FETAX has been used to assess the teratogenicity and embryotoxicity of defined binary mixtures. Both embryolethality and malformation are relevant endpoints to be evaluated when assessing mixtures, although modes of action also need to be considered. NICEATM has concluded that the potential utility of FETAX for this purpose merits additional investigation.

Except for the most recent four of five FETAX validation studies, it does not appear that blind coding was used to eliminate potential bias in any other FETAX study. Also, it does not appear that any FETAX studies were conducted in compliance with national or intemational GLP guidelines.

The effect of these two issues on the quality of the data in the FETAX database is difficult to ascertain.

6.0 PERFORMANCE CHARACTERISTICS OF FETAX

The performance characteristics (i.e., accuracy, sensitivity, specificity, positive predictivity, negative predictivity, false positive rate, and false negative rate) of FETAX compared to either rat, mice, and/or rabbit teratogenicity test results or human teratogenicity study results were determined by NICEATM. FETAX studies that did not follow the ASTM FETAX Guideline (1991, 1998), especially in regard to data presentation and analysis, were excluded from consideration of performance characteristics. The decision criteria used in determining the performance characteristics of FETAX included:

- single decision criteria (TI >1.5; MCIG/LC_{50} <0.30) for identifying teratogenic potential, as defined by the ASTM FETAX Guideline (1991, 1998);

- modified single decision criterion (TI >3.0) for identifying teratogenic potential, as used in a recent study by Fort et al. (2000a);

- multiple decision criterion (TI >1.5 plus MCIG/LC_{50} <0.30) for identifying teratogenic potential, as used in FETAX Validation Study Phase III.3 (Bantle et al., 1999); and

- multiple decision criterion (TI >3.0 plus MCIG/LC_{50} <0.30) for identifying teratogenic potential, as used in a recent study by Fort et al. (2000a).

In the ASTM FETAX Guideline (1991, 1998), a TI value greater than 1.5, an MCIG/LC_{50} ratio less than 0.30, or the presence of severe malformations was considered to be indicative of teratogenic activity. In the FETAX Phase III.3 Validation Study (Bantle et al., 1999), multiple as well as single criteria were used. When multiple criteria were used, test substances were classified as positive when both the TI value was greater than 1.5 and the MCIG/LC_{50} ratio was less than 0.3, equivocal when either but not both criteria were positive, and negative when neither criteria was positive. Where results were classified as equivocal, information on the severity of the observed malformations was used to potentially resolve the classification. In the Fort et al. (2000a) study, single and multiple criteria were used as described in the Phase III.3

Validation Study, except that the critical TI decision value was increased from 1.5 to 3.0; values between 1.5 and 3.0 were considered to be suggestive. In the performance analysis conducted by NICEATM, information on the types and incidence of malformations induced in *X. laevis* embryos were excluded from the evaluations due to the almost complete absence of quantitative data on malformations in the published FETAX literature. For substances that were evaluated multiple times in FETAX, the NICEATM consensus FETAX result was based on a simple weight-of-evidence approach; test substances with an equal number of positive and negative studies were classified as equivocal and were excluded from the performance calculations. In the performance calculations presented herein, the numbers in parenthesis after a percentage value are the number of correct results divided by the total number of test substances considered. Differences in the total number of FETAX test substances considered under apparently identical conditions are due to differences in available data or from the exclusion of test substances with an equivocal classification for that particular decision criteria. Where multiple criteria, equivocal FETAX results were encountered, performance characteristics were calculated excluding equivocal FETAX results, including equivocal FETAX results as positive, or including equivocal FETAX results as negative. The FETAX, laboratory mammal, and human teratogenicity results used in these analyses are summarized in **Appendix 5**.

6.1 Performance Characteristics of FETAX compared to Combined Rat, Mouse, and Rabbit Teratogenicity Test Results

The performance characteristics of FETAX compared to combined rat, mouse, and rabbit teratogenicity test results were determined using three approaches. Performance characteristics were calculated based on the results of FETAX studies conducted without metabolic activation only, conducted with metabolic activation only, and conducted with and without metabolic activation. In the latter 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 consensus positive response was not obtained using either exposure condition. In addition to these analysis conducted using the total FETAX database, the performance characteristics were determined by chemical and product class for

FETAX, with and without metabolic activation, compared to combined rat, mouse, and rabbit teratogenicity test results. For the evaluation of FETAX compared to teratogenicity data obtained from combined rat, mouse, and rabbit studies, a substance was classified as a laboratory mammal teratogen if a positive result was reported for any of the three species. In contrast, test substances positive in one, but not another, species were classified as equivocal by the investigators in the FETAX Phase III.3 Validation Study (Bantle et al., 1999) and in the comparative study conducted by Fort et al. (2000a).

6.1.1 Performance Characteristics of FETAX, Without Metabolic Activation, compared to Combined Rat, Mouse, and Rabbit Teratogenicity Test Results

The performance characteristics of FETAX, without metabolic activation, compared to combined rat, mouse, and rabbit teratogenicity results were calculated using both single and multiple decision criteria (**Table 5**).

<u>Single Decision Criteria:</u> Based on the use of single decision criteria (i.e., TI >1.5; TI >3.0; MCIG/LC_{50} <0.3),

- accuracy varied from 54% (40/74) to 63% (57/90),
- sensitivity from 40% (16/40) to 78% (39/50),
- specificity from 45% (18/40) to 71% (24/34),
- positive predictivity from 62% (23/37 and 16/26) to 64% (39/61),
- negative predictivity from 50% (24/48 and 26/52) to 62% (18/29),
- false positive rate from 29% (10/34) to 55% (22/40), and
- false negative rate from 22% (11/58) to 60% (24/40).

Maximal accuracy and sensitivity, but minimal specificity, occurred when the single decision criterion was a TI value greater than 1.5.

Multiple Decision Criteria: Using the multiple decision criterion (TI >1.5 plus MCIG/LC_{50} <0.3) of Bantle et al. (1999), and when equivocal results were excluded from the evaluation,

- accuracy was 63% (31/49),
- sensitivity was 67% (16/24),
- specificity was 60% (15/25),
- positive predictivity was 62% (16/26),
- negative predictivity was 65% (15/23),
- false positive rate was 40% (10/25), and
- false negative rate was 33% (8/24).

When equivocal results were re-classified as positives and included in the analysis,

- accuracy was 63% (46/73),
- sensitivity was 79% (31/39),
- specificity was 44% (15/34),
- positive predictivity was 62% (31/50),
- negative predictivity was 65% (15/23),
- false positive rate was 56% (19/34), and
- false negative rate was 21% (8/39).

When equivocal responses were re-classified as negatives and included in the analysis,

- accuracy was 55% (40/73),
- sensitivity was 41% (16/39),
- specificity was 71% (24/34),
- positive predictivity was 62% (16/26),
- negative predictivity was 51% (24/47),
- false positive rate was 29% (10/34), and
- false negative rate was 59% (23/39).

Sensitivity was increased when equivocal results were re-classified as positives and included in the analysis, while specificity was increased when equivocal results were re-classified as negatives and included in the analysis. Accuracy was not increased when equivocal calls were re-classified as positives or negatives and included in the analysis.

Using the multiple decision criterion (TI >3.0 plus MCIG/LC_{50} <0.3) of Fort et al. (2000a), when equivocal FETAX results were excluded from the evaluation,

- accuracy was 58% (36/62),
- sensitivity was 47% (14/30),
- specificity was 69% (22/32),
- positive predictivity was 58% (14/24),
- negative predictivity was 58% (22/38),
- false positive rate was 31% (10/32), and
- false negative rate was 53% (16/30).

When equivocal responses were re-classified as positives and included in the analysis,

- accuracy was 61% (44/72),
- sensitivity was 58% (22/38),
- specificity was 65% (22/34),
- positive predictivity was 65% (22/34),
- negative predictivity was 58% (22/38),
- false positive rate was 35% (12/34), and
- false negative rate was 42% (16/38).

When equivocal responses were re-classified as negatives and included in the analysis,

- accuracy was 53% (38/72),
- sensitivity was 37% (14/38),
- specificity was 71% (24/34),
- positive predictivity was 58% (14/24),
- negative predictivity was 50% (24/48),
- false positive rate was 29% (10/34), and
- false negative rate was 63% (24/38).

Accuracy appeared to be optimal when equivocal responses were re-classified as positives and included in the analysis, while sensitivity and specificity were optimal when equivocal responses were re-classified as positives or negative, respectively, and included in the analysis.

The performance characteristics for FETAX, without metabolic activation, compared to combined rat, mouse, and rabbit teratogenicity results were generally not improved by using multiple decision criteria. The use of a single decision criterion based on a TI value greater than 1.5 resulted in increased accuracy and sensitivity over one based on a TI value greater than 3.0.

6.1.2 Performance Characteristics of FETAX, With Metabolic Activation, compared to Combined Rat, Mouse, and Rabbit Teratogenicity Test Results

The performance characteristics of FETAX, with metabolic activation, compared to combined rat, mouse, and rabbit teratogenicity results were calculated using both single and multiple decision criteria (**Table 6**).

Single Decision Criteria: Based on the use of single decision criteria (i.e., TI >1.5; TI >3.0; MCIG/LC_{50} <0.3),

- accuracy varied from 42% (11/26) to 56% (15/27),
- sensitivity ranged from 20% (2/10) to 87% (13/15),
- specificity from 17% (2/12) to 70% (7/10),
- positive predictivity from 40% (2/5) to 57% (13/23),
- negative predictivity from 40% (6/15) to 50% (2/4),
- false positive rate from 30% (3/10) to 83% (10/12), and
- false negative rate from 13% (2/15) to 80% (8/10).

Maximal accuracy and sensitivity occurred when the single decision criterion was a TI value greater than 1.5. However, specificity was highest when an MCIG/LC_{50} ratio less than 0.3 was used as the single decision criterion.

Multiple Decision Criteria: Using multiple decision criterion (TI >1.5 plus MCIG/LC_{50} <0.3) of Bantle et al. (1999), when equivocal results were excluded from the evaluation,

- accuracy was 50% (4/8),
- sensitivity was 67% (2/3),
- specificity was 40% (2/5),
- positive predictivity was 40% (2/5),
- negative predictivity was 67% (2/3),
- false positive rate was 60% (3/5), and
- false negative rate was 33% (1/3).

When equivocal responses were re-classified as positives and included in the analysis,

- accuracy was 55% (11/20),
- sensitivity was 90% (9/10),
- specificity was 20% (2/10),
- positive predictivity was 53% (9/17),
- negative predictivity was 67% (2/3),
- false positive rate was 80% (8/10), and
- false negative rate was 10% (1/10).

When equivocal responses were re-classified as negatives and included in the analysis,

- accuracy was 45% (9/20),
- sensitivity was 20% (2/10),
- specificity was 70% (7/10),
- positive predictivity was 40% (2/5),
- negative predictivity was 47% (7/15),
- false positive rate was 30% (3/10), and
- false negative rate was 80% (8/10).

Accuracy and sensitivity but not specificity were maximal when equivocal calls were re-classified as positives and included in the analysis.

Using the multiple decision criterion (TI >3.0 plus MCIG/LC$_{50}$ <0.3) of Fort et al. (2000a), when equivocal FETAX results were excluded from the evaluation,

- accuracy was 40% (6/15),
- sensitivity was 13% (1/8),
- specificity was 71% (5/7),
- positive predictivity was 33% (1/3),
- negative predictivity was 42% (5/12),
- false positive rate was 29% (2/7), and
- false negative rate was 88% (7/8).

When equivocal calls were re-classified as positives and included in the analysis,

- accuracy was 37% (7/19),
- sensitivity was 22% (2/9),
- specificity was 50% (5/10),
- positive predictivity was 29% (2/7),
- negative predictivity was 42% (5/12),
- false positive rate was 50% (5/10), and
- false negative rate was 78% (7/9).

When equivocal calls were re-classified as negatives and included in the analysis,

- accuracy was 47% (9/19),
- sensitivity was11% (1/9),
- specificity was 80% (8/10),
- positive predictivity was 33% (1/3),
- negative predictivity was 50% (8/16),
- false positive rate was 20% (2/10), and
- false negative rate was 89% (8/9).

Accuracy and specificity were slightly better when equivocal results were classified as positive and included in the analysis.

Accuracy and sensitivity but not specificity were improved compared to combined rat, mouse, and rabbit teratogenicity results when a TI value greater than 1.5 rather than 3.0 was used as the decision criteria. Performance was not generally improved when multiple decision criteria were used.

6.1.3 Performance Characteristics of FETAX, With and Without Metabolic Activation, compared to Combined Rat, Mouse, and Rabbit Teratogenicity Test Results

The overall performance characteristics of FETAX, with and without metabolic activation, compared to the combined rat, mouse, and rabbit teratogenicity results were calculated using both single and multiple decision criteria (**Table 7**).

<u>Single Decision Criteria:</u> Based on the use of a single decision criteria (i.e., TI >1.5; TI >3.0; MCIG/LC$_{50}$ <0.3),

- accuracy varied from 53% (48/90) to 61% (55/90),
- sensitivity from 43% (17/40) to 82% (41/50),
- specificity from 35% (14/40) to 68% (23/34),
- positive predictivity was 61% (17/28, 23/38, and 41/67),
- negative predictivity from 48% (23/46) to 61% (14/23),
- false positive rate from 32% (11/34) to 65% (26/40), and
- false negative rate from 18% (9/50) to 58% (23/40).

Maximal accuracy and sensitivity occurred when the single decision criterion was a TI value greater than 1.5, while maximal specificity occurred when the single decision criterion was an MCIG/LC$_{50}$ ratio of less than 0.3.

Multiple Decision Criteria: Using the multiple decision criterion (TI >1.5 plus MCIG/LC$_{50}$ <0.3) of Bantle et al. (1999), when equivocal results were excluded from the evaluation,

- accuracy was 63% (29/46),
- sensitivity was 74% (17/23),
- specificity was 52% (12/23),
- positive predictivity was 61% (17/28),
- negative predictivity was 67% (12/18),
- false positive rate was 48% (11/23), and
- false negative rate was 26% (6/23).

When equivocal responses were re-classified as positives and included in the analysis,

- accuracy was 62% (45/73),
- sensitivity was 85% (33/39),
- specificity was 35% (12/34),
- positive predictivity was 60% (33/55),
- negative predictivity was 67% (12/18),
- false positive rate was 65% (22/34), and
- false negative rate was 15% (6/39).

When equivocal responses were re-classified as negatives and included in the analysis,

- accuracy was 55% (40/73),
- sensitivity was 44% (17/39),
- specificity was 68% (23/34),
- positive predictivity was 61% (17/28),
- negative predictivity was 51% (23/45),
- false positive rate was 32% (11/34), and
- false negative rate was 56% (22/39).

Accuracy and sensitivity were similar when equivocal response were excluded from the analysis or re-classified as positives and included in the analysis; specificity was optimal when equivocal responses were re-classified as positives and included in the analysis.

Using the multiple decision criterion (TI >3.0 plus MCIG/LC$_{50}$ <0.3) of Fort et al. (2000a), when equivocal FETAX results were excluded from the evaluation,

- accuracy was 58% (35/60),
- sensitivity was 48% (14/29),
- specificity was 68% (21/31),
- positive predictivity was 58% (14/24),
- negative predictivity was 58% (21/36),
- false positive rate was 32% (10/31), and
- false negative rate was 52% (15/29).

When equivocal calls were re-classified as positives and included in the analysis,

- accuracy was 62% (45/73),
- sensitivity was 62% (24/39),
- specificity was 62% (21/34),
- positive predictivity was 65% (24/37),
- negative predictivity was 58% (21/36),
- false positive rate was 38% (13/34), and
- false negative rate was 38% (15/39).

When equivocal calls were re-classified as negatives and included in the analysis,

- accuracy was 52% (38/73),
- sensitivity was 36% (14/39),
- specificity was 71% (24/34),
- positive predictivity was 58% (14/24),
- negative predictivity was 49% (24/49),
- false positive rate was 29% (10/34), and
- false negative rate was 64% (25/39).

With the exception of specificity, performance appeared to be optimal when equivocal calls were re-classified as positives and included in the analysis. In general, a FETAX decision criteria based on the use of a TI value greater than 3.0 was not as accurate as one based on using a TI value greater than 1.5.

Based on an analysis of the performance characteristics for FETAX, with and without metabolic activation, compared to combined rat, mouse, and rabbit teratogenicity results, the use of single decision criterion based on a TI value greater than 1.5 rather than 3.0 appeared to provide the most optimal approach in terms of accuracy and sensitivity. The use of multiple decision criteria did not appreciable improve FETAX performance.

6.1.4 Performance Characteristics of FETAX, With and Without Metabolic Activation, compared to Combined Rat, Mouse, or Rabbit Teratogenicity Test Results by Chemical and Product Class

The most numerically prevalent classes were alcohols (including glycols), amides, amines, halogenated organic compounds, esters, heavy metals and their salts, hydrazides and hydrazines, nitrogen heterocyclic compounds, organic (phenolic and carboxylic) acids, and salts (see **Section 3.3**). The most common product classes tested in FETAX were antimicrobials, chemical synthesis materials, cosmetics, dyes, food additives, fossil fuels, pesticides, pharmaceuticals, photographic chemicals, and polymers (including monomers). The performance characteristics of FETAX, using with and without metabolic activation studies combined were compared to combined rat, mouse, and rabbit teratogenicity results by chemical and product class using single decision criteria (i.e., TI >1.5, TI >3.0, MCIG/LC$_{50}$ <0.3) (**Tables 8**, **9**, and **10**, respectively). Analyses were limited to those chemical and product classes that included a minimum of 15 substances tested in FETAX for which there was also laboratory mammal teratogenicity test results. For comparative purposes, the corresponding performance characteristics when all FETAX data were considered are included in each table.

<u>Amides plus Hydrazides:</u> Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 56% (9/16),
- sensitivity was 78% (7/9),
- specificity was 29% (2/7),
- positive predictivity was 58% (7/12),
- negative predictivity was 50% (2/4),
- false positive rate was 71% (5/7), and
- false negative rate was 22% (2/9).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 44% (7/16),
- sensitivity was 44% (4/9),
- specificity was 43% (3/7),
- positive predictivity was 50% (4/8),
- negative predictivity was 38% (3/8),
- false positive rate was 57% (4/7), and
- false negative rate was 56% (5/9).

Due to the absence of a sufficient database, performance characteristics using a decision criterion based on an MCIG/LC$_{50}$ ratio of less than 0.3 were not determined.

<u>Amines:</u> Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 60% (9/15),
- sensitivity was 89% (8/9),
- specificity was 17% (1/6),
- positive predictivity was 62% (8/13),
- negative predictivity was 50% (1/2),
- false positive rate was 83% (5/6), and
- false negative rate was 11% (1/9).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 60% (9/15),
- sensitivity was 67% (6/9),
- specificity was 50% (3/6),
- positive predictivity was 67% (6/9),
- negative predictivity was 50% (3/6),
- false positive rate was 50% (3/6), and
- false negative rate was 33% (3/9).

Using an MCIG/LC_{50} ratio less than 0.3 as the single decision criterion,

- accuracy was 53% (8/15),
- sensitivity was 56% (5/9),
- specificity was 50% (3/6),
- positive predictivity was 63% (5/8),
- negative predictivity was 43% (3/7),
- false positive rate was 50% (3/6), and
- false negative rate was 44% (4/9).

Nitrogen Heterocyclic Compounds: Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 70% (21/30),
- sensitivity was 80% (16/20),
- specificity was 50% (5/10),
- positive predictivity was 76% (16/21),
- negative predictivity was 56% (5/9),
- false positive rate was 50% (5/10), and
- false negative rate was 20% (4/20).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 53% (16/30),
- sensitivity was 50% (10/20),
- specificity was 60% (6/10),
- positive predictivity was 71% (10/14),
- negative predictivity was 38% (6/16),
- false positive rate was 40% (4/10), and
- false negative rate was 50% (10/20).

Using an MCIG/LC_{50} ratio less than 0.3 as the single decision criterion,

- accuracy was 48% (11/23),
- sensitivity was 53% (8/15),
- specificity was 50% (4/8),
- positive predictivity was 67% (8/12),
- negative predictivity was 36% (4/11),
- false positive rate was 50% (4/8), and
- false negative rate was 47% (7/15).

<u>Organic (Phenolic and Carboxylic) Acids:</u> Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 80% (16/20),
- sensitivity was 86% (12/14),
- specificity was 67% (4/6),
- positive predictivity was 86% (12/14),
- negative predictivity was 67% (4/6),
- false positive rate was 33% (2/6), and
- false negative rate was 14% (2/14).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 60% (12/20),
- sensitivity was 43% (6/14),
- specificity was 100% (6/6),
- positive predictivity was 100% (6/6),
- negative predictivity was 43% (6/14),
- false positive rate was 0% (0/6), and
- false negative rate was 57% (8/14).

Using an MCIG/LC_{50} ratio less than 0.3 as the single decision criterion,

- accuracy was 60% (9/15),
- sensitivity was 40% (4/10),
- specificity was 100% (5/5),
- positive predictivity was 100% (4/4),
- negative predictivity was 45% (5/11),
- false positive rate was 0% (0/5), and
- false negative rate was 60% (6/10).

<u>Pharmaceuticals:</u> Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 54% (21/39),
- sensitivity was 82% (18/22),
- specificity was 18% (3/17),
- positive predictivity was 56% (18/32),
- negative predictivity was 43% (3/7),
- false positive rate was 82% (14/17), and
- false negative rate was 18% (4/22).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 50% (19/38),
- sensitivity was 43% (9/21),
- specificity was 59% (10/17),
- positive predictivity was 56% (9/16),
- negative predictivity was 45% (10/22),
- false positive rate was 41% (7/17), and
- false negative rate was 57% (12/21).

Using an MCIG/LC$_{50}$ ratio less than 0.3 as the single decision criterion,

- accuracy was 53% (18/34),
- sensitivity was 47% (9/19),
- specificity was 60% (9/15),
- positive predictivity was 60% (9/15),
- negative predictivity was 47% (9/19),
- false positive rate was 40% (6/15), and
- false negative rate was 53% (10/19).

Due to the limited FETAX database, only five chemical classes and one product class were evaluated for performance characteristics compared to the combined rat, mouse, and rabbit teratogenicity test results. Among the chemical and product classes evaluated, a decision criterion based on a TI value greater than 1.5 generally provided greater accuracy and sensitivity, but less specificity, than one based on either on a TI value greater than 3.0 or on an MCIG/LC$_{50}$ ratio less than 0.3. In general, the accuracy of FETAX compared to laboratory mammal teratogenicity test results was somewhat improved for nitrogen heterocyclic compounds, and phenolic and carboxylic acids. Performance characteristics for the other chemical classes and the single product class evaluated were not appreciable different from the performance of FETAX compared to the total database.

6.2 Performance Characteristics of FETAX compared to Individual Rat, Mouse, or Rabbit Species Teratogenicity Test Results

The performance characteristics of FETAX compared to individual rat, mouse, and rabbit species teratogenicity test results were calculated using single TI decision criteria (TI >1.5, TI >3.0) only. Comparisons using other decision criteria (i.e., MCIG/LC$_{50}$ <0.30, various multiple decision criteria) were not conducted because of the inadequate numbers of comparisons available for the analysis. In this analysis, performance characteristics were determined based on the results of FETAX studies conducted without metabolic activation only, conducted with metabolic activation only, and conducted with and without metabolic activation. Performance characteristics based on chemical and product class for FETAX compared to individual rat, mouse, and rabbit species teratogenicity test results were not determined due to the paucity of the data. For the evaluation of FETAX compared to teratogenicity data obtained from combined rat, mouse, and rabbit studies, a substance was classified as a laboratory mammal teratogen if a positive result was reported for any of the three species.

6.2.1 Performance Characteristics of FETAX, Without Metabolic Activation, compared to Individual Rat, Mouse, or Rabbit Species Teratogenicity Test Results

The performance characteristics of FETAX, without metabolic activation, compared to rat, mouse, or rabbit teratogenicity results, individually, are provided in **Table 11**.

<u>FETAX versus Rat:</u> Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 61% (46/75),
- sensitivity was 77% (30/39),
- specificity was 44% (16/36),
- positive predictivity was 60% (30/50),
- negative predictivity was 64% (16/25),
- false positive rate was 56% (20/36), and
- false negative rate was 23% (9/39).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 51% (37/73),
- sensitivity was 43% (16/37),
- specificity was 58% (21/36),
- positive predictivity was 52% (16/31),
- negative predictivity was 50% (21/42),
- false positive rate was 42% (15/36), and
- false negative rate was 57% (21/37).

<u>FETAX versus Mouse:</u> Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 68% (45/66),
- sensitivity was 83% (33/40),
- specificity was 46% (12/26),
- positive predictivity was 70% (33/47),
- negative predictivity was 63% (12/19),
- false positive rate was 54% (14/26), and
- false negative rate was 18% (7/40).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 57% (37/65),
- sensitivity was 51% (20/39),
- specificity was 65% (17/26),
- positive predictivity was 69% (20/29),
- negative predictivity was 47% (17/36),
- false positive rate was 35% (9/26), and
- false negative rate was 49% (19/39).

<u>FETAX versus Rabbit:</u> Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 47% (16/34),
- sensitivity was 64% (9/14),
- specificity was 35% (7/20),
- positive predictivity was 41% (9/22),
- negative predictivity was 58% (7/12),
- false positive rate was 65% (13/20), and
- false negative rate was 36% (5/14).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 48% (16/33),
- sensitivity was 31% (4/13),
- specificity was 60% (12/20),
- positive predictivity was 33% (4/12),
- negative predictivity was 57% (12/21),
- false positive rate was 40% (8/20), and
- false negative rate was 69% (9/13).

Using either TI decision criteria value, the performance characteristics of FETAX, without metabolic activation, compared to teratogenicity data for rats and mice were quite similar, while that for rabbits appeared to be reduced. Furthermore, the performance characteristics compared to rats and mice were not different from the corresponding performance characteristics based on combined rat, mouse, and rabbit teratogenicity data (**Table 5**). Comparing the performance characteristics for each species as a function of the TI decision criterion value, increased accuracy and sensitivity, but decreased specificity, was associated with the use of a TI value greater than 1.5 rather than 3.0.

6.2.2 Performance Characteristics of FETAX, With Metabolic Activation, compared to Individual Rat, Mouse, or Rabbit Species Teratogenicity Test Results

The performance characteristics of FETAX, with metabolic activation, compared to rat, mouse, or rabbit teratogenicity results, individually, are shown in **Table 12**.

<u>FETAX versus Rat:</u> Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 65% (15/23),
- sensitivity was 100% (11/11),
- specificity was 33% (4/12),
- positive predictivity was 58% (11/19),
- negative predictivity was 100% (4/4),
- false positive rate was 67% (8/12), and
- false negative rate was 0% (0/11).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 36% (8/22),
- sensitivity was 30% (3/10),
- specificity was 42% (5/12),
- positive predictivity was 42% (5/12),
- negative predictivity was 58% (7/12),
- false positive rate was 70% (7/10), and
- false negative rate was 30% (3/10).

<u>FETAX versus Mouse:</u> Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 52% (11/21),
- sensitivity was 85% (11/13),
- specificity was 0% (0/8),
- positive predictivity was 58% (11/19),
- negative predictivity was 0% (0/2),
- false positive rate was 100% (8/8), and
- false negative rate was 15% (2/13).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 40% (8/20),
- sensitivity was 42% (5/12),
- specificity was 38% (3/8),
- positive predictivity was 50% (5/10),
- negative predictivity was 30% (3/10),
- false positive rate was 63% (5/8), and
- false negative rate was 58% (7/12).

FETAX versus Rabbit: Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 53% (8/15),
- sensitivity was 100% (7/7),
- specificity was 13% (1/8),
- positive predictivity was 50% (7/14),
- negative predictivity was 100% (1/1),
- false positive rate was 88% (7/8), and
- false negative rate was 0% (0/7).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 50% (7/14),
- sensitivity was 33% (2/6),
- specificity was 63% (5/8),
- positive predictivity was 40% (3/5),
- negative predictivity was 56% (5/9),
- false positive rate was 38% (3/8), and
- false negative rate was 67% (4/6).

Using either TI decision criterion value, the performance characteristics of FETAX, with metabolic activation, compared to teratogenicity data for all three-laboratory species appeared to be similar. These FETAX performance characteristics were not very different from the performance characteristics based on combined rat, mouse, and rabbit teratogenicity data. Comparing the performance characteristics for each species as a function of the TI decision criterion value, increased accuracy and sensitivity, but decreased specificity, was associated with the use of a TI value greater than 1.5 rather than 3.0. However, the validity of these conclusions is suspect because of the very limited number of substances tested with metabolic activation.

6.2.3 Performance Characteristics of FETAX, With and Without Metabolic Activation, compared to Individual Rat, Mouse, or Rabbit Species Teratogenicity Test Results

The performance characteristics of FETAX, with and without metabolic activation, compared to rat, mouse, or rabbit teratogenicity results, individually, are presented in **Table 13**.

FETAX versus Rat: Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 61% (46/75),
- sensitivity was 82% (32/39),
- specificity was 39% (14/36),
- positive predictivity was 59% (32/54),
- negative predictivity was 67% (14/21),
- false positive rate was 61% (22/36), and
- false negative rate was 18% (7/39).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 49% (36/74),
- sensitivity was 42% (16/38),
- specificity was 56% (20/36),
- positive predictivity was 50% (16/32),
- negative predictivity was 48% (20/42),
- false positive rate was 44% (16/36), and
- false negative rate was 58% (22/38).

FETAX versus Mouse: Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 64% (42/66),
- sensitivity was 85% (34/40),
- specificity was 31% (8/26),
- positive predictivity was 65% (34/52),
- negative predictivity was 57% (8/14),
- false positive rate was 69% (18/26), and
- false negative rate was 15% (8/40).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 55% (36/66),
- sensitivity was 50% (20/40),
- specificity was 62% (16/26),
- positive predictivity was 67% (20/30),
- negative predictivity was 44% (16/36),
- false positive rate was 38% (10/26), and
- false negative rate was 50% (20/40).

FETAX versus Rabbit: Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 44% (15/34),
- sensitivity was 71% (10/14),
- specificity was 25% (5/20),
- positive predictivity was 40% (10/25),
- negative predictivity was 56% (5/9),
- false positive rate was 75% (15/20), and
- false negative rate was 29% (4/14).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 47% (16/34),
- sensitivity was 29% (4/14),
- specificity was 60% (12/20),
- positive predictivity was 33% (4/12),
- negative predictivity was 55% (12/22),
- false positive rate was 40% (8/20), and
- false negative rate was 71% (10/14).

Using either TI decision criterion value, the performance characteristics of FETAX, with and without metabolic activation, compared to teratogenicity data for rats, mice, and rabbits appeared to be similar. These FETAX performance characteristics were not very different from the performance characteristics based on combined rat, mouse, and rabbit teratogenicity data. Comparing the performance characteristics for each species as a function of the TI decision criteria value, increased accuracy and sensitivity, but decreased specificity, was associated with the use of a TI value greater than 1.5 rather than 3.0.

6.3 Performance Characteristics of FETAX compared to Human Teratogenicity Study Results

The performance characteristics of FETAX compared to human teratogenicity study results were determined also using three approaches. Performance characteristics were calculated based on the results of FETAX studies conducted without metabolic activation only, conducted with metabolic activation only, and conducted with and without metabolic activation. In addition to these analysis conducted using the total FETAX database, the performance characteristics were determined, where feasible, by chemical and product class for FETAX, with and without metabolic activation combined, compared to human teratogenicity study results.

6.3.1 Performance Characteristics of FETAX, Without Metabolic Activation, compared to Human Teratogenicity Study Results

The performance characteristics of FETAX, without metabolic activation, compared to human teratogenicity study results were calculated using both single and multiple decision criteria (**Table 14**).

Single Decision Criteria: Based on the use of single decision criteria (i.e., TI >1.5; TI >3.0; MCIG/LC$_{50}$ <0.3),

- accuracy varied from 48% (15/31) to 63% (17/27 and 19/30),
- sensitivity from 47% (7/15) to 67% (10/15),
- specificity from 31% (5/16) to 80% (12/15),
- positive predictivity from 48% (10/21) to 70% (7/10),
- negative predictivity from 50% (5/10) to 65% (11/17),
- false positive rate from 20% (3/15) to 69% (11/16), and
- false negative rate from 33% (5/15) to 53% (8/15).

Maximal accuracy and specificity occurred when the single decision criterion was an MCIG/LC$_{50}$ ratio less than 0.3. Maximal sensitivity occurred when the single decision criterion was a TI value greater than 1.5.

Multiple Decision Criteria: Using the multiple decision criterion (TI >1.5 plus MCIG/LC$_{50}$ <0.3) of Bantle et al. (1999), when equivocal results were excluded from the evaluation,

- accuracy was 61% (11/18),
- sensitivity was 67% (6/9),
- specificity was 56% (5/9),
- positive predictivity was 60% (6/10),
- negative predictivity was 63% (5/8),
- false positive rate was 44% (4/9), and
- false negative rate was 33% (3/9).

When equivocal responses were re-classified as positives and included in the analysis,

- accuracy was 52% (14/27),
- sensitivity was 75% (9/12),
- specificity was 33% (5/15),
- positive predictivity was 47% (9/19),
- negative predictivity was 63% (5/8),
- false positive rate was 67% (10/15), and
- false negative rate was 25% (3/12).

When equivocal responses were re-classified as negatives and included in the analysis,

- accuracy was 63% (17/27),
- sensitivity was 50% (6/12),
- specificity was 73% (11/15),
- positive predictivity was 60% (6/10),
- negative predictivity was 65% (11/17),
- false positive rate was 27% (4/15), and
- false negative rate was 50% (6/12).

Maximal accuracy occurred when equivocal results were excluded from analysis or were re-classified as negative and included in the analysis. Maximal sensitivity occurred when equivocal results were re-classified as positive and included in the analysis. Maximal specificity occurred when equivocal results were re-classified as negative and included in the analysis

Using the multiple decision criterion (TI >3.0 plus MCIG/LC_{50} <0.3) of Fort et al. (2000a), when equivocal FETAX results were excluded from the evaluation,

- accuracy was 73% (16/22),
- sensitivity was 60% (6/10),
- specificity was 83% (10/12),
- positive predictivity was 75% (6/8),
- negative predictivity was 71% (10/14),
- false positive rate was 17% (2/12), and
- false negative rate was 40% (4/10).

When equivocal calls were re-classified as positives and included in the analysis,

- accuracy was 65% (17/26),
- sensitivity was 58% (7/12),
- specificity was 71% (10/14),
- positive predictivity was 64% (7/11),
- negative predictivity was 67% (10/15),
- false positive rate was 29% (4/14), and
- false negative rate was 42% (5/12).

When equivocal calls were re-classified as negatives and included in the analysis,

- accuracy was 69% (18/26),
- sensitivity was 50% (6/12),
- specificity was 86% (12/14),
- positive predictivity was 75% (6/8),
- negative predictivity was 67% (12/18),
- false positive rate was 14% (2/14), and
- false negative rate was 50% (6/12).

Maximal accuracy occurred when equivocal results were excluded from analysis or were re-classified as negative and included in the analysis. Maximal sensitivity occurred when equivocal results were excluded or were re-classified as positive and included in the analysis. Maximal specificity occurred when equivocal results were excluded or were re-classified as negative and included in the analysis

The performance characteristics of FETAX, without metabolic activation, compared to human teratogenicity study results were maximal and similar when either a TI value greater than 3.0 or a MCIG/LC_{50} ratio less than 0.3 were used as the single decision criterion. In general, the use of multiple criteria did not increase the performance of FETAX for predicting human teratogenicity.

6.3.2 Performance Characteristics of FETAX, With Metabolic Activation, compared to Human Teratogenicity Study Results

The performance characteristics of FETAX, with metabolic activation, compared to human teratogenicity results were calculated using both single and multiple decision criteria (**Table 15**). The validity of this analysis is questionable considering the very limited number of substances tested with metabolic activation in FETAX for which there were relevant human data also.

<u>Single Decision Criteria:</u> Based on the use of single decision criterion (i.e., TI >1.5; TI >3.0; MCIG/LC_{50} <0.3),

- accuracy varied from 40% (4/10) to 100% (8/8),
- sensitivity from 50% (1/2) to 100% (3/3),
- specificity from 14% (1/7) to 100% (5/5),
- positive predictivity from 33% (3/9) to 100% (3/3),
- negative predictivity from 86% (6/7) to 100% (1/1 and 5/5),
- false positive rate from 0% (0/5) to 86% (6/7), and
- false negative rate from 0% (0/3) to 50% (1/2).

Maximal accuracy, sensitivity, and specificity occurred when the single decision criterion was an MCIG/LC$_{50}$ ratio less than 0.3.

Multiple Decision Criteria: Using the multiple decision criterion (TI >1.5 plus MCIG/LC$_{50}$ <0.3) of Bantle et al. (1999), when equivocal results were excluded from the evaluation,

- accuracy was 100% (4/4),
- sensitivity was 100% (3/3),
- specificity was 100% (1/1),
- positive predictivity was 100% (3/3),
- negative predictivity was 100% (1/1),
- false positive rate was 0% (0/1), and
- false negative rate was 0% (0/3).

When equivocal responses were re-classified as positives and included in the analysis,

- accuracy was 50% (4/8),
- sensitivity was 100% (3/3),
- specificity was 20% (1/5),
- positive predictivity was 43% (3/7),
- negative predictivity was 100% (1/1),
- false positive rate was 80% (4/5), and
- false negative rate was 0% (0/3).

When equivocal responses were re-classified as negatives and included in the analysis,

- accuracy was 100% (8/8),
- sensitivity was 100% (3/3),
- specificity was 100% (5/5),
- positive predictivity was 100% (3/3),
- negative predictivity was 100% (5/5),
- false positive rate was 0% (0/5), and
- false negative rate was 0% (0/3).

Maximal performance characteristics occurred when equivocal results were excluded from analysis or were re-classified as negative results and included in the analysis.

Using the multiple decision criterion (TI >3.0 plus MCIG/LC$_{50}$ <0.3) of Fort et al. (2000a), when equivocal FETAX results were excluded from the evaluation,

- accuracy was 100% (6/6),
- sensitivity was 100% (1/1),
- specificity was 100% (5/5),
- positive predictivity was 100% (1/1),
- negative predictivity was 100% (5/5),
- false positive rate was 0% (0/5), and
- false negative rate was 0% (0/1).

When equivocal calls were re-classified as positives and included in the analysis,

- accuracy was 100% (7/7),
- sensitivity was 100% (2/2),
- specificity was 100% (5/5),
- positive predictivity was 100% (2/2),
- negative predictivity was 100% (5/5),
- false positive rate was 0% (0/5), and
- false negative rate was 0% (0/2).

When equivocal calls were re-classified as negatives and included in the analysis,

- accuracy was 86% (6/7),
- sensitivity was 50% (1/2),
- specificity was 100% (5/5),
- positive predictivity was 100% (1/1),
- negative predictivity was 83% (5/6),
- false positive rate was 0% (0/5), and
- false negative rate was 50% (1/2).

Maximal performance characteristics occurred when equivocal results were excluded from analysis or were re-classified as negative results and included in the analysis.

The performance characteristics of FETAX, with metabolic activation, compared to human teratogenicity study results were maximal and similar when an MCIG/LC_{50} ratio less than 0.3 were used as the single decision criterion. In general, the use of multiple criteria did not increase the performance of FETAX for predicting human teratogenicity.

6.3.3 Performance of FETAX, With and Without Metabolic Activation, compared to Human Teratogenicity Study Results

The performance characteristics of FETAX, with and without metabolic activation, compared to human teratogenicity results were calculated using both single and multiple decision criteria (**Table 16**).

Single Decision Criteria: Based on the use of single decision criterion (i.e., TI >1.5; TI >3.0; MCIG/LC$_{50}$ <0.3),

- accuracy varied from 48% (15/31) to 70% (19/27),
- sensitivity from 47% (7/15) to 80% (12/15),
- specificity from 19% (3/16) to 81% (13/16),
- positive predictivity from 48% (12/25) to 70% (7/10),
- negative predictivity from 50% (3/6) to 73% (11/15),
- false positive rate from 19% (3/16) to 81% (13/16), and
- false negative rate from 20% (3/15) to 53% (8/15).

Maximal accuracy and specificity occurred when the single decision criterion was an MCIG/LC$_{50}$ ratio less than 0.3. Maximal sensitivity occurred when the single decision criterion was a TI value greater than 1.5.

Multiple Decision Criteria: Using the multiple decision criterion (TI >1.5 plus MCIG/LC$_{50}$ <0.3) of Bantle et al. (1999), when equivocal results were excluded from the evaluation,

- accuracy was 69% (11/16),
- sensitivity was 89% (8/9),
- specificity was 43% (3/7),
- positive predictivity was 67% (8/12),
- negative predictivity was 75% (3/4),
- false positive rate was 57% (4/7), and
- false negative rate was11% (1/9).

When equivocal responses were re-classified as positives and included in the analysis,

- accuracy was 52% (14/27),
- sensitivity was 92% (11/12),
- specificity was 20% (3/15),
- positive predictivity was 48% (11/23),
- negative predictivity was 75% (3/4),
- false positive rate was 80% (12/15), and
- false negative rate was 8% (1/12).

When equivocal responses were re-classified as negatives and included in the analysis,

- accuracy was 70% (19/27),
- sensitivity was 67% (8/13),
- specificity was 73% (11/15),
- positive predictivity was 67% (8/12),
- negative predictivity was 73% (11/15),
- false positive rate was 27% (4/15), and
- false negative rate was 33% (4/12).

Maximal accuracy occurred when equivocal results were excluded from analysis or were re-classified as negative and included in the analysis. Maximal sensitivity occurred when equivocal results were excluded from analysis or were re-classified as positive and included in the analysis. Maximal specificity occurred when equivocal results were re-classified as negative and included in the analysis

Using the multiple decision criterion (TI >3.0 plus MCIG/LC$_{50}$ <0.3) of Fort et al. (2000a), when equivocal FETAX results were excluded from the evaluation,

- accuracy was 76% (16/21),
- sensitivity was 67% (6/9),
- specificity was 83% (10/12),
- positive predictivity was 75% (6/8),
- negative predictivity was 77% (10/13),
- false positive rate was 17% (2/12), and
- false negative rate was 33% (3/9).

When equivocal calls were re-classified as positives and included in the analysis,

- accuracy was 70% (19/27),
- sensitivity was 75% (9/12),
- specificity was 67% (10/15),
- positive predictivity was 64% (9/14),
- negative predictivity was 77% (10/13),
- false positive rate was 33% (5/15), and
- false negative rate was 25% (3/12).

When equivocal calls were re-classified as negatives and included in the analysis,

- accuracy was 70% (19/27),
- sensitivity was 50% (6/12),
- specificity was 87% (13/15),
- positive predictivity was 75% (6/8),
- negative predictivity was 68% (13/19),
- false positive rate was 13% (2/15), and
- false negative rate was 50% (6/12).

Maximal accuracy occurred when equivocal results were excluded from analysis. Maximal sensitivity occurred when equivocal results were re-classified as positive and included in the analysis. Maximal specificity occurred when equivocal results were re-classified as negative and included in the analysis.

In general, among single decision criteria, the use of a criterion based on an MCIG/LC_{50} ratio less than 0.3 resulted in the greatest accuracy and specificity, while a TI value greater than 1.5 resulted in the greatest sensitivity for identifying human teratogenicity responses. The use of multiple decision criteria did not have an appreciable effect on the performance characteristics of FETAX.

6.3.4 Performance Characteristics of FETAX, With and Without Metabolic Activation, compared to Human Teratogenicity Study Results by Chemical and Product Class

The most numerically prevalent chemical classes were alcohols (including glycols); amides; amines; halogenated organic compounds; esters; heavy metals and their salts; hydrazides and hydrazines; nitrogen heterocyclic compounds; organic (phenolic and carboxylic) acids; and salts (see **Section 3.3**). The most common product classes tested in FETAX were antimicrobials, chemical synthesis, cosmetics, dyes, food additives, fossil fuels, pesticides, pharmaceuticals, photographic chemicals, and polymers (including monomers). The performance characteristics of FETAX, with and without metabolic activation, compared to human teratogenicity study results were determined by chemical and product class using single decision criteria (i.e., TI >1.5, TI >3.0, MCIG/LC_{50} <0.3) only (**Table 17**). Analyses were limited to those chemical and product classes that included a minimum of 15 substances tested in FETAX for which there was also human teratogenicity study results. For comparative purposes, the corresponding performance characteristics when all FETAX data were considered are included in **Table 17**.

<u>Nitrogen Heterocyclic Compounds:</u> Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 56% (9/16),
- sensitivity was 78% (7/9),
- specificity was 29% (2/7),
- positive predictivity was 58% (7/12),
- negative predictivity was 50% (2/4),
- false positive rate was 71% (5/7), and
- false negative rate was 22% (2/9).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 69% (11/16),
- sensitivity was 56% (5/9),
- specificity was 86% (6/7),
- positive predictivity was 83% (5/6),
- negative predictivity was 60% (6/10),
- false positive rate was 14% (1/7), and
- false negative rate was 44% (4/9).

Due to the absence of a sufficient database, performance characteristics using a decision criterion based on an MCIG/LC_{50} ratio of less than 0.3 were not determined.

<u>Pharmaceuticals:</u> Using a TI value greater than 1.5 as the single decision criterion,

- accuracy was 43% (9/21),
- sensitivity was 80% (8/10),
- specificity was 9% (1/11),
- positive predictivity was 44% (8/18),
- negative predictivity was 33% (1/3),
- false positive rate was 91% (10/11), and
- false negative rate was 20% (2/10).

Using a TI value greater than 3.0 as the single decision criterion,

- accuracy was 67% (14/21),
- sensitivity was 50% (5/10),
- specificity was 82% (9/11),
- positive predictivity was 71% (5/7),
- negative predictivity was 64% (9/14),
- false positive rate was 18% (2/11), and
- false negative rate was 50% (5/10).

Using an MCIG/LC_{50} ratio less than 0.3 as the single decision criterion,

- accuracy was 68% (13/19),
- sensitivity was 67% (6/9),
- specificity was 70% (7/10),
- positive predictivity was 67% (6/9),
- negative predictivity was 70% (7/10),
- false positive rate was 30% (3/10), and
- false negative rate was 33% (3/9).

Due to the limited FETAX database with corresponding human teratogenicity study results, only one chemical class and one product class were evaluated for performance characteristics compared to human teratogenicity study results. The performance characteristics of FETAX compared to human teratogenicity study results were not improved for these chemical and product classes compared to that for the total database.

6.4 Performance Characteristics of Rat, Mouse, and/or Rabbit Teratogenicity Test Results compared to Human Teratogenicity Study Results

The performance characteristics for combined rat, mouse, and rabbit teratogenicity results, as well as for each individual species, compared to human teratogenicity responses were calculated (for comparative purposes compared to FETAX with and/or without metabolic activation, these data are presented in **Tables 14** through **16**).

For combined laboratory mammal results,

- accuracy was 63% (19/30),
- sensitivity was 71% (10/14),
- specificity was 56% (9/16),
- positive predictivity was 59% (10/17),
- negative predictivity was 69% (9/13),
- false positive rate was 44% (7/16), and
- false negative rate was 29% (4/14).

When the performance characteristics for rat compared to human teratogenicity results only were determined,

- accuracy was 65% (17/26),
- sensitivity was 75% (9/12),
- specificity was 57% (8/14),
- positive predictivity was 60% (9/15),
- negative predictivity was 73% (8/11),
- false positive rate was 43% (6/14), and
- false negative rate was 25% (3/12).

When the performance characteristics for mouse compared to human teratogenicity results only were calculated,

- accuracy was 68% (19/28),
- sensitivity was 71% (10/14),
- specificity was 64% (9/14),
- positive predictivity was 67% (10/15),
- negative predictivity was 69% (9/13),
- false positive rate was 36% (5/14), and
- false negative rate was 29% (4/14)

When the performance characteristics for rabbit compared to human teratogenicity results only were calculated,

- accuracy was 53% (8/15),
- sensitivity was 50% (4/8),
- specificity was 57% (4/7),
- positive predictivity was 57% (4/7),
- negative predictivity was 50% (4/8),
- false positive rate was 43% (3/7), and
- false negative rate was 50% (4/8).

Maximal performance were obtained using rat, mouse, or combined laboratory mammal teratogenicity data. Performance characteristics for rabbit teratogenicity data were generally reduced compared to that for the other two species, but may reflect the limited database available for substances also tested in FETAX. The rat, mouse, or combined laboratory mammal performance characteristics compared to human teratogenicity study results appeared to be not much improved compared to that calculated for FETAX, with and without metabolic activation, using the MCIG/LC$_{50}$ ratio of less than 0.3 as the single decision criterion.

6.5 FETAX Results Discordant with Reference Laboratory Mammal or Human Teratogenicity Study Results

The substances tested in FETAX that are discordant with the teratogenicity results obtained for laboratory mammals and humans are listed in **Table 18**. For the purpose of collecting these data, a substance was classified as positive in FETAX based on the most commonly used decision criterion (i.e., TI >1.5) only. Furthermore, if tested with and without metabolic activation, a substance was classified as a FETAX positive if a positive response was obtained using either exposure condition, and as a FETAX negative only if negative results were obtained with and without metabolic activation. Classification of a laboratory mammal teratogenicity result as positive was based on the presence of at least one positive rat, mouse, and/or rabbit study.

Using these classification parameters:

- Twenty-four substances were discordant with laboratory mammal teratogenicity results (seven substances were FETAX positive and laboratory mammal negative; seventeen substances were FETAX negative and laboratory mammal positive);

- Eight substances were concordant with laboratory mammal teratogenicity data but discordant with human teratogenicity results (one substance was FETAX/laboratory mammal negative and human positive; seven substances were FETAX/laboratory mammal positive and human negative); and

- Eight substances were discordant with laboratory mammal and human teratogenicity results (two substances were FETAX negative and laboratory mammal/human positive; six substances were FETAX positive and laboratory mammal/human negative);

- three substances were discordant with laboratory mammal but concordant with human teratogenicity results (no substance was a FETAX/human negative and laboratory mammal positive; three substances were FETAX/human positive and laboratory mammal negative).

Table 18. FETAX Results Discordant with Reference Laboratory Mammal Data and/or Human Teratogenicity Results*

Substance	TI>1.5	Laboratory Mammal	Human
Substances Discordant with Laboratory Mammal Teratogenicity Results			
alpha.-Chaconine	-	+	
Actinomycin D	-	+	
Cycloheximide	-	+	
Dichloroacetic acid	-	+	
Formamide	-	+	
Glycerol formal	-	+	
N-Nitrosodimethylamine	-	+	
2-Butyne-1,4-diol	+	-	
Acrylamide	+	-	
Amaranth	+	-	
Atrazine	+	-	
Benzo[a]pyrene	+	-	
Cobalt chloride	+	-	
Copper chloride	+	-	
Cotinine	+	-	
Diethylene glycol	+	-	
Glycerol	+	-	
Hydrazine	+	-	
Monosodium glutamate	+	-	
Permethrin	+	-	
Propylene glycol	+	-	
Sodium acetate	+	-	
Sodium selenate	+	-	
Trichloroethylene	+	-	
Substances Concordant with Laboratory Mammal but Discordant with Human Teratogenicity Results			
p-Hydroxydilantin	-	-	+
Boric Acid	+	+	-
Cadmium chloride	+	+	-
Caffeine	+	+	-
Dichloroacetate	+	+	-
Phenytoin	+	+	-
Theophylline	+	+	-
Trichloroacetic acid	+	+	-

Substances Discordant with Laboratory Mammal and Human Teratogenicity Results			
Ethanol (L)	-	+	+
m-Hydroxydilantin	-	+	+
Acetaminophen	+	-	-
Acetone	+	-	-
Ascorbic acid	+	-	-
Diphenhydramine hydrochloride	+	-	-
Doxylamine succinate	+	-	-
Furazolidone	+	-	-
Substances Discordant with Laboratory Mammal but Concordant with Human Teratogenicity Results			
4-Hydroxycoumarin	+	-	+
Coumarin	+	-	+
Isoniazid	+	-	+

*If tested with and without metabolic activation, a substance was classified as a FETAX positive if a positive response was obtained using either exposure condition, and as a FETAX negative only if negative results were obtained with and without metabolic activation. Classification of a laboratory mammal teratogenicity result as positive was based on the presence of at least one positive rat, mouse, and/or rabbit study.
The symbols "-" and "+" signify a negative and positive response, respectively.

The bases for the discordant results (e.g., mechanistic, the use of a less than optimal decision criteria) between FETAX and the combined laboratory mammal and/or the human teratogenicity results remains to be determined.

6.6 NICEATM Analysis of FETAX Decision Criteria

The use of a single decision criterion based on a TI value greater than 1.5 appeared to provide the optimal approach in terms of accuracy and sensitivity for predicting combined laboratory mammal teratogenicity data. The use of a TI value greater than 3.0 as the single decision criterion resulted in increased specificity, but decreased sensitivity. The use of multiple decision criteria had no appreciable effect on accuracy or sensitivity but increased specificity when equivocal results were excluded from the analysis. Using either TI decision criterion value, the performance characteristics of FETAX, with and without metabolic activation, compared to teratogenicity data for rats, mice, or rabbits individually appeared to be similar. These FETAX performance characteristics were not very different from the performance characteristics based on combined rat, mouse, and rabbit teratogenicity data. Comparing the performance characteristics for each species as a function of the TI value, increased accuracy and sensitivity, but decreased specificity, was associated with the use of a decision criterion based on a TI value greater than 1.5 rather than 3.0.

In general, the use of a single decision criterion based on an MCIG/LC_{50} ratio less than 0.3 appeared to provide the optimal approach for predicting human teratogenicity data. The use of multiple decision criterion increased sensitivity when equivocal results were classified as positive, and increased specificity when equivocal results were classified as negative.

Maximal performance characteristics for laboratory mammal data compared to human results were obtained using rat, mouse, or combined laboratory mammal teratogenicity data. Performance characteristics for rabbit teratogenicity data were generally poor compared to that for the other two species. In general, the rat, mouse, or combined laboratory mammal performance characteristics compared to human teratogenicity results appeared to be similar to that calculated for FETAX using the MCIG/LC_{50} ratio less than 0.3 as the single decision criterion. However, the database for this comparison was limited to substances tested in FETAX only.

Limiting the analysis of the performance characteristics to substances for which there were, in each case, FETAX, laboratory mammal, and human results does not alter these conclusions.

6.6.1 Evaluation for the Optimal FETAX Single Decision Criterion

In an attempt to identify the optimal TI value or MCIG/LC_{50} ratio to use as a single decision criterion in evaluating FETAX data, NICEATM assessed the relationship between different TI values or MCIG/LC_{50} ratios and performance characteristics. Accuracy, sensitivity, and specificity were calculated for FETAX, without metabolic activation, compared to combined laboratory mammal (rat, mouse, and rabbit) or human teratogenicity results. In conducting these analysis, the median TI value or median MCIG/LC_{50} ratio was used for test substances where multiple studies had been conducted. The use of a median value may result in performance characteristics for

FETAX that are different from those calculated in **Sections 6.1** through **6.3**. FETAX performance characteristics in those sections were based on a weight-of-evidence approach that only evaluated whether a TI value or an MCIG/LC_{50} ratio was above or below the selected decision point.

The optimal TI value or MCIG/LC_{50} ratio to use as a single decision criterion for identifying teratogens in FETAX depends on whether the assay is to be used as a replacement for an existing *in vivo* laboratory mammal assay, or as a screen to identify substances expected to be positive in laboratory mammal assays or in humans. If used as a replacement assay, accuracy (i.e., the ability to correctly identify both positive and negative teratogens) is probably the most important performance characteristic on which to evaluate the data. In contrast, for screening purposes, sensitivity (i.e., the proportion of all positive substances that are correctly identified as positive; sensitivity is also the inverse of the false negative rate) may be the performance characteristic of primary interest.

6.6.1.1 Combined Rat, Mouse, and Rabbit Teratogenicity Test Results

Optimal TI Value: The accuracy, sensitivity, and specificity of FETAX, without metabolic activation, based on using TI values ranging from 0 to 49 as the single decision criterion, compared to combined rat, mouse, and rabbit teratogenicity results are presented graphically in **Figure 1**.

Maximal accuracy for FETAX, without metabolic activation, compared to combined rat, mouse, and rabbit teratogenicity test results was ~60% at TI values between 0 and ~2.1. At TI values between 2.1 and ~22, accuracy steady decreased to ~40% and then remained relatively constant at this value as the TI increased. Sensitivity was 85% at a TI value of 1.42; the corresponding specificity was 40%.

Optimal MCIG/LC_{50} Ratio: The accuracy, sensitivity, and specificity of FETAX, without metabolic activation, based on using MCIG/LC_{50} ratios ranging from 0 to 1.5 as the single decision criterion, compared to combined rat, mouse, and rabbit teratogenicity test results are presented graphically in **Figure 2**.

Maximal accuracy for FETAX, without metabolic activation, compared to combined rat, mouse, and rabbit teratogenicity test results was ~58% at MCIG/LC$_{50}$ ratios between 0 and 0.2. At MCIG/LC$_{50}$ ratios between 0.2 and 0.4, accuracy steadily decreased to ~40% and then remained relatively constant at MCIG/LC$_{50}$ ratios up to 1.5. Sensitivity was 85% at an MCIG/LC$_{50}$ ratio of 0.08; the corresponding specificity was 13%.

When compared to combined rat, mouse, and rabbit teratogenicity results, accuracy based on using either a TI value or an MCIG/LC$_{50}$ ratio as the single decision criterion value was never greater than ~60%. This level of accuracy does not support the use of FETAX, as currently conducted, as a possible replacement *in vitro* assay for *in vivo* laboratory mammal teratogenicity tests. Using either the TI value or the MCIG/LC$_{50}$ ratio as the single decision criterion, a sensitivity of at least 85% (i.e., positive teratogens are correctly identified 85% of the time) was accompanied by a specificity of less than 40%. This low specificity corresponds to a false positive rate of greater than 60%. The poor specificity at a sensitivity of 85% raises concerns about the use of FETAX as a screening assay.

6.6.1.2 Human Teratogenicity Study Results

Optimal TI Value: The accuracy, sensitivity, and specificity of FETAX, without metabolic activation, based on using TI values ranging from 0 to 49 as the single decision criterion, compared to human teratogenicity study results are presented graphically in **Figure 3**.

Maximal accuracy for FETAX, without metabolic activation, compared to human teratogenicity study results was ~60% at TI values around 3.0. Accuracy then decreased to ~50% at higher TI values. Sensitivity was 85% at a TI value of 1.0; the corresponding specificity was 8%.

Optimal MCIG/LC$_{50}$ Ratio: The accuracy, sensitivity, and specificity of FETAX, without metabolic activation, based on using MCIG/LC$_{50}$ ratios ranging from 0 to 1.5 as the single decision criterion, compared to human teratogenicity study results are presented graphically in **Figure 4**.

Maximal accuracy for FETAX, without metabolic activation, compared to human teratogenicity study results was ~50% at MCIG/LC_{50} ratios between 0 and 0.06 or between 1.2 and 1.5. Sensitivity was 85% at an MCIG/LC_{50} ratio between 0.06 and 0.07; the corresponding specificity was 8%.

When compared to human teratogenicity results, maximum accuracy based on using either a TI value or an MCIG/LC_{50} ratio as the single decision criterion was never greater than about 50%. This value is lower than the previously reported accuracy of 64% calculated using an MCIG/LC_{50} ratio of less than 0.3 as the decision criterion for FETAX, without metabolic activation, compared to human teratogenicity study results (**Table 14**). This difference presumably reflects the use of median values in this analysis versus the weight-of-evidence approach used to generate the data for **Table 14**. This level of accuracy does not support the use of FETAX, as currently conducted, as apotential replacement *in vitro* assay for *in vivo* laboratory mammal teratogenicity tests. Using either the TI value or the MCIG/LC_{50} ratio as the single decision criterion, a sensitivity of at least 85% (i.e., positive teratogens are correctly identified 85% of the time) was accompanied by a specificity of less than 10%. This low specificity corresponds to a false positive rate of greater than 90%. This poor specificity at a sensitivity of 85% raises concerns about the use of FETAX as a screening assay.

6.6.2 Characteristic Malformations Induced in *X. laevis* Embryos

Qualitative information on the types of malformations was reported for 35 substances (**Appendices 2** and **3)**. Three of these were environmental samples, while the remaining 32 were individual substances. Malformations reported most commonly (i.e., reported for at least ten substances) included gut miscoiling, craniofacial malformations, and microencephaly. Substances inducing such malformations are provided in **Table 19**.

Table 19. Substances Inducing Gut Miscoiling, Craniofacial Malformations, or Microencephaly in *X. laevis* Embryos

Substance	FETAX Malformation(s) Induced
5-Azacytidine	Gut miscoiling; craniofacial malformations; microencephaly
5-Fluorouracil	Gut miscoiling; microencephaly
Amaranth	Gut miscoiling; craniofacial malformations
Bisphenol A	Craniofacial malformations
Copper (1)	Gut miscoiling; craniofacial malformations; microencephaly
Copper (2)	Gut miscoiling; craniofacial malformations; microencephaly
Copper sulfate	Gut miscoiling; craniofacial malformations; microencephaly
Desisopropyl atrazine	Microencephaly
Diethylene glycol	Gut miscoiling
Glycerol	Gut miscoiling; craniofacial malformations
Hydroxyurea	Microencephaly
Maneb	Craniofacial malformations
Methotrexate	Gut miscoiling; microencephaly
Nickel chloride	Gut miscoiling; craniofacial malformations
Pentachlorophenol	Gut miscoiling; craniofacial malformations; microencephaly
Permethrin	Microencephaly
Phthalic acid	Gut miscoiling
Propylthiourea	Craniofacial malformations
Pseudoephedrine	Gut miscoiling; craniofacial malformations

Sodium arsenite	Gut miscoiling; craniofacial malformations
Sodium iodoacetate	Gut miscoiling
Zinc (1)	Gut miscoiling; craniofacial malformations; microencephaly
Zinc (2)	Gut miscoiling; craniofacial malformations; microencephaly
Zinc sulfate heptahydrate	Gut miscoiling; craniofacial malformations; microencephaly

Other malformations reported less frequently are as follows, in decreasing order of occurrence:

- microopthalmia,
- opthalmic malformations,
- pericardial edema,
- mouth deformities,
- visceral edema,
- muscular kinking,
- facial abnormalities,
- gut malformations,
- edema,
- skeletal kinking,
- blistering of the dorsal fin,
- eye malformations,
- head anomalies,
- abnormal heart coiling,
- bent tail,
- curved tail tip,
- notocord defects,
- brain abnormalities,
- improper skin pigmentation,
- visceral hemorrhage,
- anencephaly,
- dermal blisters,
- incomplete gut coiling,
- hunchback,
- hydrocephaly,
- rupture of the eye pigment vesicle,
- opthalmic edema,
- axial skeletal anomalies,
- failure of the choriod to fuse,
- hypopigmented eyes,
- fin expansion,
- malformed fins,
- heart anomalies,
- enlarged heart, and
- vertebral fusions.

In the FETAX Phase III.3 Validation Study, Bantle et al. (1999) evaluated study results based on both single and multiple decision criteria. Using multiple decision criteria, test substances were classified as equivocal when either a TI value greater than 1.5 or an MCIG/LC_{50} ratio less than 0.30 was obtained. In such situations, the types and severity of malformations in *X. laevis* embryos were examined for guidance in assessing teratogenic hazard. However, due to the subjectivity of malformation identification, a decision was made that this approach should not be made a permanent part of the decision criteria by the investigators.

Dr. D. Fort (personal communication) has recently re-evaluated the FETAX Phase III.3 Validation Study results based on limiting the analysis of the EC_{50} to malformations deemed characteristic for the substance tested, rather than using data on all malformations as described in the ASTM FETAX Guideline (1991, 1998). Using the preserved embryos, Dr. D. Fort (personal communication) has recently re-evaluated the types and incidences of malformations in the various studies conducted in the FETAX Phase III.3 Validation Study. Subsequently, Dr. Fort then limited the analysis of the EC_{50} to malformations deemed characteristic for the substance tested, rather than using data on all malformations as described in the ASTM FETAX Guideline (1991, 1998). The embryos were re-evaluated to ensure the use of a uniform criteria in identifying malformations. The premise behind the use of characteristic malformations to evaluate the potential teratogenic hazard of a test substance is that any given teratogenic agent induces a syndrome characteristic of that substance. Non-specific, or background, malformations are also found in any given study. Malformations that are characteristic of the test substance should increase in frequency and possibly severity with increasing concentrations of the test substance. Malformations that occur sporadically and do not increase in frequency or severity with respect to test substance concentration are not likely directly due to the test material itself. To evaluate FETAX studies using this criterion, both characteristic and non-characteristic malformations are determined. However, statistical evaluation of the malformation data is limited to characteristic malformation data only. Because an evaluation of malformations is subjective, a secondary review of the scoring process is recommended (D. Fort, personal communication).

A preliminary assessment of the results of the re-analysis indicated that the use of the characteristic malformation criterion resulted in decreased intra- and inter-laboratory variability, a decreased number of equivocal test calls, and increased endpoint precision. Further, since this approach considers the syndrome associated with exposure to a given substance, it provides a more accurate means of comparing results between species. This approach may or may not increase the predictive accuracy of FETAX since that depends on the responsiveness of *Xenopus* to the test material. The disadvantages include, increased time required to evaluate each test, greater knowledge required by the technical staff, and a rigid QA/QC program to enforce secondary data review. However, in this re-analysis, all data on characteristic malformations were collected by the same scorer, which would be inherently expected to reduce inter-laboratory variability. NICEATM suggests that this approach has merit and that the process by which characteristic malformations is recognized *posthoc* needs to be evaluated across multiple laboratories.

Another aspect of characteristic malformations in FETAX that has yet to be critically explored is the correlation between the types of agent-specific malformations induced in *X. laevis* and those induced by the same agent in rats, mice, and rabbits, or in humans. A very limited assessment by Sabourin and Faulk (1987) and one more recently by Fort et al. (2000a) suggested a positive correlation between the types of malformations induced in laboratory mammals and in *Xenopus* embryos. A more extensive evaluation of the correlation between the types of malformations induced in laboratory mammals and in *Xenopus* embryos is currently in progress by NTP using data collected in the FETAX Phase III.3 Validation Study. The results of this assessment may support the validity of additional research in this area.

6.6.3 Evaluation of Growth Inhibition

In FETAX, the ratio between the MCIG and the LC_{50} is used as one criterion for identifying teratogens. The MCIG is the minimal concentration to inhibit growth, as determined by comparing the mean head-to-tail length at each test concentration compared to the appropriate control value, using student's t-test. However, because an assessment of growth is not required for range-finding tests (ASTM, 1991; 1998), the test concentrations selected for the definitive tests are frequently not conducive to an adequate assessment of the MCIG. As a consequence, the MCIG has been associated with the greatest inter-laboratory variability (see **Section 7**). Dr. D. Fort (personal communication) has suggested that a point estimate for growth inhibition, rather than the MCIG, would enhance the performance characteristics of FETAX. The possible effect of this modification to the decision criteria for FETAX on performance and the possible protocol changes needed for implementation have not yet been determined.

6.6.4 The Use of Confidence Intervals

Dr. D. Fort (personal communication) has suggested that the FETAX performance characteristics would be increased if 95% confidence intervals were used for statistically identifying TI values (and other point estimates) that are significantly greater than the decision point. This approach would allow for the variability among the replicate definitive tests to be considered when identifying a positive response in FETAX. The utility of this approach has yet to be evaluated.

6.6.5 Performance of FETAX with Metabolic Activation

In the FETAX Phase III.2 Validation Study, caffeine and CP were evaluated for their teratogenic activity in both the absence and presence of an exogenous MAS. This validation study was conducted because the investigators recognized the importance of including the capacity for metabolic activation. Based on the results of this validation study, the investigators concluded that the inclusion of metabolic activation in the assay was essential if FETAX was to be used to predict developmental hazard in mammals (including humans) but that the methodology required further development. The FETAX Phase III.3 Validation Study extended the Phase III.2 Validation Study results by testing 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), with and without metabolic activation, in three laboratories with extensive FETAX experience. The rationale for the selection of the test substances was not provided in the validation report. However, it is likely that selection was based on the availability of relevant laboratory mammal data and the suitability of the test substance for testing in FETAX (e.g., water solubility, lack of volatility). It does not appear that selection was based on the known or suspected requirement for metabolic activation to be a teratogen. NICEATM evaluated the possible metabolic activation requiring status of all substances tested in FETAX with an MAS. Identification of the possible involvement of metabolic activation was based on whether the substance was positive in one or more *in vitro* genetic toxicological tests (generally the *Salmonella typhimurium* reverse mutation assay) in the presence of metabolic activation only. *In vitro* genetic toxicology data were obtained from the EPA Genetic Activity Profile (GAP) database (www.epa.gov/gapdb/) and the NTP Salmonella test database. This method for identifying substances that may require metabolic activation to be teratogenic *in vitro* assumes a common mechanism between mutagenicity and teratogenicity that may not be valid. The results of this determination are presented in **Table 20**, with substances ranked by the increasing ratio of the TI with metabolic activation to the TI without metabolic activation. Also provided in **Table 20** is the FETAX result for studies conducted with and without metabolic activation, based on the single decision criterion of a TI greater than 1.5.

Of the 35 substances tested with metabolic activation in FETAX, useful *in vitro* genetic toxicology data were located on 15 substances (43%). Of these 15 substances, 11 were genotoxic in the absence of metabolic activation and four were only genotoxic with metabolic activation. In FETAX, with and without metabolic activation, three of the 35 substances were classified as negative under both metabolic conditions, seven were positive with metabolic activation only, three were positive without metabolic activation only, and 22 were positive under both metabolic conditions. Of the four substances requiring metabolic activation to be genotoxic *in vitro*, two substances were positive in FETAX with metabolic activation only while the other two substances were active in FETAX with and without metabolic activation. Of the eleven substances that are genotoxic *in vitro* without metabolic activation, two substances were positive in FETAX with metabolic activation only, two were positive in FETAX without metabolic activation only, and the remaining seven substances were positive in FETAX with and without metabolic activation.

The information in **Table 20** was also evaluated based on the assumption that a ratio of the median TI with metabolic activation to the median TI without metabolic activation of

Table 20. Substances Tested in FETAX With Metabolic Activation: Identification of Possible Metabolic Activation Requiring Substances

Substance	Requires MA*	Result Without MA	Result With MA	TI With MA/ TI Without MA
Doxylamine succinate	No	+	+	0.01
Nicotine		+	+	0.01
Hydrazine	No	+	+	0.03
Acetylhydrazide		+	+	0.06
4-Bromobenzene		+	-	0.13
Cytochalasin D	No	+	-	0.38
Sodium iodoacetate		-	-	0.42
Theophylline	No	+	-+	0.46
Caffeine	No	+	+	0.68
Isoniazid		+	+	0.70
Sodium bromate		+	+	0.72
Solanine		-	-	0.76
Triethylene glycol dimethyl ether		+	+	0.78
Phenytoin		+	+	0.80
Isonicotinic acid		+	+	0.84
N-Ethyl-N-nitrosourea	No	+	+	0.88
Boric Acid		+	+	0.89
Tribromoacetic acid		+	+	0.92
7-Hydroxycoumarin		+	+	1.00
Ethylene glycol	No	+	+	1.00

Table 20. Substances Tested in FETAX With Metabolic Activation: Identification of Possible Metabolic Activation Requiring Substances (Continued)

Substance	Requires MA*	Result Without MA	Result With MA	TI With MA/ TI Without MA
Acrylamide	No	+	+	1.01
Phthalic acid		-	-	1.04
3-Methylxanthine		+	+	1.07
Diethylene glycol		+	+	1.08
Dichloroacetate	No	-	+	1.11
1-Methylxanthine		+	+	1.12
Glycerol		-	+	1.18
Sodium arsenite	No	-	+	1.28
2-Acetylaminofluorene	Yes	+	+	1.34
4-Hydroxycoumarin		-	+	1.67
CP	Yes	-	+	1.85
Acetaminophen		-	+	1.92
Trichloroethylene	No	+	+	2.10
Urethane	Yes	+	+	3.91
Benzo[a]pyrene	Yes	-	+	6.83

The terms “No” and “Yes” indicates chemicals that do not or do appear to require metabolic activation, respectively, to induce a positive response in an *in vitro* genetic toxicological test according to the EPA Genetic Activity Profile (GAP) database (www.epa.gov/gapdb/) and the NTP Salmonella test database. MA = metabolic activation.

*Indicates substances without relevant metabolic activation-requiring information in these two databases.

[1] Classification of the test substance in FETAX based on a weight-of-evidence approach where multiple studies had been conducted, using a TI >1.5 as the single decision criterion.

[2] Ratio of median TI value with metabolic activation to median TI value without metabolic activation.

approximately one indicates independence of metabolism, while a ratio below 0.5 indicates decreased activity with metabolic activation and a ratio above 1.5 indicates increased activity with metabolic activation. Eight of the 35 substances tested with metabolic activation exhibited a with metabolic activation/without metabolic activation TI ratio below 0.5. One of these eight substances was negative in FETAX with and without metabolic activation, three were positive in FETAX without metabolic activation only, and four were positive in FETAX with and without metabolic activation. Six of the 35 substances tested with metabolic activation exhibited a with metabolic activation/without metabolic activation TI ratio greater than 1.5. Four of these six substances were positive in FETAX with metabolic activation only, and two were positive in FETAX with and without metabolic activation.

This evaluation revealed that most of the 35 substances tested with metabolic activation were not known to require metabolic activation to be active *in vitro*, but that there was a tendency towards increased activity in FETAX with metabolic activation for those substances that required metabolic activation to be genotoxic *in vitro*. Based on the limited database, additional studies to validate the role of metabolic activation in FETAX appear to be justified.

6.7 Strengths and Limitations of FETAX in Terms of Performance Characteristics

FETAX is a 96-hour *in vitro* whole-embryo test developed to determine the teratogenic and developmental toxicity potential of chemicals, metals, and complex mixtures (ASTM, 1991; 1998; Finch, 1994). It is essentially an organogenesis test, and organogenesis is highly conserved across amphibians and laboratory mammals. The first 96 hours of embryonic development in *Xenopus* parallel many of the major processes of human organogenesis (ASTM, 1991; 1998). Thus, it was anticipated that FETAX should be useful in predicting potential human developmental toxicants and teratogens (ASTM, 1991; 1998). Due to the nature of the endpoints assessed, FETAX does not provide information on substances that may induce functional developmental deficits in mammals. Because FETAX has been concluded by the developers to be easy, rapid, reliable, and inexpensive, the test (with and without metabolic activation) has been proposed as a screening assay for potential human teratogens and developmental toxicants (ASTM, 1991; 1998). As a screening test, a positive FETAX response would indicate a potential human hazard while a negative FETAX response would not indicate the absence of a hazard. In the role of a screening assay, a negative response would be followed by *in vivo* laboratory mammal testing, while a positive response would require no further testing unless the investigator is concerned about a potential false positive response.

NICEATM evaluated the performance characteristics of FETAX, with and/or without metabolic activation, compared to teratogenicity test results in rats, mice, and/or rabbits, and compared to human teratogenicity study results. In this analysis, different decision criteria (i.e., single decision criteria based on a TI value greater than 1.5 or 3.0, or an MCIG/LC_{50} ratio less than 0.30; multiple decision criteria based on a TI value greater than 1.5 or 3.0 plus an MCIG/LC_{50} ratio less than 0.30) reported in the literature for identifying teratogenic potential in FETAX were evaluated. When the performance for FETAX, with and without metabolic activation, was determined compared to combined rat, mouse, and rabbit teratogenicity results, maximal accuracy was 60%, maximal sensitivity was 80%, and maximal specificity was 56%. These values occurred using different decision criteria. When the performance for FETAX, with and without metabolic activation, was determined compared to human teratogenicity study results, maximal accuracy was 73%, maximal sensitivity was 93%, and maximal specificity was 79%. Again, each maximal value occurred using different decision criteria.

NICEATM also evaluated the performance characteristics of FETAX, with and without metabolic activation, compared to combined rat, mouse, and rabbit teratogenicity test results and human teratogenicity study results by chemical and product class using single decision criteria (i.e., TI >1.5, TIMCIG/LC_{50} <0.3). Analyses were limited to chemical and product classes containing a minimum of 15 FETAX test substances with corresponding animal or human teratogenicity data. Only five chemical classes and one product class were evaluated for performance characteristics compared to the combined rat, mouse, and rabbit teratogenicity test results, while only one chemical class and one product class were evaluated for performance characteristics compared to human teratogenicity study results. The accuracy of FETAX compared to laboratory mammal teratogenicity test results was somewhat improved compared to that for the total database for amides, nitrogen heterocyclic compounds, and organic (phenolic and carboxylic) acids. Performance for the other chemical classes and the single product class evaluated were not different from the performance of FETAX compared to the total database. The performance characteristics of FETAX compared to human teratogenicity study results were not improved for nitrogen heterocyclic compounds and pharmaceuticals compared to that for the total database.

In response to these results, NICEATM attempted to identify the optimal TI value or MCIG/LC_{50} ratio to use as a single decision criterion in evaluating FETAX data. Performance characteristics (accuracy, sensitivity, specificity) were determined for FETAX, without metabolic activation, compared to combined rat, mouse, and rabbit teratogenicity test results or compared to human teratogenicity study results. When compared to laboratory mammal or human data, maximum accuracy based on using either a TI value or an MCIG/LC_{50} ratio as the single decision criterion value was never greater than ~60%. This level of accuracy does not support the use of FETAX, as currently conducted, as a replacement *in vitro* assay for *in vivo* laboratory mammal teratogenicity tests. Using either the TI value or the MCIG/LC_{50} ratio as the single decision criterion, a sensitivity of at least 85% (i.e., positive teratogens are correctly identified 85% of the time) was accompanied by a specificity of less than 30%. The poor specificity at a sensitivity of 85% raises concerns about the use of FETAX as a screening assay.

Based on these analyses, additional efforts to optimize the decision criteria appear to be warranted. Several modifications that are potentially useful (e.g., use of characteristic malformations, use of confidence intervals) were discussed in **Section 6.6**.

6.8 Data Interpretation Issues

As specified by the ASTM FETAX Guideline (1991, 1998), three separate decision criteria (TI>1.5; MCIG/LC_{50}<0.3, and severity of malformation) have been used to identify potential human teratogens. The ASTM FETAX Guideline (1991, 1998) concludes that any single decision criterion is sufficient to identify a potential teratogenic hazard, and that these three decision criteria are based on empirical evidence resulting from over 100 materials tested (without metabolic activation) in FETAX. In the NICEATM analysis of the performance characteristics of FETAX compared to either laboratory mammal or human teratogenicity results, these as well as multiple decision criteria were considered. The multiple decision criteria (TI >1.5 or TI >3.0 plus MCIG/LC_{50} <0.3) evaluated were those used in the most recent FETAX Validation Study (Bantle et al., 1999) and in a comparative FETAX-rat study conducted by Fort et al. (2000a). This analysis indicates that the use of a TI value greater than 1.5 and an MCIG/LC_{50} ratio below 0.3 as the single decision criteria results in the maximum accuracy for laboratory mammal and human teratogenicity results, respectively. The use of multiple decision criteria did not significantly increase the ability of FETAX to correctly identify mammalian (including human) teratogens. These analyses suggest that additional effort is warranted to investigate and optimize the methods by which FETAX data are collected and interpreted.

6.9 Section 6 Conclusions

The use of single decision criterion based on a TI value greater than 1.5 appeared to provide the most optimal approach in terms of accuracy and sensitivity for predicting combined laboratory mammal teratogenicity data. The use of multiple decision criteria had no appreciable effect on accuracy or sensitivity but increased specificity when equivocal results were excluded from the analysis.

Using either TI decision criteria value, the performance characteristics of FETAX, with and without metabolic activation, compared to teratogenicity data for rats, mice, or rabbits appeared to be similar. These FETAX performance characteristics were not very different from the performance characteristics based on combined rat, mouse, and rabbit teratogenicity data. Comparing the performance characteristics for each species as a function of the TI value, increased accuracy and sensitivity but decreased specificity was associated with the use of a TI value greater than 1.5 rather than 3.0.

In general, the use of single decision criterion based on an MCIG/LC_{50} ratio lower than 0.3 appeared to provide the most optimal approach for predicting human teratogenicity data. The use of multiple decision criteria increased sensitivity when equivocal results were classified as positive and specificity when equivocal results were classified as negative.

Five chemical classes and one product class was evaluated for performance characteristics compared to the combined rat, mouse, and rabbit teratogenicity test results. Among the chemical and product classes evaluated, a decision criterion based on a TI value greater than 1.5 generally provided greater accuracy and sensitivity, but less specificity, than one based on either on a TI value greater than 3.0 or on an MCIG/LC_{50} ratio of less than 0.3. The accuracy of FETAX compared to laboratory mammal teratogenicity test results for nitrogen heterocyclic compounds, and organic (phenolic and carboxylic) acids was somewhat improved compared to that for the total database. Performance compared to the other chemical classes and the single product class (pharmaceuticals) evaluated were not different from the performance of FETAX compared to the total database.

Maximal performance characteristics for laboratory mammal data compared to human results were obtained using rat, mouse, or combined laboratory mammal teratogenicity data. Performance characteristics for rabbit teratogenicity data were generally poor compared to that for the other two species. The rat, mouse, or combined laboratory mammal performance characteristics compared to human teratogenicity results appeared to be slightly but consistently improved over the performance of FETAX when TI was used as the single decision criterion.

NICEATM conducted an evaluation for the optimal TI value or MCIG/LC_{50} ratio to use as a single decision criterion in evaluating FETAX data. Performance characteristics (accuracy, sensitivity, specificity) were calculated for FETAX compared to combined laboratory mammal (rat, mouse, and rabbit) or human teratogenicity results. When compared to combined laboratory mammal or human teratogenicity results, accuracy based on using either a TI value or an MCIG/LC_{50} ratio as the single decision criterion value was never greater than ~60%. Using either the TI value or the MCIG/LC_{50} ratio as the single decision criterion, a sensitivity of 85% was accompanied by specificity of 40% or less. The magnitude of these values suggests that FETAX is not appropriate as a replacement for *in vivo* laboratory mammal teratogenicity tests, and that its use as a screen, based on current decision criterion, is problematic.

An analysis of FETAX database revealed 43 substances that were discordant with laboratory mammal teratogenicity results and/or human teratogenicity results, seven substances that were concordant with laboratory mammal teratogenicity data but discordant with human teratogenicity results, and three substances that were discordant with laboratory mammal but concordant with human teratogenicity results. The bases for these discordant results are not known.

The inclusion of an exogenous MAS in FETAX is considered to be essential for predicting developmental hazard in humans (ASTM, 1991; 1998). Two FETAX validation studies (Phase III.2 and Phase III.3) were conducted in which substances were tested with and without metabolic activation. However, selection of the substances tested did not appear to have been based on whether or not metabolic activation was required for teratogenic activity *in vitro*. NICEATM evaluated the possible metabolic activation requiring status of these and other substances tested in FETAX with metabolic activation. Identification of the possible involvement of metabolic activation was based on whether the substance was positive in one or more *in vitro* genetic toxicological tests (generally the *Salmonella typhimurium* reverse mutation assay) with, but not without, metabolic activation. This method for identifying substances that may require metabolic activation to be teratogenic *in vitro* assumes a common mechanism between mutagenicity and teratogenicity that may not be valid. This evaluation revealed that most of the 35 substances tested with metabolic activation were not known to require metabolic activation to be active *in vitro*, but that there was a tendency towards increased activity in FETAX with metabolic activation for those substances that required metabolic activation to be genotoxic *in vitro*. Based on the limited database, additional studies to validate the role of metabolic activation in FETAX appear to be justified.

Several approaches have been suggested for modifying the decision criteria used to distinguish between a positive and a negative FETAX response. These approaches include an evaluation of the EC_{50} based on characteristic malformations only, a point estimate rather than an MCIG for growth inhibition, and 95% confidence intervals for statistically identifying TI values (and other point estimates) that are significantly greater than the decision point. The effects of these suggested approaches on the performance characteristics of FETAX have not yet been evaluated.

Another aspect of characteristic malformations in FETAX that has yet to be critically explored in the correlation between the types of agent-specific malformations induced in *X. laevis* and those induced by the same agent in rats, mice, and/or rabbits, or in humans. An evaluation is in progress by NTP using data collected in the FETAX Phase III.3 Validation Study. The results of this assessment may indicate the appropriateness of additional research in this area.

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.

8.0 TEST METHOD DATA QUALITY

8.1 Extent of Adherence to GLP Guidelines

Studies were not conducted in compliance with national or international GLP guidelines, nor were they generally conducted at facilities at which GLP studies are normally conducted.

8.2 Results of Data Quality Audits

The NTP Quality Assurance (QA) Unit conducted a limited audit of the FETAX Phase III.3 Validation Study. In this audit, the data provided in the published report were compared for accuracy, consistency, and completeness when compared to original records of the studies, as supplied by the three participating laboratories. A number of the values tabulated in the report could not be confirmed, because of omission of corresponding data in the provided summary sheets, illegible print, or inadequate description of statistical methods used to compute the values. General findings include:

- Data trails, study records, and results analysis procedures were not sufficient to support a QA audit;

- Some calculations apparently used the formula for standard deviation of a sample while others used that for a population; and

- The presence of transcriptional errors between the published report and the original data.

8.3 Impact of GLP Deviations and/or Data Audit Non-Compliance

A review by NICEATM of the discrepancies noted in the QA data audit by the NTP QA Unit did not reveal any that significantly altered the general conclusions presented in the FETAX Phase III.3 Validation Study report (Bantle et al., 1999). However, the audit results do indicate the general lack of GLP compliance in the validation studies.

8.4 Section 8 Conclusions

Studies were not conducted in compliance with national or international GLP guidelines, nor were they generally conducted at facilities at which GLP studies are normally conducted. A review by NICEATM of discrepancies noted in the QA data audit by the NTP QA Unit did not reveal any findings that significantly altered the general conclusions presented in the FETAX Phase III.3 Validation Study report (Bantle et al., 1999). However, the audit results do indicate the general lack of GLP compliance in the validation studies. It is recommended that future validation studies be conducted in compliance with national and international GLP guidelines.

9.0 OTHER SCIENTIFIC REPORTS AND REVIEWS

9.1 Availability of Other FETAX Data

The focus of the BRD has thus far been on the use of FETAX, as defined in the ASTM FETAX Guideline (1991, 1998), as a screening assay for identifying substances that may pose a developmental hazard in humans. The sources for the FETAX data evaluated for that purpose included peer-reviewed literature (including studies accepted for publication) and non peer-reviewed book chapters. Information not considered included abstracts, manuscripts not accepted for publication, studies where test substances were not identified, studies not conducted in general compliance with the ASTM FETAX Guideline (1991, 1998), and published reports lacking appropriate quantitative FETAX data. Published information on substances not appropriately identified was excluded to avoid the possibility of duplication of results during an analysis of the performance characteristics of FETAX.

9.2 Conclusions of Other Peer Reviews

No other independent peer reviews of FETAX have been conducted. However, an evaluation of the performance of FETAX was published in 1987 by Sabourin and Faulk, based on FETAX studies conducted in their laboratory. FETAX was evaluated as a candidate *in vitro* teratogenicity assay by testing 35 chemicals listed in a consensus NTP teratogenesis chemical repository. The authors concluded that the most promising endpoints were embryo malformations and growth during the 96-hour test. In FETAX, 17 of 20 *in vivo* laboratory mammal teratogens tested positive, and 12 of 15 negative laboratory mammal teratogens tested negative, for an overall accuracy of 83%. Furthermore, the concordance between the types of malformations (e.g., skeletal, visceral, nervous, optic, osmoregulatory) detected in *Xenopus* and in mammals was 67% for 19 teratogens. The authors concluded that FETAX was a strong candidate for further consideration as a teratogen screen. However, due to the lack of quantitative FETAX malformation data for the substances considered in this review, this information was not considered in the evaluation of FETAX conducted by NICEATM.

The utility of *X. laevis* for identifying human developmental hazards was discussed also in a review by Sakamoto et al. (1992). In this review, based on analysis of seventeen substances tested in-house and a review of the literature, Sakamoto et al. concluded that the *Xenopus* embryo and larva system is a good candidate for a simple and effective test system to evaluate developmental toxicants. Due to significant protocol differences, the data provided in this review were not considered by NICEATM in the evaluation of the performance characteristics of FETAX.

Recently, Fort et al. (2000a) assessed the predictive validity of FETAX, with and without metabolic activation, for identifying the potential developmental toxicity of a group of diverse coded chemicals (fungicides, herbicides, nematocides) by comparison with results from *in vivo* teratogenicity studies in rats. A total of 12 chemicals were evaluated, three of which were classified as teratogenic *in vivo*, four of which were embryolethal but not teratogenic *in vivo*, and five that did not produce any developmental toxicity *in vivo*. The FETAX studies followed the 1991 ASTM FETAX guideline. In this study, each test chemical was judged to have developmental hazard when the TI value was greater than 3.0, the MCIG/LC_{50} ratio was less than 0.30, and/or strong characteristic malformations were induced. If the TI value was between 1.5 and 2.9, the MCIG/LC_{50} ratio was greater than 0.3, but characteristic malformations of moderate severity were induced, the chemical was classified as equivocal. The test chemical was judged not hazardous when all decision criteria fell into the non-hazard category. The investigators concluded that FETAX correctly predicted that three chemicals had strong teratogenic potential (were positive in FETAX), four had low teratogenic hazard potential but were embryolethal (i.e., were equivocal in FETAX), and five posed little if any developmental toxicity hazard (i.e., were negative in FETAX). In addition, the investigators stated that within a family of chemical analogs, the compounds could be ranked according to relative teratogenic hazard and that, for the teratogenic compounds, the types of malformations induced in *Xenopus* mimicked the abnormalities induced *in vivo* in rats. Based on these results, the investigators concluded that the results confirmed that the FETAX assay is predictive and can be useful in an integrated biological hazard assessment for the preliminary screening of chemicals.

Although supportive of a conclusion that FETAX was predictive of rat teratogenicity results, these data could not be used by NICEATM in their evaluation of the performance characteristics of FETAX. First and most importantly, the identity of each test chemical was not available; each test substance was identified by chemical class (e.g., substituted diphenyl ether) only. Furthermore, although tested with and without metabolic activation, only a single set of FETAX data were provided for each compound and information on the metabolic activation status of that data was not provided in the publication. Finally, mortality and malformation rates for the inactivated MAS and CP only negative control dishes are reported to range from 0-25% and 2.5-100%, respectively. These values exceed the 10% mortality and malformation limits that appear to be established by the ASTM FETAX guideline (1991, 1998) as being acceptable.

9.3 Section 9 Conclusions

No other independent peer reviews of FETAX were located. An evaluation of the performance of FETAX was published in 1987 by Sabourin and Faulk, based on FETAX studies conducted in their laboratory. However, due to the lack of quantitative FETAX data for the substances considered in this review, this information was not considered in the evaluation of FETAX conducted by NICEATM. The utility of *X. laevis* for identifying human developmental hazards was discussed also in a review by Sakamoto et al. (1992), in which it was concluded that the *Xenopus* embryo and larva system is a good candidate for a simple and effective test system to evaluate developmental toxicants.

10.0 ANIMAL WELFARE CONSIDERATIONS

10.1 Extent to which FETAX Will Reduce, Refine, or Replace Animal Use for Human Developmental Hazard Assessment

In terms of human developmental toxicology, FETAX is proposed as a screen for hazard identification, and thus would not totally eliminate the use of mammals in teratogenicity and developmental toxicity testing. If validated, the use of this *in vitro* assay would, however, reduce reliance on mammalian tests, and thereby reduce the number of mammals used. Each successful FETAX assay would potentially eliminate the use of approximately 190 rats and 112 rabbits in the typical segment 2 mammalian test. FETAX also offers substantial refinement in the way animals are used. Federal guidelines for teratogenicity and developmental toxicity testing recommend the use of 16 to 24 litters of rats for each dose level (U.S. EPA, 1991; U.S. FDA, 1994). In comparison, FETAX not only employs a non-mammalian alternative, but it is stated that fewer organisms are used per dose level (ASTM, 1991; 1998). In addition, embryos and not adult animals are used in FETAX, another refinement in the assay (ASTM, 1991; 1998).

The per-dose group numerical advantage stated in the ASTM FETAX Guideline (1991, 1998) for FETAX disappears when the recommended numbers of dose groups and of replicate experiments are taken into consideration. In a typical segment 2 test for one compound, ten different dose groups are tested for each species, both rodent and lagomorph. Typically, six dose groups of rodents and non-rodents are used in a pilot study, and four dose groups for each species is used in the definitive study. For the rodent segment of the segment 2 test, rats are usually used. Fifteen rats are used for each pilot study dose group, and 25 rats are used for each definitive test dose group (total number of rats = 190). For the lagomorph, rabbits are usually used, with eight rabbits used in each pilot study dose group, and 16 rabbits used in each dose group for the definitive portion of the test (total number of rabbits = 112). FETAX uses 40 to 50 embryos per dose level (80 to 100 for the concurrent control group), with a minimum of seven dose groups tested per range-finder assay, and five dose groups tested in each of three replicate tests, for a minimum of at least 1300 embryos (ASTM, 1991; 1998). Also, it is recommended that each study be conducted with and without metabolic activation, which would require a minimum of at least 2600 embryos.

FETAX would also reduce animal usage if the assay could be used:

- in the earliest stages of product development, to select for further development those compounds that are the least likely to cause developmental toxicity;

- to compare the developmental toxicity potential of a new chemical that is only a slight modification of an existing chemical that has already been tested *in vivo*; and

- to evaluate compounds for which testing is not routinely performed, usually because the anticipated exposure is very low (Spielmann, 1998).

10.2 Section 10 Conclusions

FETAX is proposed as a screen for human hazard identification, and thus will not totally eliminate the use of mammals in teratogenicity and developmental toxicity testing. However, if accepted as a screen, use of this *in vitro* assay would reduce reliance on mammalian tests, and would thereby reduce the number of mammals used.

11.0 OTHER CONSIDERATIONS

11.1 Test Method Transferability

11.1.1 Facilities and Major Fixed Equipment

As described in **Section 2.1.1**, adults should be kept in an animal room that is isolated from extraneous light that might interfere with a consistent 12-hour photoperiod. Adults can be maintained in large temperature-controlled aquaria or in fiberglass or stainless steel raceways at densities of four to six animals per 1800 cm^2 of water surface area. For conducting the actual FETAX assay, a constant temperature room or a suitable incubator for embryos is required, although a fixed photoperiod is unnecessary. The incubator or room must be capable of maintaining a temperature of 24 ± 2°C.

A binocular dissection microscope capable of magnifications up to 30x is required to count and evaluate embryos for malformations. A simple darkroom enlarger is used to enlarge embryo images two to three times for head to tail length measurements. It is also possible to measure embryo length through the use of a map measurer or an ocular micrometer. However, the process is greatly facilitated by using a digitizer interfaced to a microcomputer. The microcomputer is also used in data analysis.

Such facilities and equipment should be readily available in most laboratories.

11.1.2 Required Level of Personnel Training and Expertise

The estimated amount of technical training required for conducting the in-life portion of a FETAX study is from three to six months (D. Fort, personal communication). This training period may not be much different from that needed for conducting the in-life portion of a corresponding laboratory study using rats, mice, or rabbits. However, as noted in **Section 7.3.3**, concern was expressed during the FETAX validation studies that at least some of the inter-laboratory variability may have been caused by differences among scorers in their ability to identify malformations in *X. leavis* embryos. Although this concern was not verified when an analysis of performance characteristics was limited to the two most experienced laboratories, expertise in the recognition of malformations in embryos appears to require extensive training. The use of agent-specific characteristic malformations as a method for increasing the performance characteristics of FETAX (see **Section 6.6.2**) or to increase the level of expertise needed for malformation identification.

11.1.3 General Availability of Necessary Equipment and Supplies

The types of equipment and supplies needed to conduct FETAX are readily available from any major supplier.

11.2 Assay Costs

A complete FETAX study, with and without metabolic activation, following the ASTM FETAX Guideline (ASTM, 1998) and conducted in compliance with national/international GLP guidelines, should cost less than $25,000 per test substance (D. Fort, personal communication). In comparison, a complete rat Prenatal Developmental Toxicity Study (screening plus definitive) would cost about $120,000 (G. Jahnke, personal communication).

11.3 Time Needed to Conduct the Test

A complete FETAX study, with and without metabolic activation, following the ASTM FETAX Guideline (1998) and conducted in compliance with national/international GLP guidelines, would require less than two months to complete. This is in contrast to the six to seven months required for a complete rat study (G. Jahnke, personal communication).

11.4 Potential Effect of Tetraploidy on the Use of *X. laevis* in FETAX

The genome of *X. laevis* is tetraploid (Vogel, 1999). The potential impact of tetraploidy on the extrapolation of teratogenic changes in *X. laevis* to laboratory mammals and humans is unknown. Galitski et al. (1999) have shown recently that gene expression is modulated by ploidy in yeast cells. Isogenic strains of yeast were compared at varying ploidy ranging from haploid to tetraploid. The mRNA levels of all genes were measured during exponential growth using oligonucleotide-probe microassays. It was found that the expression of 17 genes was altered due to changes in ploidy, with ten genes being induced and seven genes being repressed with increasing ploidy. With specific regard to developmental response, the investigators monitored the effect of ploidy on invasiveness, a developmental trait in yeast. As ploidy increases, invasiveness decreases due to repression of the FLO11 gene that is responsible for the production of a cell wall protein. This developmental modulation was verified by the restoration of invasiveness with the addition of a $FLO11^+$ plasmid.

In view of this finding, it may be useful to consider the potential value of a diploid species of *Xenopus*, such as *X. (Silurana) tropicalis* in FETAX. Although currently limited in availability, this species potentially offers several advantages over *X. laevis*. These advantages include a smaller size, greater ease in housing, more rapid maturation (four or five months as opposed to the one to two years for *X. laevis*), and the ability to be altered transgenetically for developmental mechanistic studies.

11.5 *Xenopus* Microarray Technology

One recent development, which may greatly increase the utility of FETAX for identifying and prioritizing developmental hazards, is in cDNA microarray technology. A cDNA microarray is a glass slide (or other support) containing a large number of genes or expressed sequence tags in a condensed array. Using cDNA microarrays, the expression of thousands of genes can be monitored simultaneously in multiple biological samples of interest, and the expression patterns compared.

This technology may be useful in identifying toxic substances individually or in mixtures; in determining whether toxic effects occur at low doses; in evaluating susceptible tissues and cell types; and in extrapolating effects from one species to another. In FETAX, treatment with a known developmental toxicant may provide a gene expression "signature" on a microarray, which represents the cellular response to this agent. When an unknown substance is tested, the microarray response can then be evaluated to see if one or more of these standard signatures is elicited. This approach might also be used to elucidate an agent's mechanism of action, assess interactions between combinations of agents, or allow for a comparison between altered gene function in *Xenopus* with changes in analogous genes in mammalian systems.

NIEHS has developed a custom "Toxchip" that is a human cDNA clone subarray-oriented toward the expression of genes involved in responses to toxic insult, including xenobiotic metabolizing enzymes, cell cycle components, oncogenes, tumor suppressor genes, DNA repair genes, estrogen-responsive genes, oxidative stress genes, and genes known to be involved in apoptotic cell death. In addition, chips to study responses in mouse, yeast, rat, and Xenopus are available. A Xenopus Chip v 1.0, containing 1000 Xenopus genes from a normalized library, has been developed by Dr. Perry Blackshear's laboratory at NIEHS. In response to increasing interest in this technology, NIEHS has implemented a cDNA Microarray Center to:

- identify toxicants on the basis of tissue-specific patterns of gene expression (molecular signature),
- elucidate mechanisms of action of environmental agents through the identification of gene expression networks,
- use toxicant-induced gene expression as a biomarker to assess human exposure,
- extrapolate effects of toxicants from one species to another,
- study the interactions of mixtures of chemicals,
- examine the effects of low dose exposures versus high dose exposures, and
- develop a public database of expression profiles.

Information on this NIEHS cDNA Microarray Center and progress on the Xenopus microarray chip can be found on the World Wide Web at http://dir.niehs.nih.gov/microarray/.

11.6 Other *In Vitro* Assays for Mammalian Developmental Hazard Identification

A number of *in vitro* systems have been considered as alternatives or screens to *in vivo* mammalian developmental toxicity assays (Kimmel et al, 1982; Smith et al., 1983; Kimmel, 1990; Brown, 1987, Schwetz et al., 1991; Tanumiura and Sakamoto, 1995; Brown et al., 1995; Spielman, 1998). In 1991, Kavlock et al. described a prototype developmental computerized database suitable for comparing the activity profiles of developmental toxicants. The information contained in these profiles could be used to compare qualitative and quantitative results across multiple assay systems, identify data gaps in the literature, evaluate the concordance of the assays, evaluate relative potencies, and examine structure activity relationships. In addition to *in vivo* mammalian assays, eight cellular assays and six *in vitro* embryo systems, including FETAX, being considered at that time were described.

Most recently, the European Centre for the Validation of Alternative Methods (ECVAM) has sponsored a series of validation studies of three *in vitro* assays considered suitable for the detection of substances posing a mammalian developmental hazard (Brown et al., 1995; Scholz et al., 1998; Genschow et al., 1999). The three *in vitro* assays being evaluated are the postimplantation whole rat embryo culture (WEC) assay, the micromass (MM) test, and the embryonic stem cell (EST) test. The WEC assay involves the cultivation of postimplantation whole rat embryos in which both general growth retardation and specific malformations of the cultivated embryo are assessed. This assay is relatively complex, covers only a part of organogenesis, requires high technical skills, and uses mammalian tissue and serum (Spielmann, 1998). In the MM test, primary limb bud cells of rat embryos are cultured and effects on the viability are compared to effects on the differentiation of these cells into chondrocytes. The EST makes use of the differentiation of cultured ES cells into cardiomyocytes. The advantage of this latter test is the use of an established cell line without the need to sacrifice pregnant animals.

In the prevalidation phase, three test chemicals with non- (saccharin), moderate (5,5 diphenylhydantoin), and strong- (cytosine arabinoside) embryotoxicity, along with a negative (Penicillin G) and a positive (5-fluoruracil) control chemical were repeatedly tested in each test in two laboratories (Scholz et al., 1998). The investigators concluded that the *in vitro* tests could be transferred from one laboratory to another and that reproducible results could be obtained. It was also concluded that the three methods were able to discriminate among the test chemicals according to their embryotoxic potential.

In the validation study, each of the tests is being evaluated in four laboratories under blind conditions. In an initial phase of the validation process, six of 30 test chemicals comprising different embryotoxic potential (non, weak, and strong embryotoxic) were tested (Genschow et al., 1999). The results were used to revise and enhance the prediction models for the three assays. The results obtained from evaluating the complete set of 30 chemicals have not yet been published. A list of the names of the 30 chemicals being tested is provided in **Table 38**, along with information on whether the chemical has been tested in FETAX, with or without metabolic activation. The ability of these *in vitro* assays to react to substances that require metabolic activation to be embryotoxic or the potential need for an exogenous MAS in the study protocol is not well defined and may need clarification.

The relative performance, cost-effectiveness, and flexibility of FETAX against other *in vitro* assays in identifying substances with mammalian developmental toxicity was not considered in developing this BRD. Sakamoto et al. (1992) has concluded that the use of *X. laevis* as an *in vitro* model system for the detection of mammalian developmental hazards offers a number of advantages in comparison to other *in vitro* model systems. The most important advantage is that

Table 37. List of Test Chemicals for the ECVAM Validation Study of Three *In Vitro* Embryotoxicity Tests (Genschow et al., 1999)

Chemical	**CAS No.**	**Tested in FETAX**	
		Without metabolic activation	**With metabolic activation**
Strong Embryotoxicity			
5-Bromo-2′-deoxyuridine	59-14-3	No	No
Methyl mercury chloride	115-09-3	No	No
Hydroxyurea	127-07-1	Yes	No
Methotrexate	59-05-2	Yes	No
all-trans-Retinoic acid	302-79-4	Yes	No
6-Aminonicotinamide	329-89-5	Yes	No
Moderate Embryotoxicity			
Boric acid	10043-35-3	Yes	Yes
Pentyl-4-yn-VPA	-	No	No
Valproic acid (VPA)	99-66-1	No	No
Lithium chloride	7447-41-8	No	No
Dimethadione	695-53-4	No	No
Methoxyacetic acid	625-45-6	No	No
Salicylic acid sodium salt	54-21-7	No	No
No Embryotoxicity			
Acrylamide	79-06-1	Yes	Yes
Isobutyl-ethyl-VPA	-	No	No
D-(+)-camphor	464-49-3	No	No
Dimethyl phthalate	131-11-3	No	No
Diphenhydramine hydrochloride	147-24-0	Yes	No
Penicillin G sodium salt	69-57-8	No	No
Saccharin sodium hydrate	82385-42-0	No	No

the development of the *Xenopus* embryo includes a number of developmental events, including cleavage, gastrulation, neurulation, and organogenesis, that are mechanistically comparable to those of mammals. Secondly, this *in vitro* system does not involve the use of any mammals (except sporadically as a source of materials to prepare an MAS).

11.7 Section 11 Conclusions

Sufficient information on facilities and equipment for establishing FETAX as a routine test is provided in the ASTM FETAX Guideline (1991, 1998). The estimated amount of technical training required for conducting the in-life portion of a FETAX study appears to be sufficient. However, based on concerns about differences in expertise in the identification of some of the more subtle malformations induced in *Xenopus* embryos, a more extensive training period may be required for the classification of malformations. The projected cost and study duration for a GLP compliant complete FETAX study, with and without metabolic activation, following the ASTM FETAX Guideline (1998), appears to be reasonable. The potential impact of tetraploidy on the extrapolation of teratogenic changes in *X. laevis* to laboratory mammals and humans needs to be considered. Furthermore, the advantage of using a diploid species of *Xenopus*, such as *X. tropicalis*, in FETAX, should be evaluated.

A number of *in vitro* systems have been proposed as alternatives or screens to *in vivo* mammalian developmental toxicity assays. A brief description of an ECVAM-sponsored validation of three *in vitro* assays considered potentially suitable for the detection of substances posing a mammalian developmental hazard was included. The relative performance, cost-effectiveness, and flexibility of FETAX against other *in vitro* assays in identifying substances with mammalian developmental toxicity was not considered in developing this BRD. However, the most important advantage of FETAX is that the development of the *Xenopus* embryo includes cleavage, gastrulation, neurulation, and organogenesis, and that these developmental events are considered to be mechanistically comparable to those of mammals. Also, this *in vitro* system does not involve the use of any mammals (except sporadically as a source of ingredients for the MAS).

FETAX FOR ECOTOXICOLOGICAL HAZARD ASSESSMENT USING WATER/SOIL/SEDIMENT SAMPLES

Much of the information provided in the sections dealing with FETAX for human developmental hazard identification (**Sections 1** through **11**) are applicable to the use of FETAX for ecotoxicological hazard assessment using water/soil/sediment samples. However, for ease of comparison, this BRD will continue to follow the structure described in the evaluation criteria guidelines found in the *Evaluation of the Validation Status of Toxicological Methods: General Guidelines for Submissions to the Interagency Coordinating Committee on the Validation of Alternative Methods* (**Appendix 15**).

12.0 INTRODUCTION AND RATIONALE FOR THE USE OF FETAX IN ASSESSING DEVELOPMENTAL HAZARDS IN WATER/ SOIL/SEDIMENT SAMPLES

12.1 Scientific Basis for the Use of FETAX

In developing alternative testing methods for ecotoxicology, there is a need to clearly define strategies and goals when undertaking testing procedures (Walker et al., 1998). Reduction, replacement, or refinement in animal use will be served by:

- developments and improvements in assays incorporating new techniques from biochemical/molecular biology that relate to mechanisms;

- further development of nondestructive assays for vertebrates, and assays for invertebrates;

- selection of the most appropriate species, strains and developmental stages in the light of new knowledge (but no additional vertebrate species for basic testing); and

- better integrated approaches incorporating biomarker assays, ecophysiological concepts, and ecological end points.

Maximum success depends on a flexible approach and expert judgment in interpretation. Testing protocols need to be realistic, taking into account particular problems with mixtures and volatile or insoluble chemicals (Walker et al., 1998).

The purpose of FETAX for ecotoxicological assessment is to identify and prioritize aquatic environments, soils, or sediments that contain naturally occurring or anthropogenic substances, which pose a developmental hazard to living organisms. Ecotoxicological testing is usually performed using multiple test species. For each species, it is a combination of toxicants, water quality, and the susceptibility of the organism itself that defines the hazard for a specific concentration of a toxicant within defined water quality conditions. Ecotoxicological standards are generally based on the susceptibility of the adult animal, which may not provide adequate protection for embryonic development and reproduction in many species. Early embryonic and juvenile stages are often the most susceptible periods for the toxic effects of many environmental contaminants. Furthermore, it is inherently impossible to evaluate developmental toxicity without exposing animals throughout development and assessing for adverse effects at multiple life stages. Due to the sensitivity of embryonic development in amphibians to water quality, FETAX is thought to be relevant as a conservative 'sentinel' estimator of ecotoxicologic hazard (ASTM, 1991; 1998).

12.2 Intended Uses of FETAX

12.2.1 Intended Regulatory Uses and Rationale

FETAX, without metabolic activation, has been used to identify and prioritize the potential developmental hazards of contaminated surface waters, sediments, waste site soils, and industrial wastewater (Fort et al., 1995; 1996b; Fort and Stover, 1997). The rationale for use is based on the sensitivity of amphibian embryonic development to water quality. Based on this sensitivity, FETAX might be useful in estimating the chronic toxicity of a test substance to aquatic organisms (ASTM, 1998). FETAX also has potential for deriving water quality criteria for aquatic organisms, for studying bioavailability (ASTM, 1998), or for evaluating the efficacy of wastewater treatment procedures (Vismara et al., 1993).

12.2.2 Currently Accepted Water/Soil/Sediment Developmental Toxicity Testing Methods

FETAX is not currently accepted by U.S. Federal agencies. U.S. Federal and international regulations pertinent to the potential use of FETAX include the following:

Under the Clean Water Act, the EPA has developed guidance (40 CFR 132) that sets minimum water quality standards, policies, and implementation procedures for the Great Lakes System to protect aquatic life and wildlife. The methodology for collecting the required data requires acceptable acute or chronic tests with at least one species of freshwater animal in at least eight different families, including a family in the phylum Chordata (e.g., fish, amphibian). However, data from species that do not reproduce in the wild in North America are not acceptable. Although *X. laevis* is not native to North America, recent reports have indicated the occurrence of naturally reproducing populations in some portions of the United States (J. Burkhart, personal communication).

The EPA guidelines for evaluating whole effluent toxicity are provided in **Appendix 14**. This final rule amends the *Guidelines Establishing Test Procedures for the Analysis of Pollutants*, 40 CFR part 136, by adding methods for measuring the acute and short-term chronic toxicity of effluents and receiving waters.

EPA’s Significant New Alternatives Policy Program (40 CFR 82.170) identifies acceptable substitutes (compounds believed to present lower overall risks to human health and the environment) for ozone-depleting compounds. Under this program, ecotoxicological studies of a substitute and its components can include data from tests for effects on invertebrates, fish, or other animals.

12.2.3 Ability of FETAX to Assess Potential Developmental Hazards in Water/Soil/Sediment Samples

FETAX has been used to evaluate the potential developmental hazards of contaminated surface waters, sediments, waste site soils, and industrial wastewater (Fort et al., 1995; 1996b; Fort and Stover, 1997), to demonstrate the efficacy of wastewater treatment processes (Vismara et al., 1993), and to identify possible causes of malformations in natural frog populations in the United States (Burkhart et al., 1998; Fort et al., 1999a, b). However, very few comparative studies have been conducted to evaluate the relative ability of FETAX versus that of other similar bioassays to prioritize the hazard associated with contaminated water/soil/sediment samples (see **Section 16**).

12.2.4 Intended Range of Water/Soil/Sediment Samples Amenable to Test and Limits According to Physico—Chemical Factors

FETAX is considered to be applicable, with appropriate modifications, to aqueous effluents; surface and ground waters; leachates; aqueous extracts of water-insoluble materials; and solid-phase samples, such as soils and sediments, particulate matter, sediment, and whole bulk soils and sediments (ASTM, 1991; 1998). The test method is incompatible with materials (or concentrations of materials) that alter the pH, hardness, alkalinity, and conductivity of the FETAX Solution beyond the acceptable ranges specified by the ASTM FETAX Guideline (1991, 1998). Testing of solids is generally limited by the water solubility of the constituents. The effects of other physical/chemical properties (e.g., nitrate levels) on *Xenopus* embryonic development needs to be fully evaluated.

12.3 Section 12 Conclusions

The scientific basis for FETAX and its intended use(s) in ecotoxicology are adequately described. Test limits are defined but only limited information is available on the complete range of environmental samples amenable to test. The test method is incompatible with environmental samples that alter the pH, hardness, alkalinity, and conductivity of the FETAX

Solution beyond the acceptable ranges. Testing of solid environmental samples is generally limited by the water solubility of the constituents. The effects of other physico-chemical properties (e.g., nitrate levels) on *Xenopus* embryonic development needs to be evaluated.

13.0 FETAX TEST METHOD PROTOCOL

13.1 Standard Detailed Protocol

ASTM published a comprehensive guideline for FETAX in 1991 (**Appendix 10**); a revised guideline was published in 1998 (**Appendix 11**). The procedures presented in the ASTM FETAX Guideline (1991, 1998) are considered to be applicable, with modification, to the use of FETAX for conducting tests on the effects of surface and ground waters, solid phase samples such as soils and sediments, and whole bulk soils and sediments. Specific modifications relevant to the testing of water/soil/sediment samples (ASTM, 1991; 1998) are described briefly in the following subsections. Other aspects of the assay are described in **Section 2.1.1**.

13.1.1 Materials, Equipment, and Supplies

Information on materials, equipment, and supplies needed to support the standard FETAX assay, as described in the ASTM FETAX Guideline (1991, 1998), are presented in **Section 2.1.1** of this BRD.

13.1.2 Detailed Procedures for FETAX

General information on procedures for FETAX, including criteria for assay acceptance, as described in the ASTM FETAX Guideline (1991, 1998), are discussed in **Section 2.1.2** of this BRD. For water/soil/sediment samples, information on the identities and concentrations of major ingredients and major impurities, solubility and stability in water, the estimated toxicity to humans, and recommended safe-handling procedures will generally be lacking. However, at a minimum, the pH, hardness, alkalinity, and conductivity of such samples should be measured (ASTM, 1998).

The 1998 ASTM Guideline outlines procedures for solid phase sample testing (ASTM, 1998). Approximately one kilogram of soil or sediment should be collected and expediently sent to the laboratory to minimize holding time. Prior to testing, soil or sediment subsamples should be thoroughly homogenized. Subsamples are collected with a non-reactive sampling device and placed in a non-reactive storage container. Subsamples are mixed and stirred until texture and color are uniform. The samples are then stored at 4^{o}C until FETAX testing is initiated. It is recommended that samples be tested within two weeks of receipt unless specific circumstances delay testing

FETAX studies should be performed with the following modification for whole soil or sediment testing. Testing may be performed in 250 mL specimen bottles or similar capped vessels equipped with a 55 mL glass tube with Teflon mesh insert as the exposure chamber. For screening tests, 35 g of sediment (dry weight) should be placed in the bottom of the vessel, with the Teflon mesh insert added, and should be filled with 140 mL of FETAX Solution. It is essential that the dilution soil be non-toxic and as chemically and physically similar to the test soil as possible. Care must be taken in interpreting results of soil/sediment dilution experiments in that toxicity results may be altered because of the nature of the soil/sediment used for dilution. The sample must be equilibrated. The top edge of the glass tube must be higher than the water level to prevent larvae from swimming out after day two. This represents four parts of dilution water to one part of soil or sediment. Blastulae stage embryos are placed directly on the mesh insert that rests directly over the top of the soil or sediment in the sediment/water interface region. The test consists of 25 embryos placed in each of four replicates (total of 100 embryos exposed to FETAX Solution), a minimum of 25 embryos exposed to blasting sand (artificial sediment) in each of three replicates (minimum of 50 embryos total), and 25 embryos exposed to the soil or sediment sample in each of three replicates (minimum of 50 embryos total). Blasting or beach sand should be extensively tested beforehand to ensure that it produced less than 10% mortality or malformations after 96 hours. There should also be a reference soil/sediment tested that is non-toxic but represents the soil/sediment characteristics of the site. Dilutions of the soil or sediment should be prepared by mixing the sample with uncontaminated site soil or laboratory reference soil. Four to six dilutions ranging from 0 to 100% soil sample and a FETAX Solution control are typically tested. Each sample should be tested in triplicate. Solutions and soils or sediments should be changed every 24 hours of the four-day test by moving the insert containing the embryos to a fresh jar of diluent water and soil/sediment sample. Dead embryos are removed at this time. Dissolved oxygen and pH should be measured prior to renewal and in waste solutions from each successive day. Dissolved oxygen, pH, conductivity, hardness, alkalinity, ammonia-nitrogen, and residual chlorine should be measured on separate aliquots of the batches of FETAX Solution used during the study. The measurements must be conducted after the conclusion of the exposure period and oxygen content must be greater than 5.5 mg/L.

13.1.3 Dose-Selection Procedures—Screening Test

Screening tests (control and 100% sample) may be performed prior to multi-concentration definitive testing (ASTM, 1998). Determinations of LC_{50} and EC_{50} values are not possible with this approach and responses may be reported as a percent effect.

13.1.4 Endpoints Measured

The same three endpoints of mortality, malformations, and embryonic growth, as described in **Section 2.1.4** of this BRD are applicable to water/soil/sediment studies using FETAX.

13.1.5 Duration of Exposure

Information on appropriate exposure duration for FETAX, as described in the ASTM FETAX Guideline (1991, 1998), are presented in **Section 2.1.5** of this BRD.

13.1.6 Known Limits of Use

With appropriate modifications, FETAX can be used to conduct tests on aqueous effluents; surface and ground waters; leachates; aqueous extracts of water-insoluble materials; and solid-phase samples, such as soils and sediments, particulate matter, sediment, and whole bulk soils and sediments. The test method is incompatible with substances that alter the pH, hardness, alkalinity, and conductivity of the FETAX solution beyond the acceptable range specified by the ASTM FETAX Guideline (1991, 1998).

13.1.7 Nature of the Responses Assessed

Relevant information is provided in **Section 2.1.7** of this BRD.

13.1.8 Appropriate Vehicle, Negative, and Positive Controls

Relevant information on vehicle, negative, and positive controls are provided in **Section 2.1.8** of this BRD. In addition, blasting or beach sand (artificial sediment), extensively tested beforehand to ensure that it produced less than 10% mortality or malformations after 96 hours, should be included as one of the negative controls. There should also be a reference soil/sediment tested that is non-toxic but represents the soil/sediment characteristics of the site. Historically, FETAX studies using water/soil/sediment samples have not employed metabolic activation. Information on the use of 6-AN as a reference compound for FETAX is provided in **Section 2.1.8**. The need and/or usefulness of incorporating metabolic activation into FETAX studies with environmental samples has not been explored.

13.1.9 Acceptable Range of Negative and Positive Control Responses

Relevant information on the acceptable range of vehicle, negative, and positive control response are provided in **Section 2.1.9** of this BRD.

13.1.10 Data Collection

Data collection, as described in the ASTM FETAX Guideline (1991, 1999), is reviewed in **Section 2.1.10** of this BRD.

13.1.11 Data Storage Media

Information on data storage media is described in **Section 2.1.11** of this BRD.

13.1.12 Measures of Variability

Information on measures of variability, as described in the ASTM FETAX Guideline (1991, 1999), is described in **Section 2.1.12** of this BRD. However, formal evaluations of intra- and inter-laboratory variation in FETAX connected with its application to environmental samples have not been published. Also, no measures of variability for historical negative and positive control data were located.

13.1.13 Statistical and Non-Statistical Methods

Information on statistical and non-statistical methods for analyzing FETAX data are discussed in **Section 2.1.13** of this BRD. In a number of FETAX ecotoxicological studies, however, the magnitude of the response, as measured by the incidence of malformations only, has been used rather than a TI value or a MCIG/LC_{50} ratio, to assess relative developmental hazard (see **Section 16**). The decision criteria described for these studies are often based on non-statistical methods. For screening tests, statistical evaluation of differences in responses between the control and the single-treated group may be evaluated using parametric or non-parametric hypothesis tests for the mortality and malformation responses, and a grouped student's-test for the growth data (p-value <0.05 for all tests).

13.1.14 Decision Criteria

The decision criteria for FETAX, as described in the ASTM FETAX Guideline (1991, 1998), are described in **Section 2.1.4** of this BRD. However, as indicated in **Section 13.1.13**, these decision criteria are seldom used in studies involving environmental samples. Rather, relative activity based on the incidence of malformations seems to be the most common approach for evaluating the results of such studies.

13.1.15 Test Report Information

The information that should be included in the test report for an acceptable FETAX study, as described in the ASTM FETAX Guideline (1991, 1998), is summarized in **Section 2.1.15** of this BRD. Items specific to defined substances need not be considered. Information on dissolved oxygen, pH, conductivity, hardness, alkalinity, ammonia-nitrogen, and residual chlorine measured on separate aliquots of the batches of FETAX Solution used during the study should be provided. Additionally, specifics on sample moisture fraction determination and extract preparation, if conducted, are required.

13.2 Commonly Used Variations in the FETAX Standard Protocol and Rationale

13.2.1 Use of Alternative Species

Although FETAX was designed expressly for the use of *X. laevis*, it might be appropriate to use an endemic species when required by regulations or other considerations. Users are cautioned that many naturally occurring species of frogs are threatened by pollution and habitat loss and the user should carefully consider the ecotoxicological consequences of large-scale collection of local anuran species. Deviations from standard procedures must be reported. The ASTM FETAX Guideline (1991, 1998) states that it will be difficult to compare data from FETAX with data obtained using an alternative species. However, the sensitivity of *Rana pipiens* and *X. laevis* to several developmental toxicants may be quite similar (D. Fort, personal communication). Members of the family Ranidae (e.g., *R. pipiens*) and Bufonidae (e.g., *Bufo fowleri*) might be best suited for FETAX, because the number of eggs or the seasonal availability, or both, are less limited than for other species. Seasonal availability can be extended by two to three months using human chorionic gonadotropin injection. *R. catesbiena* and *B. americanus* are as well suited as *R. pipiens* and *B. fowleri*. High egg production, geographical range, short hatching periods, and other factors would indicate that these four species could serve as alternatives. Comparative sensitivities to inorganic mercury have been reported for some of these species (Birge and Black, 1979; Birge et al., 1979).

The ASTM FETAX Guideline (1991, 1998) suggested that reported differences in sensitivity to inorganic mercury should be taken into account when comparing data among amphibian species.

13.2.2 Additional Data and Alternative Exposure Protocols

Other types of data that can be collected in FETAX and alternative exposure protocols are discussed in **Section 2.3** of this BRD.

13.3 Basis for Selection of FETAX

The basis for selection of FETAX is discussed in **Sections 2.3** and **12.1.2** of this BRD.

13.4 Confidentiality of Information

Original data was not sought by NICEATM for any publication involved with the application of FETAX to the identification of developmental hazards in water/soil/sediment samples.

13.5 Basis for FETAX Decision Criteria

See **Section 2.5** for a discussion of the standard FETAX decision criteria. However, as indicated in **Section 13.1.13**, these decision criteria are seldom used in studies involving environmental samples. Rather, relative activity based on the incidence of malformations, seems to be the most common approach for evaluating the results of such studies. Relative activity was used because the studies evaluated were used generally to prioritize sites by relative importance for further investigation and/or remediation.

13.6 Basis for Numbers of Replicates and Repeat Tests in FETAX

In contrast to the ASTM FETAX assay (1991, 1998), most studies with environmental samples have used two, rather than three, definitive tests to define a FETAX study. The relative merit of two versus three replicate definitive tests has not been determined. Each definitive test is conducted using embryos from a different male/female pair of *X. laevis*. Each test consists of several different concentrations of the test substance with two replicate dishes at each test concentration and four replicate dishes for each control. Each dish contains 20 or 25 embryos. The number of embryos per dish, the number of replicate dishes per sample dilution, and the number of replicate tests per study have not been based on a formal scientific analysis.

13.7 Validation Study Based Modifications to the Standard Protocol

Published information on FETAX validation studies conducted using environmental samples was not located. Modifications to the standard FETAX protocol arising from validation studies employing defined chemicals are described in **Section 2.7** of this BRD.

13.8 Section 13 Conclusions

The 1991 and the revised and expanded 1998 FETAX Guideline published by ASTM are detailed, comprehensive, and well structured. Adequate information is provided on the necessary materials, equipment, and supplies; screening and definitive tests; endpoint (mortality, malformations, and embryonic growth) assessment; nature of the responses assessed; the duration of exposure; data collection and data storage media; measures of variability; statistical and non-statistical methods; test report information; commonly used protocol variations and rationale; the use of alternative species; and the basis for selection of FETAX.

Known limits of use for FETAX with water/soil/sediment samples were not described, except it was stated that the test method is incompatible with environmental samples that alter the pH, hardness, alkalinity, and conductivity of the FETAX Solution beyond the acceptable range specified by the ASTM FETAX Guideline (1991, 1998).

The three decision criteria used to distinguish between a teratogen and a non-teratogen in FETAX were well described in the ASTM FETAX Guideline (1991, 1998). However, as discussed in **Section 2.8** of this BRD, additional effort to evaluate and optimize the standard FETAX decision criteria appears to be warranted.

Selection of the number of embryos per dish (i.e., 20 or 25), the number of replicate dishes per test concentration (i.e., two), and the number of replicate tests per FETAX study (i.e., three) were based on the best scientific judgement of the developers/users of the assay at the time the ASTM FETAX Guideline (1991, 1998) was developed. However, FETAX studies with water/soil/sediments samples have been published based on the use of two replicate definitive tests only (see **Section 16**). A formal analysis of the relative power of FETAX based on two versus three identical definitive tests would be useful.

14.0 CHARACTERIZATION OF ENVIRONMENTAL WATER/SOIL/SEDIMENT SAMPLES TESTED IN FETAX

FETAX test data from 10 publications involving 124 water/soil/sediment samples were located, reviewed, extracted, and entered into the NICEATM FETAX Environmental Sample Database (**Appendix 8**). Sources for these data included peer-reviewed literature (including studies accepted for publication) and non peer-reviewed book chapters. Excluded from consideration was information provided in abstracts, manuscripts not accepted for publication, and publications that did not provide quantitative data.

14.1 Rationale for Water/Soil/Sediment Samples Selected for FETAX Validation Studies

Published FETAX validation studies involving water/soil/sediment samples were not located. Validation studies involving defined substances were discussed in **Section 3.0** of this BRD.

14.2 Rationale for the Number of Water/Soil/Sediment Samples Tested in FETAX Validation Studies

Not applicable.

14.3 Description of Chemical/Product Classes Evaluated in FETAX

Not applicable.

14.4 Coding Used in FETAX Validation Studies

Not applicable.

14.5 Section 14 Conclusions

Given that FETAX validation studies for evaluating developmental hazard in water/soil/sediment samples were not located, such an investigation is warranted. It is recommended that the ICCVAM Submission Guidelines be followed inn the design, conduct, and reporting of such studies.

15.0 REFERENCE DATA USED FOR AN ASSESSMENT OF FETAX PERFORMANCE CHARACTERISTICS

15.1 Description of Available Reference Data Sources

In some FETAX testing circumstances (e.g., ecotoxicological hazard assessment), reference data for laboratory mammals may not be appropriate. In other testing situations, reference data from naturally occurring anuran populations may prove useful. Such data were not located. Reference laboratory mammal teratogenic data were sought from the several general sources and databases described in **Section 4.1** of this BRD. With very few exceptions, teratogenic data for laboratory mammals exposed to water/soil/sediment samples were not located, while relevant data for humans was nonexistent.

15.2 Reference Data

The laboratory mammal reference data for environmental samples are provided in **Appendix 4**. Laboratory mammal (mouse, rat, or rabbit) teratogenicity data were obtained for one of the 124 water/soil/sediment samples evaluated in FETAX. Where available, descriptive information on the types of malformations observed was included in the database.

Appropriate reference data for non-mammalian aquatic species was limited to a direct comparison between FETAX and a developmental assay using *Pimephales promelas* (fathead minnow) in one sediment study (Dawson et al., 1988) and in two-related soil extract studies (Fort et al., 1995; 1996). Appropriate reference data for other species were not located.

15.3 Availability of Original Reference Test Data

The availability of original test data for the reference assays is not known.

15.4. Reference Data Quality

The quality of any reference data in terms of accuracy and whether the studies were conducted in compliance with national/international Good Laboratory Practice (GLP) Guidelines is not known.

15.5 Availability and Use of Human Teratogenicity Data

No human teratogenicity data for the water/soil/sediment samples tested in FETAX were located.

15.6 Section 15 Conclusions

It is suggested that future assessments of FETAX for evaluating developmental hazards in water/soil/sediment samples include tests on at least one reference species. Such studies should follow standardized protocols for the reference species. It is recommended also that GLP guidelines be followed.

16.0 FETAX TEST METHOD DATA AND RESULTS

16.1 Availability of a Detailed FETAX Protocol

As described in **Section 2** of this BRD, a comprehensive ASTM Guideline for FETAX was published in 1991 and a revised guideline suitable for testing environmental samples was published in 1998 (ASTM 1991; 1998). The 1998 ASTM FETAX Guideline is provided in **Appendix 11**.

16.2 Availability of Original and Derived FETAX Data

No attempt was made to obtain original data for any environmental sample study considered in this BRD.

16.3 Statistical and Non-Statistical Approaches used to Evaluate FETAX Data

The statistical and non-statistical methods used by the individual investigator to analyze FETAX data obtained are described in **Section 2.1.13**. In a number of environmental sample studies, however, the magnitude of the response, as measured by the incidence of malformations only, has been used rather than a TI value or a MCIG/LC_{50} ratio to assess relative hazard. The decision criteria described for these studies is also not based on statistical methods. For screening tests, statistical evaluation of differences in responses between the control and the single treated group may be evaluated using parametric or non-parametric hypothesis tests for the mortality and malformation responses, and a grouped t-test for the growth data.

16.4 FETAX Test Results for Individual Water/Soil/Sediment Samples Evaluated to Assess Developmental Toxicity

FETAX test data from 10 publications involving 124 samples were located, reviewed, extracted, and entered into the NICEATM Environmental Sample Database (**Appendix 8**). All 124 samples had been tested using FETAX without metabolic activation; no environmental sample was tested also with metabolic activation. All environmental samples were tested only once. Qualitative data on malformations observed in *X. laevis* were available for 36 chemicals (29%); quantitative malformation data were not provided for any study.

16.4.1 FETAX Testing with Water/Sediment Samples

In an early environmental study conducted by Dawson et al. (1985), discharges from abandoned lead and zinc mines were evaluated for their ability to induce malformations in FETAX. The discharges were characterized as having high concentrations of zinc, iron, and other metals, in addition to having a low pH and low oxygen content. Typical kinds of malformations induced included gut coiling, pericardial and ventral fin edema, microopthalmia, and tail kinking. The authors concluded that the observed effects were likely due to the observed alterations in oxygen, metal content, and pH.

In a related study, Dawson et al. (1988) evaluated the effects of extracts of metal-contaminated sediments and a reference metal toxicant (zinc sulfate) on the development of exposed *P. promelas* and *X. laevis*, using a standard FETAX protocol. Sediments from two contaminated stream sites were extracted with reconstituted culture water at various pH values for 24 hours, and evaluated for developmental toxicity. Developing *P. promelas* and *X. laevis* embryos were exposed for six and four days, respectively. The endpoints assessed were the EC_{50}, LC_{50}, TI, and MCIG. The investigators concluded that zinc was the major developmental toxicant in the sediment extracts, malformations were a more sensitive endpoint that growth inhibition, the pH used during extraction affected the toxicity of the extracted sample, and *P. promelas* was slightly more sensitive than *X. laevis.*

Contaminated groundwater is potentially hazardous to wildlife and humans (Bruner et al., 1998). Using FETAX, ground and surface water samples collected near a closed municipal landfill south of Norman, Oklahoma, demonstrated elevated developmental toxicity risk. More than 35 volatile and 40 semi-volatile and non-volatile compounds were identified in these samples. Many of the contaminants were known xenobiotics and carcinogens. Toxicity was significantly correlated with cumulative rainfall and relative humidity during the three days prior to sampling, but was negatively correlated with the weather conditions for days four to seven preceding sampling. Mortality was positively correlated with solar radiation and net radiation. The results suggest that solar radiation is low during rain events when toxicants are being diluted; however, toxicants are concentrated during periods of high solar radiation as evaporation increases.

Zaga et al. (1998) investigated the possible interaction between ultraviolet radiation and toxicants by examining the effects of exposure to a carbamate pesticide, carbaryl, and ultraviolet radiation on *X. laevis* embryos. The toxicity of 7.5 mg/L carbaryl increased by 10-fold in the presence of ultraviolet-B radiation, indicating photoenhancement of the toxicity of carbaryl. In another study, La Clair et al. (1998) found that a common insect growth regulator, S-methoprene, can react with sunlight, water, and microorganisms and disrupt the normal development of *X. laevis*. When embryos were exposed to S-methoprene degradates, malformations including eye defects and neural tube defects were observed.

Another environmental concern is the direct discharges from industries and municipal wastewater treatment plants. These wastewaters are complex mixtures that contain organic and inorganic compounds. Ciccotelli et al. (1998) investigated the biochemical alterations, such as glutathione-S-transferase (GST) activity, in *X. laevis* caused by exposure to various concentrations of wastewater from a treatment plant. Results of the investigation support the use of *X. laevis* in measuring biochemical alterations that serve as early indicators of environmental hazards. Vismara et al. (1993) used FETAX to evaluate a water purification system by testing the input and output waters from a chemical company for the presence of toxicants. Under the conditions of the test, the percentage of dead embryos following exposure to input water (i.e., untreated wastewater) was 100%, whereas the percentage of dead embryos following exposure to output water(i.e., treated wastewater) was 6.7%.

Malformations and abnormalities have been observed in various species of frogs inhabiting bodies of water throughout the United States. Malformations identified include missing and partial hind limbs, missing or misplaced eyes, microencephaly, ectromelia, ectrodactyly, and internal abnormalities (Fort et al., 1999a, b). A number of factors have been proposed as potential contributors to malformations in natural amphibian populations. These include the presence of developmental toxicants, ionic imbalances, nutritional deficiencies, mineral depletion (e.g., calcium and magnesium), disease (e.g., parasite infestation), UV radiation, and weather conditions (e.g., air temperature, humidity, rainfall) (Bruner et al., 1998; Fort et al., 1999a; Burkhart et al., 1998). In response to the general concern generated by the widespread prevalence of frog populations with a high incidence of malformations, Burkhart et al. (1998) conducted an extensive evaluation of water quality using FETAX. Water and sediment samples were collected from ponds in Minnesota with high incidences of frog malformations and from ponds with unaffected frog populations. Pond water from affected sites produced a high frequency of malformations and mortality in *X. laevis*. Removal of microbial contamination by boiling and filtration had no affect on the results. The teratogenic/toxic activities of the water samples were reduced or eliminated when samples were passed through activated carbon (Burkhart et al., 1998). The results of the studies excluded ion concentration, the presence of metals, and infectious organisms as causal factors of abnormal development, and suggested that the water contained one or more unknown agents that induce developmental abnormalities.

Fort et al. (1999b) used FETAX to evaluate the causal factors associated with developmental anomalies in *X. laevis* exposed to these pond water and aqueous sediment extracts. The craniofacial defects and abnormal eye and mouth development were reduced when some pond water and sediment extract samples underwent microfiltration and/or C_{18}-SPE treatments. Ion exchange was also effective in reducing the malformation-inducing activity of some samples. Results suggested that a mixture of naturally occurring compounds (e.g., pesticides) and anthropogenic organic compounds were primarily responsible for the abnormalities observed (Fort et al., 1999b).

16.4.2 FETAX Testing with Soil Samples

FETAX was used to assess the comparative environmental hazard of soil samples from two waste located in the state of Washington, U.S. (Fort et al., 1996). One waste site was contaminated with polycyclic aromatic hydrocarbons (PAHs), while the other was contaminated with heavy metals. An integrated hazard assessment study was conducted with the aqueous extracts of these samples using FETAX, the conventional *Pimephales promelas* 7-day teratogenicity test, and an abbreviated *P. promelas* teratogenicity test using the general FETAX protocol. Because inadequate sample volumes were available to perform definitive testing sufficient to define the EC_{50}, LC_{50}, or MCIG, the decision criteria used was based on rates of mortality, malformations, and growth inhibition. Zinc, copper, and pentachlorophenol were used as reference toxicants. Results from the studies with the aqueous soil extracts indicated that each of the two sample sites induced a contaminant-related increase in the rates of malformations and mortality in both species. Extracts from the site contaminated with PAHs tended to induce greater levels of embryo mortality in both species, with *P. promelas* being somewhat more sensitive. The types of malformations induced by the aqueous extracts from the PAH and heavy metal contaminated sites in *X. laevis* were characteristic of those induced by pentachlorophenol and zinc, respectively. Concurrent with these studies, a battery of bioassays, including lettuce seed (germination), earthworm (survival), *Daphnia* (survival) and larval *Pimephales* (survival) were also performed. In comparison with the other bioassays, the investigators concluded that FETAX and the *P. promelas* developmental toxicity test appeared to be the most predictive of the contaminated samples. Based on the results obtained, Fort et al. (1996) concluded that FETAX is useful as a component of a multi-testing approach to ecotoxicological hazard assessment.

In an extension of the Fort et al. (1996) study describe above, FETAX was used to evaluate the developmental toxicity of aqueous extracts of soil samples collected at six selected waste sites in the state of Washington (Fort et al., 1995). The waste sites included two sites contaminated with metals (copper, lead, zinc; and arsenic, lead, and mercury, respectively), one site contaminated with PAHs, two sites contaminated by petroleum products, and one site contaminated with organochlorine pesticides. Three to five samples from each site, representing baseline and increasing levels of contamination, were collected. Aqueous extracts of the soil samples were prepared and tested in FETAX.

FETAX was conducted in general compliance with the ASTM FETAX Guideline (ASTM, 1991). The concurrent controls consisted of FETAX Solution, whole blasting sand, whole reference soil, extracted blasting sand, and extracted reference soil. Because inadequate sample volumes were available to perform definitive testing sufficient to define the EC_{50}, LC_{50}, or MCIG, the decision criteria used was based on rates of mortality, malformations, and growth inhibition. Samples collected from the PAH- and petroleum product-contaminated sites were more toxic, although malformations were observed also. The metal-contaminated sites induced more malformations, but less toxicity, than the other sites. The organochlorine pesticide-contaminated site samples caused significant levels of embryonic deformities but not mortality.

Consistent with the other study (Fort et al., 1996), FETAX was concluded by the investigators to be more sensitive than other bioassays (lettuce seed, earthworm, *Daphnia*, larval *Pimephales*) in detecting ecotoxicological hazard. Fort et al. (1995) also concluded that FETAX was sensitive enough to detect low levels of developmental abnormalities but robust enough to be suitable for the testing of aqueous soil extracts.

In an effort to determine the significance of experimental design on the results of laboratory soil toxicity studies with FETAX, two different sample preparations were evaluated from three contaminated waste sites (Fort and Stover, 1997). Whole soil and aqueous soil extracts from each site were evaluated. Site 1 soil was characterized as loamy with a relatively high total organic carbon (TOC), moisture fraction (MF), and sulfide content. This site was contaminated with organochlorine pesticides. Site 2 soil, contaminated with PAHs and pentachlorophenol, was characterized as silt/clay with low/moderate TOC, MF, and sulfides. The Site 3 soil sample consisted of two separate sub-site samples. The first sub-site sample “a” was characterized as loamy with a relatively high TOC, moisture fraction (MF), and sulfide content. The second sub-site sample “b” was characterized as a mixture of silt/clay and sand with relatively low TOC, MF, and sulfide content. Both sub-site samples were contaminated with heavy metals, including copper, lead, and zinc.

The FETAX studies followed the ASTM FETAX Guideline (ASTM, 1991). The concurrent controls consisted of FETAX Solution, whole blasting sand, whole reference soil, extracted blasting sand, and extracted reference soil. A FETAX response was considered indicative of developmental hazard if the TI value was greater than 1.5 or if the MCIG/LC_{50} ratio was less than 0.30. Types of malformations induced were also considered. FETAX testing of the Site 1 sample indicated that substantially greater levels of developmental toxicity were induced by the aqueous extraction of the sample than by the whole bulk soil. Tests with Site 2 samples suggested that both the aqueous extract and the whole bulk sample were capable of inducing comparable rates of developmental toxicity. Tests with sub-site sample "a" of Site 3 indicated that the aqueous extract of the sample induced greater levels of developmental toxicity than the whole soil sample. Toxicity tests with sub-site sample "b" produced variable results that seemed to suggest that the aqueous extract induced greater toxicity than the whole bulk preparations. Fort and Stover (1997) concluded that the results from these studies suggested the importance of experimental design in evaluating potential ecological hazards of contaminated soils, particularly to in regard to amphibian species.

16.4.3 FETAX Testing with Other Environmental Samples

The method was considered useful by Dumont et al. (1983) for determining the teratogenic/embryotoxic potency for a group of chemical mixtures with similar composition that may be expected to be found as environmental pollutants. Tests with five fossil fuel mixtures were conducted using FETAX. The mixtures included a coal-derived fuel oil (Comparative Research Material [CRM]-1), a shale-derived crude (CRM-2), a coal gasifier electrostatic precipitator tar (CRM-4), an aromatic natural petroleum crude (CRM-3), and an aliphatic natural petroleum crude (CRM-5). The experiments were conducted using dilutions of stock solutions made of aqueous extracts of the material. Based on the EC_{50} values, the coal-derived materials were the most teratogenic followed by shale-derived, aromatic, and aliphatic petroleums (**Table 37)**. With respect to embryolethality, CRM-1 and CRM-4 were the most toxic with LC_{50} values of 1.48% and 0.83% respectively followed by CRM-2 with a LC_{50} of 6.97%. CRM-3 and CRM-5 were essentially non-toxic.

These studies demonstrate the utility of FETAX in ecotoxicological hazard assessment, as a means for detecting and prioritizing sites with increased developmental risks. To increase the validity of the interpretation of such data, it would be useful to further evaluate the influence of the physico-chemical properties of environmental samples on the frequency of malformations in

Table 38. Summary of Results for Complex Fossil Fuel Mixtures using FETAX (Dumont et al., 1983).

Comparative Research Material (CRM)	96 Hour Results		TI	Observed Effects
	LC_{50} (%)	EC_{50} (%)		
CRM-1	1.48	0.96	1.54	Growth = 81%; developmental stage attained = 44; pigmentation and motility reduction
CRM-2	6.97	3.36	2.07	Growth = 96%; developmental stage attained = 45/46; pigmentation and motility reduction
CRM-3	33.38	31.10	1.07	Growth = 90%; developmental stage attained = 46
CRM-4	0.83	0.48	1.73	Growth = 87%; developmental stage attained = 45/47; pigmentation and motility reduction
CRM-5 (90% aqueous extract)	--	--	<1.00	Growth = 100%; developmental stage attained = 46/47

FETAX. Additionally, further research is needed on the utility of the current FETAX protocol as an effective assay for assessing water and sediment quality and detecting changes that can have adverse effects on the ecosystem.

16.5 Use of Coded Environmental Samples and Compliance with GLP Guidelines

Generally, coded water/soil/samples were used for ease of identification and chain of custody. These studies were not conducted in compliance with national or international GLP guidelines, nor were they generally conducted at facilities at which GLP studies are normally conducted.

16.6 Availability of Non-Audited FETAX Data

Original data was not sought by NICEATM for any FETAX study involving environmental samples.

16.7 Section 16 Conclusions

Based on the studies evaluated, FETAX appears to be useful in ecotoxicological hazard assessment, and as a means for detecting and prioritizing sites with increased developmental risks. Studies including other bioassays as part of a battery with FETAX (i.e., lettuce seed, earthworm, *Daphnia*, larval *Pimephales*) indicated that FETAX was sensitive enough to detect low levels of developmental abnormalities, but robust enough to be suitable for testing aqueous soil extracts. A comprehensive ASTM protocol for use in such assessments is available (ASTM, 1998). To increase the validity of the interpretation of such data, it would be useful to further evaluate the influence of the physico-chemical properties of environmental samples on the frequency of malformations in FETAX. Additionally, further research on the performance of the current FETAX protocol as an effective assay for assessing water and sediment quality and detecting changes that can have adverse effects on the ecosystem may provide further insight that could optimize ecotoxicological assessments. It may also be helpful to further evaluate how FETAX could best fit into a test battery for prioritizing of sites for further testing and remediation.

17.0 PERFORMANCE CHARACTERISTICS OF FETAX WITH WATER/ SOIL/SEDIMENT SAMPLES

Given the lack of sufficient reference data for comparison, the performance characteristics of FETAX, based on tests conducted using water/soil/sediment samples, could not be determined. Appropriate reference data for non-mammalian aquatic species was limited to a direct comparison between FETAX and a developmental assay using *P. promelas* in one sediment study (Dawson et al., 1988) and in two-related soil extract studies (Fort et al., 1995; 1996). Both species gave similar results, with *P. promelas* exhibiting slightly greater sensitivity in the sediment study (Dawson et al., 1988).

18.0 TEST METHOD RELIABILITY (REPEATABILITY/ REPRODUCIBILITY)

Due to the lack of validation studies to evaluate FETAX reliability with water/soil/sediment samples, an assessment of test method reliability with environmental samples could not be conducted.

18.1 Summary of Historical Positive and Negative Control Data

No historical control data for FETAX studies conducted with water/soil/sediment samples were located.

18.2 Limitations of FETAX in Regard to Test Method Reliability

Limitations in regard to test method reliability for studies conducted with environmental samples could not be identified.

18.3 Data Interpretation Issues

One potential issue affecting data interpretation connected with water/soil/sediment samples is the lack of an exogenous MAS incorporated into the FETAX assay. This is relevant when results are being extrapolated to estimate effects on adult organisms of the same species. Other data interpretation issues include the appropriateness of data handling and processing, and the lack of reference data for comparison.

18.4 Section 18 Conclusions

To provide data for evaluation, a validation study designed to evaluate the ecotoxicological applicability of FETAX would be helpful. Such a study should include assessments by several laboratories, and should include the testing of both common samples and environmental samples collected independently. Data from at least one reference species should be collected, and the study design should include sites with different gradients of developmental toxicity. Studies focusing on data interpretation issues could also be helpful in further optimizing the assay. Potential issues to address include the decision criteria used for ranking samples in regard to developmental hazard, and the appropriateness of sample handling and processing techniques.

18.0 TEST METHOD RELIABILITY (REPEATABILITY/ REPRODUCIBILITY)

Due to the lack of validation studies to evaluate FETAX reliability with water/soil/sediment samples, an assessment of test method reliability with environmental samples could not be conducted.

18.1 Summary of Historical Positive and Negative Control Data

No historical control data for FETAX studies conducted with water/soil/sediment samples were located.

18.2 Limitations of FETAX in Regard to Test Method Reliability

Limitations in regard to test method reliability for studies conducted with environmental samples could not be identified.

18.3 Data Interpretation Issues

One potential issue affecting data interpretation connected with water/soil/sediment samples is the lack of an exogenous MAS incorporated into the FETAX assay. This is relevant when results are being extrapolated to estimate effects on adult organisms of the same species. Other data interpretation issues include the appropriateness of data handling and processing, and the lack of reference data for comparison.

18.4 Section 18 Conclusions

To provide data for evaluation, a validation study designed to evaluate the ecotoxicological applicability of FETAX would be helpful. Such a study should include assessments by several laboratories, and should include the testing of both common samples and environmental samples collected independently. Data from at least one reference species should be collected, and the study design should include sites with different gradients of developmental toxicity. Studies focusing on data interpretation issues could also be helpful in further optimizing the assay. Potential issues to address include the decision criteria used for ranking samples in regard to developmental hazard, and the appropriateness of sample handling and processing techniques.

19.0 TEST METHOD DATA QUALITY

19.1 Extent of Adherence to GLP Guidelines

Ecotoxicological studies were not conducted in compliance with national or international GLP guidelines, nor were they generally conducted at facilities at which GLP studies are normally conducted.

19.2 Results of Data Quality Audits

No data audits were conducted on studies testing environmental samples.

19.3 Impact of GLP Deviations and/or Data Audit Non-Compliance

Not applicable.

19.4 Section 19 Conclusions

Ecotoxicological studies were not conducted in compliance with national or international GLP guidelines, nor were they generally conducted at facilities at which GLP studies are normally conducted. No data audits were conducted on studies testing environmental samples. It might be useful to compare the original sampling data from the ecotoxicological studies conducted with FETAX with the data provided in the published reports.

20.0 OTHER SCIENTIFIC REPORTS AND REVIEWS

20.1 Availability of Other Relevant FETAX Data

The focus of this section of the BRD was on the use of FETAX as an assay for evaluating the development hazard of water/soil/sediment samples. The sources for the FETAX data evaluated for that purpose included peer-reviewed literature (including studies accepted for publication) and non peer-reviewed book chapters. Information provided in abstracts, manuscripts not accepted for publication, and studies not conducted in general compliance with the ASTM FETAX Guideline (1991, 1998) were excluded from consideration.

20.2 Conclusions of Other Peer Reviews

No independent peer reviews of FETAX as applied to environmental samples were located.

20.3 Section 20 Conclusions

No independent peer reviews of FETAX as an ecotoxicological assay were located.

21.0 ANIMAL WELFARE CONSIDERATIONS

21.1 Extent to which FETAX Will Reduce, Refine, or Replace Animal Use for Ecotoxicological Hazard Assessment using Water/Soil/Sediment Samples

Multiple species are generally used in ecotoxicology studies. Use of this *in vitro* assay would reduce reliance on tests involving adult organisms.

22.0 OTHER CONSIDERATIONS

22.1 Test Method Transferability

22.1.1 Facilities and Major Fixed Equipment

Information on the facilities and major fixed equipment needed for FETAX are provided in **Section 11.1.1** of this BRD.

22.1.2 Required Level of Personnel Training and Expertise

Information on the level of personnel training and expertise needed for FETAX are provided in **Section 11.1.2** of this BRD.

22.1.3 General Availability of Necessary Equipment and Supplies

The equipment and supplies needed to conduct FETAX are readily available from any major supplier.

22.2 Assay Costs

Assay costs for a complete FETAX study, without metabolic activation only, following the ASTM FETAX Guideline (1998) and conducted in compliance with national/international GLP guidelines, should cost less than $12,500 per test substance (D. Fort, personal communication). No attempt was made to obtain costs for other biological-based assays used to assess developmental hazards in water/soil/sediment samples.

22.3 Time Needed to Conduct the Test

A complete FETAX study, without metabolic activation, following the ASTM FETAX Guideline (1998) and conducted in compliance with national/international GLP guidelines, would require less than two months to complete.

22.4 Section 22 Conclusions

Sufficient information on facilities and equipment for setting up FETAX is provided in the ASTM FETAX Guideline (1991, 1998). The estimated amount of technical training required for conducting the in-life portion of a FETAX study appears to be sufficient. However, based on concerns the level of expertise needed for the proper identification of malformations induced in *Xenopus* embryos, more intensive training may be needed for this aspect of the assay. The projected cost and study duration for a GLP compliant complete FETAX study, without metabolic activation, following the ASTM FETAX Guideline (1998), appears to be reasonable.

23.0 OTHER APPLICATIONS FOR *XENOPUS*

23.1 *Xenopus* Tail Resorption Assay

The *Xenopus* Tail Resorption Assay is an endocrine (thyroid) disruption assay using advanced *Xenopus* larvae to screen materials that may disrupt thyroid function (Fort et al., 2000b). In this assay, tadpoles are exposed for approximately 14 days from Developmental Stages between 58 and 60 through Developmental Stage 66. Ten tadpoles at the "just bud" stage (average tail length between four to five centimeters) are exposed to varying concentrations of the test material. Test organisms are fed twice daily and dead organisms removed and mortality counted. At a minimum, all solutions are renewed and exposure containers cleaned on Days 4, 7, and 10. Photographic images of the test organisms are taken on Days 0, 4, 7, 10, and 14 and tail length determined. At the completion of the 14-day exposure period, any malformed organisms are maintained for further observation. All other test organisms are euthanized. Data on gross effects on tail resorption are collected. The test period may be extended if tail resorption in the control organisms is not complete at the end of the 14 days. Also, the renewal and photographic days may be modified as necessary to meet differing test requirements.

In a recent evaluation of this assay (Fort et al., 2000b), short-term static-renewal studies were performed on *X. laevis* embryos with 16 selected test materials from day 50 (Developmental Stage 60) to day 64 (Developmental Stage 66) (14-day test). The test materials were 6-AN, acetyl hydrazide, cadmium, copper, endosulfan, iodine, lindane, methimazole, methoprene, nonylphenol, pentachlorophenol, perchlorate, propylthiouracil, semicarbizide, thyroxin, and triiodothyronine (T_3). Of these 16 test materials, ten (acetyl hydrazide, cadmium, endosulfan, lindane, methimazole, methoprene, pentachlorophenol, perchlorate, propylthiouracil, and semicarbizide) were found to significantly inhibit the rate of tail resorption. Four test materials (iodine, nonylphenol, thyroxin, and T_3) were found to stimulate metamorphosis. Two test materials (6-AN and copper) had no appreciable effect on the rate of metamorphosis. In an effort to determine if the morphological effects observed were related to an alteration in thyroid activity, measurement of T_3 in the treated embryos, and co-administration studies using thyroxine (agonist) or propylthiouracil (antagonist) were performed based on the morphological response noted during tail resorption. Of the ten compounds found to inhibit the rate of tail resorption, eight were found to reduce the levels of T_3. In each case, the inhibitory response could be at least partially alleviated by the co-administration of thyroxine. Larvae exposed to the four stimulatory agents had somewhat elevated levels of T_3 and were responsive to propylthiouracil antagonism. Twelve of the 14 compounds tested in this study that altered the rate of tail resorption did so via the thyroid axis. The investigators concluded that *Xenopus* might be a suitable system for evaluating the impact of environmental agents and chemical products on thyroid function.

This methodology has been used to evaluate the effects of two sulfonylurea herbicides, sulfometuron methyl and nicosulfan, on tail resorption (Fort et al., 1999c). The analytically impure, but not the pure, sulfonylurea herbicides slowed tail resorption rates significantly. Also, Fort et al. (1999a) demonstrated that pond water, sediment, and sediment extracts from ponds inhabited by mal-developed frogs was capable of inhibiting tail resorption in *Xenopus*.

23.2 *Xenopus* Vitellogenin Assay

Another endocrine disruption assay involving *Xenopus* is based on the detection of vitellogenin in the blood of treated males (Palmer and Palmer, 1995). One of the most important and sensitive responses to estrogen is the upregulation of protein production. A particularly well known estrogenic response in all oviparous and ovoviviparous vertebrates is the induction of the lipoprotein vitellogenin in liver cells. In females, vitellogenin is transported in the blood to the ovaries, where it is incorporated into the developing ovarian follicles as yolk. Due to their normally low levels of endogenous estrogens, male *X. laevis* have no detectable levels of vitellogenin in blood. However, their liver is capable of synthesizing and secreting vitellogenin into the blood in response to exogenous estrogen stimulation. In the *Xenopus* Vitellogenin assay, adult males are given intraperitoneal injections of a test substance daily for 7 days, and plasma is collected on day 14 (Palmer and Palmer, 1995). The estrogenic activity of the test substance is determined by measuring the induction of plasma vitellogenin. Vitellogenin is identified by precipitation, electrophoresis, Western blot, and enzyme-linked immunosorbant assay (ELISA).

Vitellogenin may prove useful as a biomarker in *Xenopus* for identifying xenobiotics with estrogenic activity. The test is relatively noninvasive, requiring only small (microliter) quantities of plasma or serum. It can be used in the laboratory to identify substances with *in vivo* estrogenic activity and *in situ* to indicate the presence of environmental pollutants with estrogenic activity. The expression of vitellogenin is through known physiological and biochemical pathways. The induction of vitellogenin is sensitive to any estrogenic contaminant, and the response is quantifiable. Finally, the assay for vitellogenin can be performed relatively easily and inexpensively. A major limitation of this assay is that it provides no direct information regarding the female or developing embryo. However, if estrogen receptors are being stimulated in the liver of males, receptors in other organs such as the testes and prostate gland of males and reproductive tissues of females and embryos may likewise be affected. Also, in ecotoxicological studies, vitellogenin production does not indicate what substance(s) may be causing the effect. However, the assay may be used as a rapid, sensitive, and economical initial screen, followed (as indicated by positive vitellogenic responses) by more costly screens to identify the specific contaminating substances.

23.3 Evaluation of Reproductive Toxicity using *Xenopus*

Fort et al. (1999d) has evaluated the utility of *X. laevis* for assessing reproductive toxicity. Cadmium, boric acid, and ethylene glycol monomethyl ether (EGME) were evaluated for reproductive and developmental toxicity in *X. laevis.* Eight reproductively mature adult male and eight superovulated female *X. laevis* were exposed to at least five separate sublethal concentrations of each material via the culture water for 30 days. Four respective pairs were mated and the offspring evaluated for developmental effects; an evaluation of reproductive status was performed on the remaining four specimens. Ovary health, oocyte count, oocyte maturity and maturation capacity, and necrosis were evaluated in the female, while testis health, sperm count, dysmorphology, and motility were studied in the male. Based on this assessment, each test material exerted reproductive toxicity in *X. laevis*, but with varying potencies. The investigators concluded that this model appears to be a useful tool in the initial assessment and prioritization of potential reproductive toxicants for further testing. In a related series of studies, low boron levels in the diet were associated with adverse reproductive performance (Fort et al. 1999e, f, g).

23.4 *Xenopus* Limb Bud Assay

The *Xenopus* Limb Bud Assay is a test method for exploring limb mal-development, including possible mechanisms of action (D. Fort, personal communication). This assay is proposed as a model for screening materials that may cause limb deformities in the workplace or the environment. The assay uses blastula stage *Xenopus* embryos raised to about Developmental. Stage 58 to 59. The first four days of the test are similar to the standard FETAX test. However, at the end of the 96-hour exposure period, the developing embryos are transferred to larger containers. Chemical solution renewal and tub cleaning are reduced to a minimum of two times per week. However, the frequency of renewals may be increased if the test chemical in solution is easily degraded or volatilized. The pH is maintained between 7.8-8.0, and each container is aerated throughout the remainder of test. Test organisms are fed twice daily; dead organisms are removed and mortality counted. The length of time to complete the assay will vary and is dependent on the rate of hind limb development in the control sets (generally 45 to 60 days). The assay can be stopped when greater than 80% of the control organisms reach Developmental Stage 58 to 59 with developed hind limbs (femur, tibia, fibula, and foot with digits or toes and the beginning formation of claws). At the end of the exposure period, the incidence of malformations, survival, and total organism counts are determined. This methodology has been used to identify agents associated with limb mal-development in pond water and sediment collected in Minnesota, U.S. (Fort et al., 1999a), to evaluate the developmental toxicity of thalidomide (Fort et al., 2000c), and to evaluate the effects of two sulfonylurea herbicides, sulfometuron methyl and nicosulfan, on limb development (Fort et al., 1999c). The analytically impure, but not the pure, sulfonylurea herbicides induced abnormal limb development.

23.5 Section 23 Conclusions

Other tests using *Xenopus* are being evaluated for their ability to identify substances or environmental samples that may disrupt endocrine function (the *Xenopus* Tail Resorption Assay, Vitellogenin Assay), for assessing reproductive toxicity, and for exploring limb mal-development, including possible mechanisms of action (*Xenopus* Limb Bud Assay). These developing test methods require appropriate validation.

24.0 REFERENCES

16 CFR 150.135—US Consumer Product Safety Commission. Code of Federal Regulations. Title 16, Commercial Practices. Chapter II, Consumer Product Safety Commission. Part 1500—Hazardous substances and articles; administration and enforcement regulations. Subpart 135. Summary of guidelines for determining chronic toxicity.

40 CFR 79.63—U.S. Environmental Protection Agency. Code of Federal Regulations. Title 40, Subchapter C—Air Programs, Part 79, Registration of fuel and fuel additives, Subpart F—Testing Requirements for Registration, Section 79.63 Fertility assessment/teratology.

40 CFR 82.170—U.S. Environmental Protection Agency. Code of Federal Regulations. Title 40, Subchapter C—Air Programs, Part 82, Protection of stratospheric ozone, Subpart D—Significant New Alternatives Policy Program, Section 82.170 Purpose and scope.

40 CFR 132—U.S. Environmental Protection Agency. Code of Federal Regulations. Title 40, Subchapter D—Water Programs, Part 132, Water Quality Guidance for the Great Lakes System.

40 CFR 136—U.S. Environmental Protection Agency. Code of Federal Regulations. Title 40, Subchapter D—Water Programs, Part 199, Whole Effluent Toxicity: Guidelines Establishing Test Procedures for the Analysis of Pollutants.

40 CFR 158.202—U.S. Environmental Protection Agency. Code of Federal Regulations. Title 40, Subchapter E—Pesticide Programs, Part 158, Data Requirements for Registration, Subpart D—Data Requirement Tables, Section 158.202 Purposes of the registration data requirements.

40 CFR 158.340—U.S. Environmental Protection Agency. Code of Federal Regulations. Title 40, Subchapter E—Pesticide Programs, Part 158, Data Requirements for Registration, Subpart D—Data Requirement Tables, Section 158.340 Toxicology data requirements.

40 CFR 795.250—U.S. Environmental Protection Agency. Code of Federal Regulations. Title 40, Subchapter R—Toxic Substances Control Act, Part 795, Provisional Test Guidelines, Subpart D—Provisional Health Effects Guidelines, Section 795.250, Developmental neurotoxicity screen.

40 CFR 798.4350—U.S. Environmental Protection Agency. Code of Federal Regulations. Title 40, Subchapter R—Toxic Substances Control Act, Part 798, Health Effects Testing Guidelines, Subpart E—Specific Organ/Tissue Toxicity, Section 798.4350 Inhalation developmental toxicity study.

40 CFR 798.4700—U.S. Environmental Protection Agency. Code of Federal Regulations. Title 40, Subchapter R—Toxic Substances Control Act, Part 798, Health Effects Testing Guidelines, Subpart E—Specific Organ/Tissue Toxicity, Section 798.4700 Reproduction and fertility effects.

40 CFR 798.4900—U.S. Environmental Protection Agency. Code of Federal Regulations. Title 40, Subchapter R—Toxic Substances Control Act, Part 798, Health Effects Testing Guidelines, Subpart E—Specific Organ/Tissue Toxicity, Section 798.4900 Developmental toxicity study.

40 CFR 799.9370—U.S. Environmental Protection Agency. Code of Federal Regulations. Title 40, Subchapter R—Toxic Substances Control Act, Part 799, Identification of Specific Chemical Substance and Mixture Testing Requirements, Subpart H—Health Effects Test Guidelines, Section 799.9370 TSCA prenatal developmental toxicity.

40 CFR 799.9380—U.S. Environmental Protection Agency. Code of Federal Regulations. Title 40, Subchapter R—Toxic Substances Control Act, Part 799, Identification of Specific Chemical Substance and Mixture Testing Requirements, Subpart H—Health Effects Test Guidelines, Section 799.9380 TSCA reproduction and fertility effects.

ASTM (American Society for Testing and Materials). 1991. Standard Guide for Conducting the Frog Embryo Teratogenesis Assay—*Xenopus* (FETAX). ASTM E1439—91. In: Annual Book of ASTM Standards, Philadelphia.

ASTM (American Society for Testing and Materials). 1992. Standard Practice of Conducting an Interlaboratory Study to Determine the Precision of a Test Method. ASTM E691—92. In: Annual Book of ASTM Standards, Philadelphia.

ASTM (American Society for Testing and Materials). 1998. Standard Guide for Conducting the Frog Embryo Teratogenesis Assay—*Xenopus* (FETAX). ASTM E1439—98. In: Annual Book of ASTM Standards, Philadelphia.

Bantle, J.A. 1995. FETAX – A Developmental Toxicity Assay Using Frog Embryos. In: Fundamentals of Aquatic Toxicology, 2nd ed., G. M. Rand (Ed.), Taylor and Francis, Washington, D.C., pp. 207-230.

Bantle, J.A., D.J. Fort, and B.L. James. 1989. Identification of developmental toxicants using the Frog Embryo Teratogenesis Assay—*Xenopus* (FETAX). Hydrobiologia 188/189:577-585.

Bantle, J.A., D.J. Fort, J.R. Rayburn, D.J. DeYoung, and S.J. Bush. 1990. Further validation of FETAX: Evaluation of the developmental toxicity of five known mammalian teratogens and non-teratogens.

Bantle, J.A., D.T. Burton, D.A. Dawson, J.N. Dumont, R.A. Finch, D.J. Fort, G. Linder, J.R. Rayburn, D. Buchwalter, M.A. Maurice, and S.D. Turley. 1994a. Initial interlaboratory validation study of FETAX: Phase I testing. J. Appl. Toxicol. 14:213-223.

Bantle, J.A., D.T. Burton, D.A. Dawson, J.N. Dumont, R.A. Finch, D.J. Fort, G. Linder, J.R. Rayburn, D. Buchwalter, A.M. Gaudet-Hull, M.A. Maurice, and S.D. Turley. 1994b. FETAX interlaboratory validation study: Phase II testing. Environ. Toxicol. Chem. 13:1629-1637.

Bantle, J.A., R.A. Finch, D.T. Burton, D.J. Fort, D.A. Dawson, G. Linder, J.R. Rayburn, M. Hull, M. Kumsher-King, A. M. Gaudet-Hull, and S.D. Turley. 1996. FETAX interlaboratory validation study: Phase III, part 1 testing. J. Appl. Toxicol. 16:517-528.

Bantle, J.A., J.N. Dumont, R.A. Finch, G. Linder, and D.J. Fort. 1998. Atlas of Abnormalities: A Guide for the Performance of FETAX, 2nd ed.. Oklahoma State University Press, Stillwater, OK, 68 pp.

Bantle, J.A., R.A. Finch, D.J. Fort, E.L. Stover, M. Hull, M. Kumsher-King, and A.M. Gaudet-Hull. 1999. Phase III interlaboratory study of FETAX, Part 3—FETAX validation using 12 compounds with and without an exogenous metabolic activation system. J. Appl. Toxicol. 19(6):447-472.

Birge, W.J., and J.A. Black 1979. Effects of Copper on Embryonic and Juvenile Stages of Aquatic Animals. In: Copper in the Environment, Pt II. Health Effects, J.O. Nriagu (Ed.), Wiley, New York, pp. 373-399.

Birge, W. J., J.A. Black, A.G. Westerman, and J.E. Hudson 1979. Effects of Mercury on Reproduction of Fish and Amphibians. In: The Biogeochemistry of Mercury in the Environment. Elsevier/North-Holland Biomedical Press, J.O. Nriagu (Ed.), pp. 629-655

Brown, N.A. 1987. Teratogenicity testing in vitro: Status of validation studies. Arch. Toxicol. Suppl. 11:105.

Brown, N.A., H. Spielmann, R. Bechter, O.P. Flint, S.J. Freeman, R.J., Jelinek, E. Koch, H. Nau, D.R. Newall, A.K. Palmer, J.-Y Renault, M. Repetto, R. Vogel, and R. Wiger. 1995. Screening chemicals for reproductive toxicity: The current alternatives. The report and recommendations of an ECVAM/ETS workshop (ECVAM workshop 12). ATLA 23:868-882.

Bruner, M.A., M. Rao, J.N. Dumont, M. Hull, T. Jones, and J.A. Bantle. 1998. Ground and surface water developmental toxicity at a municipal landfill: Description and weather-related variation. Ecotoxicol. Environ. Saf. 39:215-226.

Budavari, S. (Ed.). 1996. The Merck Index, 12th ed. Merck & Co., Inc., Whitehouse Station, NJ.

Burkhart, J.G., J.C. Helgen, D.J. Fort, K. Gallagher, D. Bowers, T. L Propst, M. Gernes, J. Magner, M. D. Shelby, and G. Lucier. 1998. Induction of mortality and malformation in Xenopus laevis embryos by water sources associated with field frog deformities. Environ. Health Perspect. 106(12):841-848.

Ciccotelli, M., S. Crippa, and A. Colombo. 1998. Bioindicators for toxicity assessment of effluents from a wastewater treatment plant. Chemosphere 37(14-15):2823-2832.

Copp, A.J., F.A. Brook, P. Estibeiro, A.S.W. Shum, and D.L. Cockroft. 1990. The embryonic development of mammalian neural tube defects. Prog. Neurobiol. 35:363-403.

Courchesne, C.L., and J.A. Bantle. 1985. Analysis of the activity of DNA, RNA, and protein synthesis inhibitors on *Xenopus* embryo development. Teratogen. Carcinogen. Mutagen. 5:177-193.

Dawson, D.A. 1991. Additive incidence of developmental malformation for *Xenopus* embryos exposed to a mixture of ten aliphatic carboxylic acids. Teratology 44(5):531-546.

Dawson, D.A., and J.A. Bantle. 1987. Development of a reconstituted water medium and preliminary validation of the frog embryo teratogenesis assay—Xenopus. J. Appl. Toxicol. 7(4):237-244.

Dawson, D.A., and T.S. Wilke. 1991a. Initial evaluation of developmental malformation as an end point in mixture toxicity hazard assessment for aquatic vertebrates. Ecotoxicol. Environ. Safety. 21(2):215-226.

Dawson, D.A., and T.S. Wilke. 1991b. Evaluation of the Frog Embryo Teratogenesis Assay: *Xenopus* (FETAX) as a model system for mixture toxicity hazard assessment. Environ. Toxicol. Chem. 10(7):941-948.

Dawson, D.A., C.A. McCormick, and J.A. Bantle. 1985. Detection of teratogenic substances in acidic mine water samples using the Frog Embryo Teratogenesis Assay—*Xenopus* (FETAX). J. Appl. Toxicol 5(4):234-244.

Dawson, D.A., E.F. Stebler, S.L. Burks, and J.A. Bantle. 1988. Evaluation of the developmental toxicity of metal-contaminated sediments using short-term fathead minnow and frog embryo-larval assays. Environ. Toxicol. Chem. 7:27-34.

Dawson, D.A., D.J. Fort, G.J. Smith, D.L. Newell, and J.A. Bantle. 1988b. Evaluation of the developmental toxicity of nicotine and cotinine with frog embryo teratogenesis assay: Xenopus. Teratogen. Carcinogen. Mutagen. 8:329-338.

DeYoung, D.J., J.A. Bantle, and D.J. Fort. 1991. Assessment of the developmental toxicity of ascorbic acid, sodium selenate, coumarin, serotonin, and 13-cis retinoic acid using FETAX. Drug Chem. Toxicol. 14:127-141.

Dumont, J.N., T.W. Schultz, M.V. Buchanon, and G.L. Kao. 1982. Frog embryo teratogenesis assay: Xenopus (FETAX)--A short-term assay applicable to complex environmental mixtures. Government Reports Announcements and Index. Number 22. NTIS.

Dumont, J.N., T.W. Schultz, M.V. Buchanan, and G.L. Kao. 1983. Frog Embryo Teratogenesis Assay: Xenopus (FETAX)—A short-term assay applicable to complex environmental mixtures. In “Short-term Bioassays in the Analysis of Complex Environmental Mixtures III." M. Waters, S. Sandhu, J. Lewtas, L. Claxton, N. Chernoff, and S. Nesnow (Eds.), Plenum, New York, NY, pp. 393-405.

Dresser, T.H., E.R. Rivera, F.J. Hoffmann, and R.A. Finch. 1992. Teratogenic assessment of four solvents using the frog embryo teratogenesis assay--Xenopus (FETAX). J. Appl. Toxicol. 12(1):49-56.

Fentem, J.H., G.E.B. Archer, M. Balls, P.A. Botham, R.D. Curren, L.K. Earl, D.J. Esdaile, H.-G. Holzhutter, and M. Liebsch. 1998. The ECVAM international validation study on *in vitro* tests for skin corrosivity. 2. Results and evaluation by the management team. Toxicol. In Vitro 12:483-524.

Finch, R.A., H.S. Gardner, Jr., and J.A. Bantle. 1994. Frog embryo teratogenesis assay—Xenopus: A nonmammalian method for developmental toxicity assessment. In: Symposium on Current Concepts and Approaches on Animal Test Alternatives, H. Salem (Ed.), pp. 297-313.

Fort, D.J., and J.A. Bantle. 1990. Analysis of the mechanism of isoniazid-induced developmental toxicity with frog embryo teratogenesis assay: Xenopus (FETAX). Teratogen. Carcinogen. Mutagen. 10:463-476.

Fort, D.J., and E.L. Stover. 1997. Assessing Ecological Hazard to Amphibian Populations. Purdue Industrial Waste Conference Proceedings, pp. 351-358

Fort, D.J., B.L. James, and J.A. Bantle. 1989. Evaluation of the developmental toxicity of five compounds with the frog embryo teratogenesis assay: Xenopus (FETAX) and a metabolic activation system. J. Appl. Toxicol. 9(6):377-388.

Fort, D.J., J.R. Rayburn, D.J. DeYoung, and J.A. Bantle. 1991. Assessing the efficacy of an Aroclor 1254-induced exogenous metabolic activation system for FETAX. Drug. Chem. Toxicol. 14:143-160.

Fort, D.J., J.R. Rayburn, and J.A. Bantle. 1992. Evaluation of acetaminophen-induced developmental toxicity using FETAX. Drug Chem. Toxicol. 15(4):329-350.

Fort, D.J., E.L. Stover, J.R. Rayburn, M. Hull, and J.A. Bantle. 1993. Evaluation of the developmental toxicity of trichloroethylene and detoxification metabolites using Xenopus. Teratogen. Carcinogen. Mutagen. 13:35-45.

Fort, D., E.L. Stover, and D. Norton. 1995. Ecological hazard assessment of aqueous soil extracts using FETAX. J. Appl. Toxicol. 12(3):183-191.

Fort, D.J., E.L. Stover, and J.A. Bantle. 1996. Integrated ecological hazard assessment of waste site soil extracts using FETAX and short-term fathead minnow teratogenesis assay. In: La Point, T. W., F. T. Price, and E. E. Little (Eds.). ASTM STP, 1262. Environmental Toxicology and Risk Assessment. Fourth Symposium on Environmental Toxicology and Risk Assessment. Montreal, Quebec, Canada. April 11-13, 1994. American Society for Testing and Materials, Philadelphia, pp. 93-109.

Fort, D.J., E.L. Stover, J.A. Bantle, J.R. Rayburn, M.A. Hull, R.A. Finch, D.T. Burton, S.D. Turley, D.A. Dawson, G. Linder, D. Buchwalter, M. Kumsher-King, and A.M. Gaudet-Hull. 1998. Phase III interlaboratory study of FETAX, part 2: Interlaboratory study of an exogenous metabolic activation system for Frog Embryo Teratogenesis Assay—*Xenopus* (FETAX). Drug Chem. Toxicol. 21:1-14.

Fort, D.J., T.L. Propst, E.L. Stover, J.C. Helgen, R. Levey, K. Gallagher, and J.G. Burkhart. 1999a. Effects of pond water, sediment, and sediment extracts from Minnesota and Vermont on early development and metamorphosis in *Xenopus*. Environ. Toxicol. Chem. 18(10):2316-2324.

Fort, D.J., R. Rogers, H. Copley, L. Bruning, E.L. Stover, J. Helgen, and J.G. Burkhart. 1999b. Progress toward identifying causes of mal-development induced in *Xenopus* by pond water and sediment extracts from Minnesota. Environ. Toxicol. Chem. 18(10):2316-2324.

Fort, D.J., R. Rogers, H. Copley, L. Bruning, E.L. Stover, and D. Rapaport. 1999c. Effect of sulfometuron methyl and nicosulfuron on development and metamorphosis in *Xenopus laevis*: Impact of purity. Environ. Toxicol. Chem. 18(12):2934-2940.

Fort, D.J., E.L. Stover, J.A. Bantle, J.N. Dumont, E.D. Clegg, R.A. Finch, and D.L. Danley. 1999d Evaluation of a reproductive toxicity assay using X*enopus laevis*: Boric acid, cadmium, and ethylene glycol monomethyl ether. J. Appl. Toxicol. Submitted for publication.

Fort, D.J., E.L. Stover, P.L. Strong, F.J. Murray, and C.L. Keen. 1999e Chronic feeding of a low boron diet adversely affects reproduction and development in *Xenopus laevis*. J. Nutr. 129(11):2055-2060.

Fort, D.J., T.L. Propst, E.L. Stover, and P.L. Strong. 1999f Adverse effects from low dietary and environmental boron exposure on reproduction, development, and maturation in *Xenopus laevis*. J. Trace Elements Exp. Med. 12:175-185.

Fort, D.J., E.L. Stover, P.L. Strong, and F.J. Murray. 1999g Effect of boron deprivation on reproductive parameters in *Xenopus laevis*. J. Trace Elements Exp. Med. 12:187-204.

Fort, D.J., T.L. Propst, E.L. Stover, D.R. Farmer, and J.K. Lemen. 2000a. Assessing the predictive validity of Frog Embryo Teratogenesis Assay: *Xenopus* (FETAX). Teratogenesis Carcinogen. Mutagen. 20(2):87-98.

Fort, D.J., R.L. Rogers, H.F. Copley, L.A. Morgan, M.F. Miller, P.A. Clark, J.A. White, R.R. Paul, and E.L. Stover. 2000b. Preliminary validation of a short-term morphological assay to evaluate adverse effects on amphibian metamorphosis and thyroid function using *Xenopus laevis*. J. Appl. Toxicol. In press.

Fort, D.J., E.L. Stover, J.A. Bantle, and R.A. Finch. 2000c. Evaluation of the developmental toxicity of thalidomide using Frog Embryo Teratogenesis Assay: *Xenopus* (FETAX): Biotransformation and detoxification. Teratogen. Carcinogen. Mutagen. 20(1):35-47.

Friedman, J.M., and J.E. Polifka. 1994. Teratogenic Effects of Drugs. A Resource for Clinicians (TERIS). Johns Hopkins University Press, Baltimore, MD. 703 pp.

Friedman, M., J.R. Rayburn, and J.A. Bantle. 1991. Developmental toxicology of potato alkaloids in the frog embryo teratogenesis assay—Xenopus (FETAX). Food Chem. Toxicol. 29(8):537-547.

Friedman, M., J.R. Rayburn, and J.A. Bantle. 1992. Structural relationships and developmental toxicity of Solanum alkaloids in the frog embryo teratogenesis assay--Xenopus. J. Agric. Food Chem. 40(9):1617-1624.

Gatlitski, T., A.J. Saldanha, C.A. Styles, E.S. Lander, and G.R. Fink. 1999. Ploidy regulation of gene expression. Science 285(5425):251.

Genschow, E., G. Scholz, N.A. Brown, A.H. Piersma, M. Brady, N. Clemann, H. Huuskonene, F. Paillard, S. Bremer, and H. Spielmann. 1999. *Die Entwicklung von Prdiktionsmodellen drei in vitro Embryotoxizittstets im Rahmen einer ECVAM Validierungsstudie* (Development of prediction models for three in vitro embryotoxicity tests which are evaluated in an ECVAM validation study). ALTEX 16(2):73-83.

Goldey, E.S., H.A. Tilson, and K.M. Crofton. 1995. Implications of the use of neonatal birth weight, growth, viability, and survival data for predicting developmental neurotoxicity: A survey of the literature. Neurotoxicol. Teratol. 17(3):313-332.

Kavlock, R.J., J.A. Greene, G.L. Kimmel, R.B. Morrissey, E. Owens, J.M. Rogers, T.W. Sadler, H.F. Stack, M.D. Waters, and F. Welsch. 1991. Activity profiles of developmental toxicity: Design considerations and pilot implementation. Teratogenesis 43:150-185.

Kimmel, G.L. 1990. *In vitro* assays in developmental toxicology: their potential application in risk assessment. In: *In Vitro* Methods in Developmental Toxicology: Use in Defining Mechanisms and Risk Parameters, G.L. Kimmel and D.M. Kochbar (Eds.), CRC Press, Boca Raton, FL, pp 163-173.

Kimmel, G.L., K. Smith, D.M. Kochhar, and R.M. Pratt. 1982. Overview of *in vitro* teratogenicity testing: Aspects of validation and application to screening. Teratogen. Carcinogen. Mutagen. 2:221-229.

Kononen, D.W., and R.A. Gorski. 1997. A method for evaluating the toxicity of industrial solvent mixtures. Environ. Toxicol. Chem. 16(5):968-976.

La Clair, J.J., J.A. Bantle, and J. Dumont. 1998. Photoproducts and metabolites of a common insect growth regulator produce developmental deformities in *Xenopus*. Environ. Sci. Technol. 32(10):1453-1461.

Luo, S.-Q., M.C. Plowman, S.M. Hopfer, and F.W. Sunderman. 1993. Embryotoxicity and teratogenicity of Cu^{2+} and Zn^{2+} for Xenopus laevis, assayed by the FETAX procedure. Ann. Clin. Lab. Science 23(2):111-120.

Morgan, M.K., P.R. Scheuerman, C.S. Bishop, and R.A. Pyles. 1996. Teratogenic potential of atrazine and 2,4-D using FETAX. J. Toxicol. Environ. Health 48:151-168.

National Institute for Occupational Safety and Health (NIOSH). RTECS® (Registry of Toxic Effects of Chemical Substances). On: the TOXNET® system. Internet Resource (http://sis.nlm.nih.gov/sis1/).

National Library of Medicine (NLM). HSDB® (Hazardous Substances Data Bank). On: the TOXNET® system. Internet Resource (http://sis.nlm.nih.gov/sis1/).

Nieuwkoop, P.D., and J. Faber. 1975. Normal tables of *Xenopus laevis* (Daudin), 2nd ed., North Holland Press, Amsterdam.

OECD (Organization for Economic Cooperation and Development). 1981. OECD guideline for testing of chemicals 414: Teratogenicity. OECD, Paris.

OECD (Organization for Economic Cooperation and Development). 1983. OECD guideline for testing of chemicals 415: One-generation reproduction toxicity study. OECD, Paris.

OECD (Organization for Economic Cooperation and Development). 1983. OECD guideline for testing of chemicals 416: Two-generation reproduction toxicity study. OECD, Paris.

OECD (Organization for Economic Cooperation and Development). 1995. OECD guideline for testing of chemicals 421: Reproduction/developmental toxicity screening test. OECD, Paris.

OECD (Organization for Economic Cooperation and Development. 1996. OECD guideline for testing of chemicals 422: Combined repeated dose toxicity study with the reproduction/ developmental toxicity screening test. OECD, Paris.

Palmer, B.D., and S.K. Palmer. 1995. Vitellogenin induction by xenobiotic estrogens in the red-eared turtle and African clawed frog. Environ. Health Perspect. 103 (Supple. 4): 19-25.

Pöch, G., and D.A. Dawson. 1996. Average empirical effects of mixtures of differently acting teratogenic agents. Arch. Complex Environ. Stud. 8(1-2):33-39.

Propst, T.L., D.J. Fort, E.L. Stover. 1997. Evaluation of the developmental toxicity of benzo[a]pyrene and 2-acetylaminofluorene using Xenopus: Modes of biotransformation. Drug Chem. Toxicol. 20:45-61.

Rayburn, J.R., D.J. Fort, R. McNew, and J.A. Bantle. 1991. Synergism and antagonism induced by three carrier solvents with t-retinoic acid and 6-aminonicotinamide using FETAX. Bull. Environ. Contam. Toxicol. 46:625-632.

Rayburn, J.R., J.A. Bantle, and M. Friedman. 1994. Role of carbohydrate side chains of potato glycoalkaloids in developmental toxicology. J. Ag. Food Chem. 42:1511-1515.

Riggin, G.W., and T.W. Schultz. 1986. Teratogenic effects of benzoyl hydrazine on frog embryos. Trans. Am. Microsc. Soc. 105:197-210.

Sabourin, T.D., and R.T. Faulk. 1987. Comparative evaluation of a short-term test for developmental effects using frog embryos. In: Banbury Report 26: Developmental Toxicology: Mechanisms and Risk, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY. pp. 203-223.

Sakamoto, M.K, S. Mima, and T. Tanimura. 1992. An assay system for developmental toxicity using embryos and larvae of *Xenopus laevis*. ALTEX 1:172-177.

Scholz, G., I. Pohl, A. Seiler, S. Bremer, N.A. Brown, A.H. Piersma, H.G. Holzhütter, and H. Spielmann. 1998. *Ergebnisse der ersten Phase des ECVAM-Projektes zur Prävalidierung und Validierung von drei in vitro Embryotoxizitästes* (Results of the first phase of the ECVAM project "prevalidation and validation of three in vitro embryotoxicity tests"). ALTEX 15(1):3-8.

Schultz, T.W., and D.A. Dawson. 1995. Developmental hazard assessment with FETAX: Aerobic metabolites in bacterial transformation of naphthalene. Bull. Environ. Contam. Toxicol. 54:662-667.

Schuytema, G.S., A.V. Nebeker, and W.L. Griffis. 1994. Toxicity of Guthion® and Guthion® 2S to Xenopus laevis embryos. Arch. Environ. Contam. Toxicol. 27:250-255.

Schardein, J.L. 1993. Chemically Induced Birth Defects, 2nd ed. Marcel Dekker, New York.

Schwetz, B.A., and M.W. Harris. 1993. Developmental toxicology: Status of the field and contribution of the national toxicology program. Environ. Health Perspectives 100:269-282.

Schwetz, B.A., R.E. Morissey, F. Welsch, and R.A. Kavlock. 1991. *In vitro* teratology. Environ. Health Perspect. 94:265-268.

Shepard, T.H. 1995. Catalog of Teratogenic Agents. 8th ed., John Hopkins University Press, Baltimore, MD.

Smith, M.K., G.L. Kimmel, D.M. Kochhar, T.H. Shepard, S.P. Spielberg, and J.G. Wilson. 1983. A selection of candidate compounds for *in vitro* teratogenesis test validation. Teratogen. Carcinog. Mutagen. 3:461-480.

Spielman, H. 1998. Reproduction and development. Environ. Health Perspect. 106:571-576.

Sunderman, Jr., F.W., M.C. Plowman, and S.M. Hopfer. 1991. Embryotoxicity and teratogenicity of cadmium chloride in Xenopus laevis, assayed by the FETAX procedure. Ann. Clin. Lab. Science 21(6):381-391.

Szabo, K.T. 1989. Congenital Malformations in Laboratory and Farm Animals. Academic Press, Inc., New York, NY.

Tanumiura, T., and M.K. Sakamoto. 1995. Alternatives to animal experiments in developmental toxicity tests. Acta Med. Kinki Univ. 20(3):135-148.

U.S. Environmental Protection Agency. 1991. Guidelines for Developmental Toxicity Risk Assessment. Federal Register 56(231):63798-63826.

U.S. Food and Drug Administration. 1994. International Conference on Harmonisation; Guideline on Detection of Toxicity to Reproduction for Medicinal Products; Availability; Notice. Federal Register 59(140):48749.

U.S. Food and Drug Administration. 1993. Redbook II. Toxicological Principles for the Safety Assessment of Direct Food Additives and Color Additives Used in Food.

Vismara, C., G. Bernardini, P. Bonfanti, A. Colombo, and M. Camatini. 1993. The use of *in vitro* fertilization in the Frog Embryo Teratogenesis Assay in Xenopus (FETAX) and its applications to ecotoxicology. Sci. Total Environ. Suppl. Pt. 1:787-790.

Vogel, G. 1999. Frog is a prince of a new model organism. Science 285(5424):25.

Walker, C., K. Kaiser, W. Klein, L. Lagadic, D. Peakall, S. Sheffield, T. Soldan, and M. Yasuno. 1998. 13th Meeting of the Scientific Group on Methodologies for the Safety Evaluation of Chemicals (SGOMSEC): Alternative Testing Methodologies for Ecotoxicity. Environ. Hlth. Perspect. 106(S2):441-451.

Zaga, A., E.E. Little, C.F. Rabeni, and M.R. Ellersieck. 1998. Photoenhanced toxicity of a carbamate insecticide to early life stage anuran amphibians. Environ. Toxicol. Chem. 17(12):2543-2553.