OECD Environment, Health and Safety Publications

Series on Testing and Assessment

No. 43

DRAFT GUIDANCE DOCUMENT ON REPRODUCTIVE TOXICITY TESTING AND ASSESSMENT

Environment Directorate

Organisation for Economic Co-operation and Development

November 10, 2004
(1st version)

Also published in the Series on Testing and Assessment:

No. 1, *Guidance Document for the Development of OECD Guidelines for Testing of Chemicals (1993; reformatted 1995)*

No. 2, *Detailed Review Paper on Biodegradability Testing (1995)*

No. 3, *Guidance Document for Aquatic Effects Assessment (1995)*

No. 4, *Report of the OECD Workshop on Environmental Hazard/Risk Assessment (1995)*

No. 5, *Report of the SETAC/OECD Workshop on Avian Toxicity Testing (1996)*

No. 6, *Report of the Final Ring-test of the Daphnia magna Reproduction Test (1997)*

No. 7, *Guidance Document on Direct Phototransformation of Chemicals in Water (1997)*

No. 8, *Report of the OECD Workshop on Sharing Information about New Industrial Chemicals Assessment (1997)*

No. 9, *Guidance Document for the Conduct of Studies of Occupational Exposure to Pesticides during Agricultural Application (1997)*

No. 10, *Report of the OECD Workshop on Statistical Analysis of Aquatic Toxicity Data (1998)*

No. 11, *Detailed Review Paper on Aquatic Testing Methods for Pesticides and industrial Chemicals (1998)*

No. 12, *Detailed Review Document on Classification Systems for Germ Cell Mutagenicity in OECD Member Countries (1998)*

No. 13, *Detailed Review Document on Classification Systems for Sensitising Substances in OECD Member Countries 1998)*

No. 14, *Detailed Review Document on Classification Systems for Eye Irritation/Corrosion in OECD Member Countries (1998)*

No. 15, *Detailed Review Document on Classification Systems for Reproductive Toxicity in OECD Member Countries (1998)*

No. 16, *Detailed Review Document on Classification Systems for Skin Irritation/Corrosion in OECD Member Countries (1998)*

No. 17, *Environmental Exposure Assessment Strategies for Existing Industrial Chemicals in OECD Member Countries (1999)*

No. 18, *Report of the OECD Workshop on Improving the Use of Monitoring Data in the Exposure Assessment of Industrial Chemicals (2000)*

No. 19, *Draft Guidance Document on the Recognition, Assessment and Use of Clinical Signs as Humane Endpoints for Experimental Animals used in Safety Evaluation (1999)*

No. 20, *Revised Draft Guidance Document for Neurotoxicity Testing (in preparation)*

No. 21, *Detailed Review Paper: Appraisal of Test Methods for Sex Hormone Disrupting Chemicals (2000)*

No. 22, *Guidance Document for the Performance of Out-door Monolith Lysimeter Studies (2000)*

No. 23, *Guidance Document on Aquatic Toxicity Testing of Difficult Substances and Mixtures (2000)*

No. 24, *Guidance Document on Acute Oral Toxicity Testing (2001)*

No. 25, *Detailed Review Document on Hazard Classification Systems for Specifics Target Organ Systemic Toxicity Repeated Exposure in OECD Member Countries (2001)*

No. 26, *Revised Analysis of Responses Received from Member Countries to the Questionnaire on Regulatory Acute Toxicity Data Needs (2001)*

No 27, *Guidance Document on the Use of the Harmonised System for the Classification of Chemicals Which are Hazardous for the Aquatic Environment (2001)*

No 28, *Guidance Document for the Conduct of Skin Absorption Studies (2004)*

No 29, *Draft Guidance Document on Transformation/Dissolution of Metals and Metal Compounds in Aqueous Media (2001)*

No 30, *Detailed Review Document on Hazard Classification Systems for Mixtures (2001)*

No 31, *Detailed Review Paper on Non-Genotoxic Carcinogens Detection: The Performance of In-Vitro Cell Transformation Assays (draft)*

No. 32, *Guidance Notes for Analysis and Evaluation of Repeat-Dose Toxicity Studies (2000)*

No. 33, *Harmonised Integrated Classification System for Human Health and Environmental Hazards of Chemical Substances and Mixtures (2001)*

No. 34, *Guidance Document on the Development, Validation and Regulatory Acceptance of New and Updated Internationally Acceptable Test Methods in Hazard Assessment (in preparation)*

No. 35, *Guidance notes for analysis and evaluation of chronic toxicity and carcinogenicity studies (2002)*

No. 36, *Report of the OECD/UNEP Workshop on the use of Multimedia Models for estimating overall Environmental Persistence and long range Transport in the context of PBTS/POPS Assessment (2002)*

No. 37, *Detailed Review Document on Classification Systems for Substances Which Pose an Aspiration Hazard (2002)*

No. 38, *Detailed Background Review of the Uterotrophic Assay Summary of the Available Literature in Support of the Project of the OECD Task Force on Endocrine Disrupters Testing and Assessment (EDTA) to Standardise and Validate the Uterotrophic Assay (2003)*

No. 39, *Guidance Document on Acute Inhalation Toxicity Testing (in preparation)*

No. 40, *Detailed Review Document on Classification in OECD Member Countries of Substances and Mixtures Which Cause Respiratory Tract Irritation and Corrosion (2003)*

No. 41, *Detailed Review Document on Classification in OECD Member Countries of Substances and Mixtures which in Contact with Water Release Toxic Gases (2003)*

No. 42, *Draft Guidance Document on Reporting Summary Information on Environmental, Occupational and Consumer Exposure (in preparation)*

No. 43, *Draft Guidance Document on Reproductive Toxicity Testing and Assessment (in preparation)*

No. 44, *Descriptions of Selected Key Generic Terms used in Chemical Hazard/Risk Assessment (2003)*

No. 45, *Draft Guidance Document on the Use of Multimedia Models for Estimating Overall Environmental Persistence and Long-range Transport (2004)*

About the OECD

The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organisation in which representatives of 30 industrialised countries in North America, Europe and the Asia and Pacific region, as well as the European Commission, meet to co-ordinate and harmonise policies, discuss issues of mutual concern, and work together to respond to international problems. Most of the OECD's work is carried out by more than 200 specialised committees and working groups composed of member country delegates. Observers from several countries with special status at the OECD, and from interested international organisations, attend many of the OECD's workshops and other meetings. Committees and working groups are served by the OECD Secretariat, located in Paris, France, which is organised into directorates and divisions.

The Environment, Health and Safety Division publishes free-of-charge documents in nine different series: **Testing and Assessment**; **Good Laboratory Practice and Compliance Monitoring**; **Pesticides and Biocides**; **Risk Management**; **Harmonisation of Regulatory Oversight in Biotechnology**; **Safety of Novel Foods and Feeds; Chemical Accidents; Pollutant Release and Transfer Registers; and Emission Scenario Documents.** More information about the Environment, Health and Safety Programme and EHS publications is available on the OECD's World Wide Web site (http://www.oecd.org/ehs/).

This publication was produced within the framework of the Inter-Organisation Programme for the Sound Management of Chemicals (IOMC).

The Inter-Organisation Programme for the Sound Management of Chemicals (IOMC) was established in 1995 following recommendations made by the 1992 UN Conference on Environment and Development to strengthen co-operation and increase international co-ordination in the field of chemical safety. The participating organisations are FAO, ILO, OECD, UNEP, UNIDO, UNITAR and WHO. The World Bank and UNDP are observers. The purpose of the IOMC is to promote co-ordination of the policies and activities pursued by the Participating Organisations, jointly or separately, to achieve the sound management of chemicals in relation to human health and the environment.

This publication is available electronically, at no charge.

For this and many other Environment, Health and Safety publications, consult the OECD's World Wide Web site (www.oecd.org/ehs/)

or contact:

OECD Environment Directorate,
Environment, Health and Safety Division

2 rue André-Pascal
75775 Paris Cedex 16
France

Fax: (33-1) 45 24 16 75

E-mail: ehscont@oecd.org

TABLE OF CONTENTS

EXECUTIVE SUMMARY

[TO BE DRAFTED AFTER CONSENSUS HAS BEEN REACHED ON THE DOCUMENT]

I. INTRODUCTION

History of the document

1. The OECD Working Group on Reproduction and Developmental Toxicity met in Copenhagen, Denmark in June 1995. The meeting recognised that there was a need for a Guidance Document on developmental and reproduction toxicity testing, covering testing strategies, approaches for tier testing, relationship with neurotoxicity testing and data interpretation procedures. The 7th WNT meeting approved to develop the Guidance Document in October 1996 and the US EPA volunteered to take the lead in this activity.

2. The drafting group met several times to co-ordinate the drafting of the document. In June 2003, the US EPA hosted a Meeting for the drafting group in Washington DC to resolve the last problems and assign writing tasks.

Purpose of the document

3. This document is intended to provide guidance on methodological aspects, interpretation of data and strategy for testing of chemicals for potential reproductive toxicity. The primary objective of this guidance is to ensure that necessary and sufficient data are obtained to enable adequate evaluation of the risk of reproductive toxicity arising from exposure to a chemical.

4. A stepwise assessment/testing strategy is recommended. To minimise animal usage and optimise allocation of resources, data should be assessed following each step of testing to decide if they are adequate for the evaluation of the risk arising from the intended use of the chemical, or if further testing is needed. Because assessment of data is an essential part of the overall strategy, this document provides a definition of reproductive toxicity, and discusses different types of effect *i.e.,* adverse effects on sexual function and fertility in adult males and females as well as developmental toxicity. Guidance is provided on test methods, because familiarity with these methods is required to assess data or to select methods to identify and/or characterise effects on reproductive toxicity.

5 The document constitutes an essential supplement to existing OECD Test Guidelines that can be used to obtain information on the potential reproductive toxicity of chemicals. Specific OECD Test Guidelines include the one- and two-generation toxicity study (TG 415 and 416), prenatal developmental toxicity study (TG 414), developmental neurotoxicity study (draft TG 426) and the reproduction/developmental toxicity screening tests (TG 421 and 422). However, data from other toxicity studies *e.g.,* repeated dose toxicity studies for systemic toxicity (TG 407, 408 and 409) may indicate potential reproductive toxicity and should be considered in the assessment as well as data from human exposure.

6. The term Endocrine Disruption (ED) denotes the ability of exogenous chemicals to alter function(s) of the endocrine system and consequently cause adverse health effects in an intact organism, or its progeny, or (sub) populations. Extensive information has become available on ED in the recent years (CSTEE 1999), OECD has initiated a specific programme to develop new Test Guidelines, and to revise the existing ones for detection of ED. The current Test Guidelines may enable the identification of ED-related effect if conducted with specific attention to this issue. However, current Test Guidelines for toxicological testing could be improved with the aim of increasing sensitivity and specificity of detection of ED-related effects. This Guidance document will include considerations on ED-related effects but ED will not be devoted a specific paragraph.

7 A number of organisations in several countries have written documents on the evaluation of reproductive toxicity data, *e.g.*, ECETOC 1983, 1992, 2002; EHC 1984; LST 1989; NIOH (AMI) 1994; DEPA 1995; IPCS-WHO 2001 and OECD 2002a.

Definition of developmental/reproductive toxicity and principles of reproductive toxicity hazard assessment

8. Reproduction is the biological process that ensures the continuation of the species. Existing genetic material is passed to the next generation.

9. Reproductive toxicity includes adverse effects on sexual function and fertility in adult males and females as well as developmental toxicity in the offspring.

10. Adverse effects on reproductive ability or capacity include alterations in the female and male reproductive organs or related endocrine systems *i.e.*, effects on:

- onset of puberty;
- gamete production and transport;
- reproductive cycle normality;
- sexual behaviour;
- fertility;
- parturition;
- premature reproductive senescence, or;
- modifications in other functions that are dependent on the integrity of the reproductive systems.

11. Developmental toxicity taken in its widest sense includes any effect which interferes with normal development of the conceptus, either before or after birth, and resulting from exposure of either parents prior to conception, or exposure of the developing offspring during prenatal development, or post-natally, to the time of sexual maturation. These effects can be manifested at any point in the life-span of the organism. The major manifestations of developmental toxicity include;

- death of the developing organism;
- structural abnormality;
- altered growth, and;
- functional deficiency.

12. The definition of reproductive toxicity hinges on interpretation of the word adverse (OTA 1990). A practical definition of an "adverse change" is: "*any treatment-related alteration from baseline that diminish an organism's ability to survive, reproduce or adapt to the environment*" (US EPA 1994).

13. Reproductive toxicity may result from single or repeated doses of a chemical. Since both single and repeated exposures are possible scenarios for human exposure, the reproductive toxicity assessment/testing strategy must ideally consider both situations.

14. The assessment of reproductive toxicity will be discussed in the following sections on specific endpoints. The available methodologies, normal range of variation for test species, interrelationship among endpoints, relationship of endpoint to functional outcome, differences among test species, relationship with maternal toxicity, reversibility of outcome, and relevance to humans will be discussed in the sections on prenatal endpoints, postnatal endpoints, and adult reproductive toxicity.

IDENTIFICATION AND ASSESSMENT OF THE EFFECTS OF CHEMICALS ON REPRODUCTION AND DEVELOPMENT

Reproductive Toxicity Tests

15. The reproductive toxicity tests constitute an important part of the toxicological testing program in the health assessment of chemicals. The data from the toxicological testing shall provide the necessary information with respect to:

- Hazard identification
- Dose-effect assessment (if possible estimation of a NOAEL)
- Extrapolation (prediction of adverse effects in other species, particularly in humans)
- Prediction of safe levels of exposure in humans

16. Tests for reproductive toxicity should ideally be able to identify one or more of the following effects on the reproduction:

- Impairment of male and female reproductive functions or capacity, *i.e.*, adverse effects on libido, sexual behaviour, any aspect of spermatogenesis or oogenesis, or hormonal activity or physiological response which would interfere with the capacity to fertilise, fertilisation itself or development of the fertilised ovum up to and including implantation.

- Induction of non-heritable harmful effects on the progeny, *i.e.*, in the widest sense, any effect interfering with normal development, both before and after birth up to puberty should be included. Both morphological malformation(s) and functional disturbances (*e.g.*, hormonal reproductive effect, neurological) should be evaluated.

17. Many different experimental methods for investigating toxic effects of chemicals on reproduction and development are in use (OECD 2002b; Toppari *et al.*, 1996; ECETOC 2002; Meyer and Svendsen 2003). Several tests are standardised and Test Guidelines have been issued by various governmental agencies and international organisations. The following sections will focus mainly on standardised and regulatory accepted Test Guidelines, some of which are presented in table 1. Other test than those included in the table 1 can reveal effects which indicate a potential of a chemical to interfere with normal reproduction, *e.g.*, the dominant lethal test, fertility assessment by continuous breeding, and repeated dose toxicity testing where the gonads are subjected to pathological examination.

Table 1. Overview of *in vivo* tests for reproductive toxicity testing

Test	Exposure period	Endpoints in offspring	Guideline(s)
Generation studies	Continuously over one, two or several generations	Growth, development and viability. Histopathology of sex organs, brain and target organs Fertility Proposal: oestrus cyclicity and sperm quality	TG 415: One-generation Study TG 416 :Two-generation Study
Prenatal Developmental Toxicity Study (Teratology study)	Usually during organogenesis Proposal: from implantation to the day before birth	Resorptions Embryonic development Foetal growth Morphological variations and malformations.	TG 414: Prenatal Developmental Study

Developmental Neurotoxicity Study (Behavioural teratology studies)	During pregnancy and lactation	Birth and pregnancy length Physical and functional maturation Behavioural changes due to CNS and PNS effects Brain weights and neuropathology	TG 426 (draft): Developmental Neurotoxicity Study
Reproduction/ Developmental toxicity screening test	At least three dose levels from 2 weeks prior to mating until day 4 post-natally	Fertility Pregnancy length and birth Foetal and pup growth and survival until day 3	TG 421 and 422

18. During recent years many *in vitro* test systems have been proposed as alternatives to animal testing for developmental toxicity (ATLA 2002). These tests usually address single events of the reproductive cycle and are therefore as such insufficient for the assessment of adverse *in vivo* effects, but they do not replace animal testing in the risk assessment of chemicals. However, these tests may be useful for screening of closely related chemicals and for elucidating the mechanisms underlying the effects. They may also be essential elements of stepwise testing and assessment strategies (see Chapter 5).

Comparative Assessment of Effects in Adult Animals

19. Repeated dose toxicity testing in adult animals provide information on the potential for systemic toxicity by investigations of growth, clinical symptoms, haematology, biochemistry, organ weights, pathology and histopathology of organs.

20. Reproductive toxicity testing can provide information on a number of developmental effects, such as malformations, growth retardation, foetal and postnatal death, fertility, and functional effects on the CNS. The tests for reproductive toxicity are unique in that *e.g.*, information on developmental effects on fertility and sex organs are only provided in the two-generation study, while effects on brain development and function is investigated only in the developmental neurotoxicity study.

21. However, the investigations of systemic effects in reproductive toxicity tests are not similar to those of repeated dose toxicity studies in adults, since *e.g.*, haematology and biochemistry is not normally investigated. In addition, the investigations of organ weight, pathology and histopathology are limited to the brain, sex organs and identified target organs. Consequently, systemic effects induced during pre- or postnatal development on *e.g.*, liver and kidneys may not be investigated.

22. In order to have a sufficient background to determine the sensitivity of the developmental period compared to adulthood there is a need for studies where end points are investigated similarly for both age groups. Ideally, this would require a two-generation study incorporating developmental neurotoxicity end points and supplemented with similar investigations of systemic effects in offspring as in repeated dose toxicity studies.

Significance of Experimental Data and their Relevance to Humans (Extrapolation)

23. The ultimate proof that a substance is a human reproductive toxicant can come only from information on the consequences of human exposure. This statement is valid regarding toxicological effects in general. However the experience within the area of developmental toxicology comparing data from human exposure and data from dosing experimental animals (irrespective the difficulties comparing two different kinds of data sets), and the "biology" as expressed by Calabrese (1983);

> *"Cellular structure and biochemistry are remarkably alike across the entire animal kingdom, starting with the lipoprotein cell membrane, which affects the absorption of xenobiotics into the cell to metabolic processes like glycolysis, the Krebs' cycle, and numerous other aspects of intermediary metabolism. The similarity among animals on the cellular level is so apparent that it serves as the basis upon which scientists have extrapolated or inferred functions from one species to another"*

In general, this serves as the basis for the use of laboratory animals in predicting possible consequences for humans. Thus, findings in experimental studies in laboratory animals, demonstrating developmental toxicity is indicative of a potential human response.

24. Concerning the predictive value of *in vitro* results and their relevance for human, the "extrapolation" between the response in isolated parts of the organism and that of the intact organism has to be considered.

25. There are sufficient data in the scientific literature to consider a positive effect in an *in vivo* study for developmental toxicity as indicative of a human response (Jakobsen and Meyer 1989; Schardein *et al.*, 1985; IPCS 2001) Well-conducted studies in laboratory animals that show developmental effects at levels of exposure not substantially greater than humans might experience are sufficient to raise substantial questions about the potential risk to humans. However, no single laboratory test species can be said to predict accurately the true human response to a given chemical. Tests in multiple species may increase the predictive reliability of laboratory animal test data, but specific differences between the test species and humans must be considered in evaluating the relevance of particular tests to humans (Jakobsen and Meyer 1989).

26. Protocols for fertility studies in males have been less extensively studied than those for females (Barlow and Sullivan 1982). However, effects on fertility in rodents seem to be a good indicator for effects in humans, and most work on contraceptive agents in humans stems from original studies in rodents (Barlow and Sullivan 1982).

27. Particular strengths and limitations will be included in the following description of the individual tests for reproductive toxicity.

Single and Multi-Generation Studies

28. Test Guidelines for carrying out single and multi-generation studies have been published by the OECD and other organisations. In the EU, the two-generation study is requested for new and existing industrial chemicals (*i.e.,* introduced after 1980) with a production volume reaching 100 tons per year (see the EU REACH (Registration, Evaluation and Authorisation of Chemicals) proposal (EU 2003a).

29. The purpose of these studies is to examine successive generations to identify possible effects of a substance on fertility of male and female animals; pre-, peri-, and post-natal effects on the ovum, foetus, and progeny, including teratogenic and mutagenic effects; and peri- and post-natal effects on the mother.

30. The various Test Guidelines have a number of common requirements. The preferred species are the rat and the mouse. Other species may be used if relevant (*e.g.,* differences in toxicokinetics between the

preferred species and man, to clarify ambiguous results or further study observed effects). The test substances are administered by the relevant route to groups of animals (number of animals per group sufficient to yield about 20 pregnant animals at or near term). In general, the chemical is administered over at least one spermatogenic cycle and the last stages of oocyte maturation before the parent generation animals are mated. The exposure of the females is continued throughout the mating period and the gestation up to weaning of the last generation. At least three treatment groups and a control group (untreated or vehicle in the highest dose used) are be used. Ideally, unless limited by the physico-chemical nature or biological effects of the test substance, the highest dose level should induce toxicity but not mortality in the parental animals. The low dose should ideally not induce any observable adverse effects on the parents or offspring.

31. Birth weight, postnatal growth and survival are recorded, but it is important in the evaluation of the data to consider variations due to different litter sizes or different sex distribution in the litters. A change in offspring body weight is a sensitive indicator of developmental toxicity, in part because it is a continuous variable. In some cases, weight reduction in offspring may be the only indicator of developmental toxicity in a generation study. While there is always a question remaining as to whether weight reduction is a permanent or transitory effect, little is known about the long-term consequences of short-term foetal or neonatal weight changes. Therefore, weight reduction should be considered as a relevant effect in establishing the NOAEL.

32. In the two-generation reproductive toxicity study (TG 416), assessment of effects on sperm quality and oestrus cyclicity in offspring have been included. In addition the assessment of the offspring has been expanded with the recording of developmental milestones including some optional behavioural parameters and histopathology of sex organs, brain and identified target organs.

33. The single and multi-generation studies include exposure of both sexes during all stages of reproduction, and the F1 animals in the multi-generation studies are unique considering F1 equivalent of an *in utero* derived repeated dose toxicity study.

Prenatal Developmental Toxicity Study, Teratology Study

34. The prenatal developmental toxicity study (TG414) is the method for examining embryo-foetal toxicity as a consequence of exposure during pregnancy (*e.g.*, growth retardation, anatomical variations, teratogenicity and lethality).

35 Originally, the teratology test focused on malformations. In the past, there was a tendency to consider only malformations or malformations and death as relevant endpoints in teratology studies. Today the teratology test focuses on the four manifestations of developmental toxicity being on concern (death, structural abnormalities, growth alterations) Consequently, the title of the TG 414 has changed to “Prenatal Developmental Toxicity Study”.

36. Various international organisations and countries have drafted guidelines for this test. These guidelines have a number of common requirements. The preferred species include rodents (*e.g.*, the rat and mouse) and non-rodents (*e.g.*, the rabbit). Other species may be used if relevant (*e.g.*, differences in toxicokinetics between the preferred species and humans, to clarify ambiguous results or to further study observed effects).

37. Young mature virgin females are artificially inseminated or mated with males. The time of mating is established by observation of mating (*e.g.*, rabbits), identification of a plug (mixture of sperm and cellular material from the vagina of rats and mice, coagulation gland and vaginal mucus), vaginal smear (in rats) or by noting the time of insemination (*e.g.*, for pigs and rabbits). Normally, three dose levels and a control group (untreated or vehicle control; the group size is 20 pregnant animals, however 16 as a minimum in TG 414) are used in order to establish a dose-effect relationship. Until recently, the pregnant female rats were up to recently

exposed at least during the period of organogenesis, *i.e.*, between day six when implantation occurs, and day 15. (The corresponding periods for mice and rabbits are days 6-15 and days 6-18, respectively). This period has been found to be the most sensitive to the induction of structural, anatomical malformations (the corresponding sensitive period for humans is between the 18th and 60th day of pregnancy). However, development of *e.g.*, sex organs and brain continues after day 15 and consequently malformations of such organs will not be discovered if exposure is stopped on day 15. In TG 414 the dosing period shall cover the period from implantation to scheduled caesarean section. If preliminary studies, when available, do not indicate a high potential for pre-implantation loss, treatment may be extended to include the entire period of gestation, from mating to the day prior to scheduled kill. The animals are observed daily for clinical changes. Body weight is recorded and food consumption is recorded throughout the gestation. The day before anticipated birth the uterus is removed by caesarean section and the uterus and the foetuses are examined. The dam is examined macroscopically for any structural abnormalities or pathological changes. If dosing is initiated before or at the time of implantation, the pre-implantation loss, *i.e.*, the number of embryos lost prior to implantation is evaluated.

38. The total number of implantations, *i.e.*, living embryos, dead embryos and resorptions (embryos that die early and are re-assimilated) are noted. The degree of resorption (*i.e.*, the extent to which the embryo has been resorbed i*.e.*, total, early, and late resorptions) is recorded in order to establish the time of death of the embryo during the pregnancy.

39. The foetuses are sexed, weighed and examined for gross malformations. Retarded growth and effects on visceral and skeletal development are evaluated.

40. The Prenatal Developmental Toxicity Study (TG 414) is very suitable for the demonstration of intra-uterine death after implantation (resorptions). In studies where dosing is started before implantation, pre-implantation loss may also be assessed.

41. Foetal weight can be assessed rather exact, but it is important to include variations due to different litter sizes or sex distribution in controls versus exposed groups in the analysis.

Developmental Neurotoxicity Studies, Postnatal Studies

42. Developmental exposure to chemicals may lead to a range of functional disturbances in the offspring. A number of chemicals are known to produce developmental neurotoxic effects in humans and other species. For example lead and methyl mercury affect brain development but effects on fertility, the immune system, metabolism of foreign substances and development as a whole have been observed.

43. A recent proposal for an OECD guideline for “Developmental Neurotoxicity Study”, (draft TG 426) has been developed based upon an US EPA guideline. Developmental neurotoxicity studies are designed to develop data on the potential functional and morphological hazards to the nervous system arising in the offspring from exposure of the mother during pregnancy and lactation. These studies identify changes in behaviour due to effects on the central nervous system (CNS) and the peripheral nervous system (PNS). As behaviour is affected by the function of other organs such as liver, kidneys and the endocrine system, toxic effects on these organs in offspring may also be reflected in general changes in behaviour. No single test is able to reflect the entire complex and intricate function of behaviour. For testing behaviour, therefore, a range of parameters, a “test battery”, is used to identify changes in individual functions. Methodology employed in behavioral teratology is described in several reviews (Adams 1986; Francis *et al.*, 1990; IPCS 2001). Behavioral teratology tests may generally be grouped into tests of physical development, simple reflexes, motor function, sensory development and functions, spontaneous activity, learning and memory, and functions of the neurotransmitter systems.

44. The preferred species are the rat and the mouse. The Test Guidelines generally recommend groups of 20 animals with dosing from day 15 of gestation to day 21 post gestation *i.e.*, spanning foetogenesis and the entire lactation period. However, the recommended dosing period does not cover all events since the CNS is also susceptible to abnormal development during the period of organogenesis. Consequently, dosing is often started earlier, for example on day one or day six of the pregnancy. In the draft Developmental Neurotoxicity Test Guideline (draft TG 426) the dosing period include the entire gestation and lactation period. After birth, the number of progeny is recorded and the litters may be adjusted so that each contains the same number of pups. In order to determine whether the chemical substance tested affects the offspring directly via the mother's milk or indirectly, either via a change in milk production or as a result of a change in the behaviour of the exposed mothers, cross-fostering may be employed. Cross fostering is a method where litters from exposed mothers are reared by control mothers and vice versa.

45. The evaluation of the offspring consists of observations to detect gross neurological and behavioural abnormalities, assessment of physical development, reflex ontogeny, motor activity, motor and sensory function, and learning and memory, and evaluation of brain weights and neuropathology during postnatal development and adulthood.

46. The draft Test Guideline 426 is designed to be used as a separate study, however, the observations and measurements can also be incorporated into *e.g.*, a two-generation study. The limitations mentioned in section 1.3.4 on single- and multi-generation studies concerning endpoints such as neonatal death, malformations, pre-implantation loss and resorptions apply also to developmental neurotoxicity studies.

Reproduction/Developmental Toxicity Screening Tests

47. In recent years, new screening tests for reproductive and developmental toxicity of shorter duration and using fewer resources have been developed. By definition, a screening test is limited in scope compared to a conventional test. Data from a screening test indicating a possible toxic potential of a substance identify the substance as one of high priority for further evaluation or risk reduction.

48. Recently, the OECD introduced Test Guidelines for screening tests for reproductive toxic effects. The “Reproduction/Developmental Toxicity Screening Test”, (TG 421) and the “Combined Repeated Dose Toxicity Study with the Reproduction/Developmental Toxicity Screening Test” (TG 422) cover part of the Screening Information Data Sets (SIDS) for high production volume (HPV) chemicals. The TG 422 is a combination of a 28-days toxicity study and a reduced one-generation study whereas the TG 421 is a reduced one-generation study.

49. The purpose of the tests is to generate limited information concerning the effects of a test substance on male and female reproductive performance such as gonadal function, mating behaviour, conception, development of conceptus and parturition. It is not suggested as an alternative to or as a replacement for the existing Test Guidelines for single and multi-generation and developmental toxicity studies.

50. The dosing of the animals is initiated two weeks prior to mating and continues until the end of the study on postnatal day four. The number of animals per group is at least ten animals of each sex and is normally sufficient to provide at least eight pregnant females per group. Effects on fertility and birth are registered. Live pups are counted and sexed and litters weighed on days one and four postpartum. The parameters include among others a detailed histological examination on the ovaries, testes and epididymides of at least the highest dosed and the control animals.

Other *In Vivo* Toxicity Tests

51. Other toxicity tests than those mentioned above could reveal effects that indicate a potential of a chemical to interfere with normal reproduction. Thus, in all the toxicological tests involving repeated dosing, including the carcinogenicity test, the gonads and accessory sex organs are subjected to pathological examination including histopathology.

52. Fertility assessment by continuous breeding has been used to study the depletion of oocytes from the ovary in mice exposed to procarbazine. In this study, pre-natally treated mice were continuously housed with untreated male mice, and the cumulative number of offspring was measured by removing the female when noticeably pregnant and then returning the female to the males cage immediately after delivery to establish a pattern of forced repetitive breeding (Chapin *et al.,* 1997; Lamb (1985).

II. PRENATAL TOXICITY END POINTS

Overview of endpoints assessed

Laboratory Animal Studies

53. For the evaluation of hazard following exposure to chemical substances (including pesticides), a number of Test Guidelines are available that include peri-natal assessment of endpoints following prenatal exposure. These include the prenatal developmental toxicity study (TG 414), the one-generation reproduction study (TG 415), the two-generation reproduction study (TG 416), the developmental neurotoxicity study (draft TG 426), and two reproductive/developmental toxicity screening tests (TG 421 and 422).

Critical Windows of Exposure

54. In the Test Guidelines described, *in utero* exposure to the offspring could occur during specified phases (*e.g.,* from implantation until one day prior to anticipated delivery, or during the major period of organogenesis and sexual differentiation, as in prenatal developmental toxicity studies) or throughout the entire gestation period (*e.g.,* from conception to delivery, as in single- or multi-generation reproduction studies or the reproductive/developmental toxicity screening tests). These broad testing scenarios are designed to screen for a variety of adverse outcomes of prenatal exposure by ensuring *in utero* exposure to the foetus across a number of critical windows of vulnerability. While it is possible in theory to identify a day or range of days during which prenatal exposure to test animals would result in a specific adverse developmental effect (Wilson, 1965; Selevan *et al.,* 2000), it is seldom that a Test Guideline could provide sufficient information to precisely identify the critical window of exposure that resulted in a particular adverse outcome. That being the case, it is assumed by default that any observed prenatal developmental effect could have resulted from a single exposure to the test substance at some time point during the exposure period.

Latent Effects

55. Latent (delayed) effects of developmental exposure are indirectly assessed in the developmental neurotoxicity study, since neurobehavioral and neuropathological assessments are conducted following a period of non-treatment (from approximately postnatal day 21 to 60). However, latent developmental effects are not assessed in the prenatal developmental toxicity study, since the foetal offspring are taken for examination prior to birth. In a multi-generation reproduction study, second (and subsequent) generations

of animals are unique in that they have been exposed to the test substance throughout *in utero* and postnatal development, and assessed for toxicity through most of their early adult life, allowing for the potential expression of latent effects in adulthood (as limited in duration by study design). It is, however, extremely difficult to distinguish between latent adverse effects that result from prenatal exposure and those that have resulted from postnatal or adult exposures, or an accumulation of multiple exposures over the various life stages of the animal. A comparison of the responses between offspring of successive generations on the same study may provide some limited information to address this issue.

Methodological issues

Corpora Lutea

56. The maternal ovaries are removed and examined at the time of necropsy and *corpora lutea* are counted. This information is useful in the interpretation of foetal viability data (see below, pre-implantation loss). *Corpora lutea* counts can be conducted with fresh or fixed tissue, and are generally performed with the aid of a dissecting microscope. Minimal dissection of the ovarian tissue is required. Counting accuracy can be affected by size differences among species (mice are somewhat challenging), but technical difficulties are not insurmountable. In prenatal developmental toxicity studies, maternal necropsies are conducted prior to delivery of the offspring; however, in reproduction studies, maternal necropsy does not occur until the offspring are weaned (postnatal day 21), resulting in further technical challenges in the accurate counting of *corpora lutea*. Additionally, multiple litters produced by a single dam will compromise *corpora lutea* count. Review of study data can provide indicators of possible inaccuracies in count for example when the number of implantations exceeds the number of *corpora lutea* within a litter; individual values from such a litter are generally not included in the statistical evaluation of group data.

Pre- and Post-Implantation Loss

57. Prenatal mortality may be grossly evident as a reduction in live litter size at the time of caesarean section or parturition. Additionally, for prenatal developmental toxicity studies, a detailed examination of uterine contents will reveal incidences of early and late resorptions, which are evidence of previous intrauterine deaths that have occurred after implantation. In prenatal developmental toxicity studies at caesarean section, very early resorptions may be observed and described as “empty implantation sites”. The uteri of dams that are necropsied at lactation day 21 in reproduction studies will appear similarly. For females that do not have visible evidence of implantation sites, further visual examination of the uterus is recommended, using techniques such as pressing the uterine tissue between two glass slides or staining with ammonium sulphide. Comparison of the number of implantation sites with the number of live and dead foetuses or neonates for each litter provides a means of quantifying post- implantation loss. Pre-implantation loss is quantified through a comparison of the numbers of corpora lutea and implantation sites for each dam. Increased pre-implantation loss could result from an adverse effect on gamete transport, fertilisation, the uterus, the developing organism, or the process of implantation. Consideration of the exposure paradigm and the pattern of dose response may provide information that assists in the interpretation of data that indicate the presence of pre-implantation loss, but it is unlikely that the specific mechanism of toxicity can be identified without further study.

External Foetal Examinations

58. A visual external examination of foetal/neonatal animals is conducted at caesarean section or parturition, enabling the identification of gross defects in structure or activity, which may have resulted from *in utero* exposure to the test substance. Viability is assessed, sometimes with the use of lung flotation as a criterion to distinguish live and dead neonates. Individual body weight data, and sometimes crown-to-

rump measurements, are recorded. Gender is generally determined at this time, by visual examination of the external genitalia and in some cases with measurement of anogenital distance, and may later be reconfirmed (*e.g.*, at soft tissue evaluation of foetuses). The integrity and closure of the palate can be confirmed by gently opening the mouth of each foetus and examining the interior.

Visceral Foetal Examinations

59. Approximately half of the foetuses from each litter on the prenatal developmental toxicity study in rodents are killed and placed into a fixative that facilitates the firming of soft tissues and the decalcification of the skeletal structure. Examination of each foetus may be performed by serial sectioning procedures or by a combination of gross organ evaluation and selected organ sectioning. Due to their relatively large size rabbit foetuses are not easily serial sectioned. For that reason, most laboratories will prefer to perform a detailed necropsy on each foetus to evaluate the structure and integrity of the thoracic and abdominal organs, with the addition of specialised procedures to examine the cranium (*e.g.*, using a coronal sectioning technique to visualise the internal structure of the brain, the eyes, and the nasal passages). Additionally, specialised procedures (*e.g.*, Staples, 1974), may be used, either as a standard method or to further explore possible developmental alterations in a known target organ. Histopathological evaluation of foetal tissues is not included in a standard guideline protocol, but could prove useful to further evaluate observed or anticipated anomalies.

Skeletal Examinations

60. For a prenatal developmental toxicity study in rodents, approximately one-half of the litter *(i.e.*, comprised of those foetuses that are not assigned to soft tissue examination) is processed for skeletal evaluation. For rabbits that are not serial sectioned (rabbit foetuses: 50% visceral examination with head; 50% visceral examination without head), all foetuses can be examined for skeletal abnormalities, after the soft tissue examination has been performed; rabbit foetuses must be eviscerated, skinned, and excess fat deposits (when present) removed from the shoulder area in order to facilitate proper processing. The foetuses (both rodent and non-rodent) are placed in a dehydrating fixative (*e.g.*, ethanol or acetone), macerated in potassium hydroxide solution, stained with calcium-specific Alizarin red S, and placed in a glycerine-based clearing solution to remove excess stain from soft tissues. The evaluation of foetal cartilage is not yet required (but recommended) by any Test Guideline, but can provide useful information on the cartilaginous precursors to ossified skeletal structures. Cartilage evaluation could entail double staining (involving an additional pre-clearing step in processing the foetuses, in which they are stained with cartilage-specific alcian blue) or alternative techniques, such as the use of differential Alizarin red S staining gradients. Following sufficient clearing of stain(s), the foetuses are examined for abnormalities of the cranial, thoracic, pelvic, and axial skeletal structures. For each foetus, the extent of ossification, normal structural formation (shape, size, integrity), position and articulation are evaluated.

Data interpretation

Malformations versus Variations

61. Classification of foetal and neonatal observations into malformations and variations is a common practice. A commonly used definition of a malformation is a permanent structural change, which may adversely affect survival, development, or function, while a common definition of a variation is a divergence beyond the usual range of structural constitution, which may not adversely affect survival or health (US EPA, 1991). There is no generally accepted classification of malformations and variations, in part because there is a continuum of responses between normal and abnormal development. It is also noted that a given observation may be classified as a malformation in one species, and a variation in another, or

the classification may change depending on the gestation day of examination. Additionally, some laboratories may attempt to further subdivide these definitions, e.g., by distinguishing between variations that are developmental delays and those that could be considered structural alterations. Therefore, the terms must be carefully and clearly defined by each individual laboratory and within the context of each study report.

International Harmonisation of Terminology

62. The International Federation of Teratology Societies (IFTS) Committee on International Harmonization of Nomenclature in Developmental Toxicology developed and published a glossary of internationally accepted common nomenclature to use when describing observations of foetal and neonatal morphology (Wise *et al.*, 1997) (Annex 1). The purpose of this effort was to advance the harmonisation of terminology, and to reduce confusion and ambiguity in the description of developmental effects, particularly in submissions to regulatory agencies world-wide. Familiarity with the IFTS terminology for external, visceral, and skeletal observations, and appropriate use of the terminology in data collection, reporting, and review, is encouraged. It is recognised, however, that although the common nomenclature developed by this effort has been widely available and internationally accepted, there is no guarantee that the terminology has been uniformly used by all laboratories that conduct studies for chemical hazard assessment.

Interrelationship of Endpoints

Maternal Toxicity

63. Studies intended to assess prenatal hazard are generally designed to include at least one dose group that elicits some degree of maternal toxicity. Endpoints of maternal toxicity, which are defined by study protocol, could include, for example, morbidity or mortality, gestation length, clinical observations, body weight, body weight change, food or water consumption, organ weights, gross necropsy, and/or histopathology data. There is a high degree of correlation between maternal condition and the status of the litter, which is particularly obvious at very toxic dose levels. However, with few exceptions (one being when maternal terminal body weight is corrected for gravid uterine weight and demonstrates a primarily maternal versus intrauterine effect), it is not possible to fully distinguish between those effects on *in utero* development which are attributable to direct foetal exposure to the toxicant versus those effects which are due to, or exacerbated by, maternal toxicity. Additionally, adverse effects on the developing organisms are, regardless of the cause, still toxic manifestations of treatment. For that reason, evidence of maternal toxicity does not automatically negate the observation of foetal toxicity at a similar dose level. In some cases, however, the presence of maternal toxicity must be weighed into the data interpretation for the study, for example, when maternal toxicity is so severe that foetal well-being is compromised and information on developmental effects may be difficult to interpret and of limited value.

Mortality and Incidence of Malformations

64. The prenatal developmental toxicity study is designed to ensure that malformed foetuses are not lost to maternal cannibalism (as could happen in a reproduction study). Nevertheless, even the prenatal developmental toxicity study does not allow the researcher to distinguish the source or cause of prenatal mortality. Intrauterine deaths may be the result of malformations that are incompatible with continuing viability. However, only those that occur in later stages of foetal development might potentially provide evidence of developmental anomalies (via the evaluation of late resorptions that are not in advanced stages of maceration). While it may not be possible to make this distinction, the contribution of malformed foetuses to overall effect on litter viability can be appropriately analysed by combining the litter incidence

of foetuses that are malformed, resorbed (early and late), and dead (full term but nonviable at caesarean section) and performing statistical comparisons of group values.

65. The sensitivity of the test for detection of rare events such as malformations is limited, due to the use of a relatively small number of animals. With the normal group sizes of 20 pregnant rats, it is not possible to identify any increase in major malformations unless high dose levels are administered or the substance studied is highly embryo/foetotoxic (Palmer 1981). To assess the developmental toxicity of a chemical, it is therefore important to include information on other developmental effects such as minor anomalies, variations, foetal death and growth. In addition, malformations of organs developing after the period of major organogenesis, *e.g.,* the sex organs and the brain may not be detected if the study is performed according to the prior guideline for the teratology study. An example is the suspected endocrine disrupter dibutylphthalate where exposure during the period of male sexual differentiation resulted in major disturbances in the morphological and functional development of the male reproductive system (Mylchrest *et al.*, 1999). Consequently, in the updated TG 414 the dosing period has been extended to cover at least the period from implantation to one day prior to the day of scheduled kill, which is one day before the expected day of delivery. In the paragraph Principle of the Test is stated that if appropriate, the dosing period should cover the entire period of gestation to the day of scheduled kill in order to include the examination of effects from the pre-implantation period, however, a potential short time high dose response may not be picked up by TG 414.

Litter Size and Mean Foetal Body Weight

66. The affect of litter size on individual and mean foetal body weight for that litter is well established. In polytocous animals, foetal or neonatal weights are generally inversely correlated with litter size, and the upper end of the dose- response curve may be affected by smaller litters and increased foetal or neonatal weight. While the average mean live litter size for untreated animals often fall within a narrow range, wide variations in litter size can occur spontaneously (MARTA and MTA, 1995, 1996). Trends towards larger litter sizes have been observed to occur over time, in part due to breeding programs that select for this trait. In addition to natural variation, litter size at caesarean section or parturition can be affected by prenatal chemical exposure, *e.g.,* through germ cell toxicity, implantation failure, or embryo-lethality. In the statistical analysis of mean foetal or neonatal body weight data, the litter weight should be statistically adjusted for the size of the litter, for example by using co-variate analysis techniques.

Time of Study Termination and Incidence of Skeletal Variations

67. In the prenatal developmental toxicity study, skeletal development is assessed for approximately half of the foetuses in each litter. The extent of skeletal ossification at the time of death (*i.e.,* at caesarean section of the dam, which is scheduled for approximately one day prior to expected delivery) is evaluated for each foetus. Skeletal development progresses along a standard and predictable time line (Spark and Dawson, 1928). Nevertheless, there can be normal variability in the schedule of ossification, *e.g.,* among various laboratory strains. For that reason, it is critical to establish a scientifically justified gestational day for caesarean section. Additionally, it is important to control, as much as possible, the time of day at which caesarean sections are performed, across control and treated groups, since the incidence of delayed ossification (generally classified as skeletal variations) can be directly related to the gestational age of the foetus and may not be an adverse effect of treatment. The epitome of scheduling would entail observation of the exact time of mating and performance of caesarean section at the precisely equivalent gestation duration (in hours) for each dam on study; this is, however, not practicable for most laboratories that conduct large guideline prenatal developmental toxicity studies.

Statistical Power

68. The power of a study, that is, the probability that a study will demonstrate a true effect, is important in the evaluation of prenatal toxicity data. Factors that may influence the statistical power include the sample size used in the study (with the assumption that the litter is the basic unit of analysis), the background incidence of the finding, the variability in the incidence of the endpoint, the robustness of the data, and the method of analysis. With a study size that includes 20 litters per dose group, the minimum detectable change is:

- an increased incidence of malformations 5 to 12 times above control levels,
- an increase 3 to 6 times the *in utero* death rate, and a
- decrease 0.15 to 0.25 times the foetal weight (US EPA, 1991).

69. Statistical significance does not need to be present to validate the biological significance of treatment-related effects. This is particularly true of findings with low incidence (i.e., rare malformations) or high variability, or in situations where the concurrent control data have an unusual incidence profile.

Concurrent/Historical Controls

70. Concurrent control data are required for every study. On the other hand, historical control data, which are generally comprised of well- characterised negative (vehicle) control data from multiple studies, are not required, but may nevertheless be available and considered useful and appropriate for interpreting study findings. Comparison of concurrent study control data with the data from treated animals should always take precedence over comparison with historical control data. If historical control data are used, the most appropriate of these are from studies conducted in the same laboratory, within a reasonable amount of time prior to the study being interpreted (*e.g.,* ± 2 years) in order to avoid genetic drift in the laboratory animal population, and under the same study conditions (*e.g.,* identical species, strain, source, age, vehicle, route and duration of administration, technical personnel, etc.). It is important that the data include sufficient information to render it meaningful in the context of the concurrent study. For example, definitions of terminology should be provided; incidental and continuous data should be fully characterised and summarised with appropriate data ranges, maximum, minimum, median, and mean values; data variance should be addressed. Historical control information that is compiled by animal suppliers or through surveys (or including genetic drift) across multiple laboratories (Clemens *et al.,* 1994, MARTA and MTA, 1995, 1996) can also be useful in some situations and under the appropriate caveats. Overall, the interpretation and use of historical control data requires careful consideration, and the application of scientific judgement and expertise. If historical control data are demonstrably different from concurrent control data, it may be an indication that the study contains some fatal flaw. In the most egregious case, it may not be appropriate to utilise the historical control data in the interpretation of data from treated groups; rather, the more appropriate response might be to repeat the study.

III. POSTNATAL ENDPOINTS: NEONATAL GROWTH, DEVELOPMENTAL LANDMARKS AND FUNCTIONAL/BEHAVIOURAL NEUROTOXICITY

Overview of endpoints assessed

71. A number of guidelines are available that include assessment of postnatal endpoints induced during development. These include the "One- and Two-Generation Study" (TG 415 and 416), the "Reproductive/Developmental Toxicity Screening Tests" (TG 421 and 422), and the "Developmental Neurotoxicity Study" (draft TG 426).

72. The endpoints assessed using the guideline studies are shown in table 2.

Table 2

Endpoints	**TG 415**	**TG 416**	**TG 421 & 422**	**Draft TG 426**
Birth weight	+	+	+	+
Survival-perinatal period	+	+	+	+
Survival-lactation period	+	+	-	+
Survival-adult	-	+	-	+
Growth-perinatal period	+	+	+	+
Growth-lactation period	+	+	-	+
Growth-adult	-	+	-	+
Physical development - sexual maturation	-	+	-	+
Functional development	-	(+)	-	+
Behaviour	-	(+)	-	+
Neuropathology	-	?	-	+
Reproductive functions	-	+	-	-

+ required; (+) optional

Methodological issues

Standardisation of Litter Size

73. Standardisation of litter size by random selection to yield, as nearly as possible, 8 pups (4 females, 4 males), so-called "culling", are often used and the procedure is described in the Test Guidelines for generation studies and perinatal studies. Adjustment of litter size is optional in the TG 416, while the draft TG 426 recommends that litter size should be adjusted to a number close to the mean litter size for the animals used in the study.

74. An argument in favour of standardisation of litter size is that since pup weight is related to litter size it might lead to a more uniform pup weight at weaning. In a study of data from approximately 500 litters standardisation of litter size seemed only to increase the mean pup weight in the litters, while the variation remained unaffected (Palmer 1986). Also, culling may result in elimination of 25-40% of the offspring and may introduce bias for example by random of elimination of runts (Palmer 1986).

75. The rate of growth and maturation of offspring may vary with litter size, but it is not clear whether variations in litter size also affects behaviour (Barlow and Sullivan 1975). Some studies have reported an increase in emotionality in offspring from large litters compared to offspring from small litters, while others studies have found no differences (Lore and Avis 1970, all cited by Barlow and Sullivan 1975). A difference in litter size of 2-3 pups only was not found to be of any significance in any of these studies.

76. If litters are not standardised, litter size should be included in the statistical analysis of the data as a covariant.

Route of Exposure

77. Whenever possible, the route of exposure should mimic the primary expected exposure of human beings, *e.g.*, the test substance can be administered by the oral (via diet, water, or by gavage), dermal, or

inhalation route. Each route poses unique considerations and challenges, in regard to dose selection, technical implementation, and data interpretation.

78. Pharmacokinetic and pharmacodynamic data on the test substance, when available, can be useful in the selection of the most appropriate route of administration, as well as in the selection of dose levels. Actual exposure of the study animals to the test substance is affected by multiple factors; these can include not only the route of administration and the pharmacokinetics and dynamics of the chemical, but also the life stage of the animal over the course of the study, maternal/litter interactions, and toxic responses to treatment. Study paradigms that include reproductive phases are generally subject to wide variations in exposure to the test subjects, as well as uncertainties regarding actual exposures to the offspring. Generally, even when more specific data are not available (*e.g.*, tissue levels of test chemical), it is assumed that an adverse treatment-related response in the parental or immature animals demonstrates adequacy of dose. Often, even in the absence of treatment-related effects in the offspring, it may be assumed that the offspring have been exposed to the test material during gestation (via maternal circulation) and lactation (via maternal milk). With most exposure routes, direct exposure of offspring to test substance will occur to some extent (for example, pups will begin to eat treated feed during their second week of the lactation period, and on a mg-test-substance per kg-body-weight basis may actually be consuming a higher dose than adults; in inhalation studies that utilise chambers which expose the whole litter, pups receive inhaled and dermal doses). However, when there is a need to ensure adequate exposure of the offspring during critical stages of development (*e.g.*, when assessing effects on the developing nervous or immune system) and/or to quantify exposure to pups, it may be necessary to consider the use of direct gavage dosing of pups during some stages of the pre-weaning period**.** Careful consideration should be given to the impact of such procedures on toxic response and data interpretation. When dosing via food or drinking water, the treatment is easily continued through the period of birth and in the neonatal period. As the pups gradually starts to consume food and water from around postnatal day 14, the exposure in the last part of the lactation period will be partly indirect via maternal milk and partly direct. Oral gavage or parenteral administration may have to be postponed for one or a few days around the period of birth.

79. For industrial chemicals with a high vapour pressure, *e.g.*, organic solvents, the major route of work-place exposure is inhalation. With inhalation exposure there is no first-pass effect in the liver, *i.e.*, the chemical in the blood stream reach the placenta before the liver. Two inhalation exposure techniques may be used: whole body exposure and head/nose-only exposure. The major disadvantage of whole body exposure is that other routes of exposure cannot be excluded: dermal exposure is evident but also oral exposure as a result of grooming activities is possible. The biggest advantage of this exposure technique is the freedom of movement that is allowed to the animals. With head/nose-only exposure techniques additional exposure by the dermal or oral route are avoided, but it is necessary to restrain the animals. This may lead to unacceptable stress during prolonged exposures and influence the results of prenatal exposure. Therefore, whole body exposure is recommended.

80. Inhalation exposure of the dam or pups only may be a possibility since limited studies have indicated that removing the dam from the litter for six hours a day only causes a slight and reversible weight decrease during the first days. However, more studies illuminating the consequences and practical problems related to inhalation exposure of the dam only are needed before this possibility can be favoured.

Cross-Fostering

81. Postnatal effects of pre- and postnatal treatment may be the result of interference with the offspring prenatally, postnatally, or both. Developmental changes may not only be caused by the chemical affecting the developing organism before and after birth, but may also be induced via effects on the mother. For example, lactation or maternal care may be affected and potentially any alteration in maternal physiology or behaviour

may in turn alter the behaviour of the offspring. To control for this, cross-fostering techniques have been developed where the prenatally exposed litters are reared by non-exposed mother and vice versa. This obviously requires more animals and demands more resources, and is therefore not required in screening studies but used in further studies.

Age of Pups

82. The age of the offspring may be decided from the time of birth, *i.e.*, post partum age, or from the time of mating, *i.e.*, gestational age. Studies have shown gestational age to be a better predictor of age of appearance of developmental landmarks in the pre-weaning period than postnatal age, especially if the length of gestation periods differs among groups (Hughes 1986, Raimondo and Draghetti 1990). Gestational age has been used in several studies of developmental neurotoxic effects (Goodlett *et al.*, 1987, Kelly *et al.*, 1988, Hass *et al.*, 1994a, b, Hass *et al.*, 1995). If the expected delivery day, *e.g.*, gestation day 22 for Wistar rats, is designated postnatal day 0 for all offspring the gestational age is used, but the "age" also relates to the time of birth.

Time of Testing

83. The response of an animal in a behavioural test will depend on the time of day the test is carried out. Rats are nocturnal animals and waking of animals during their normal sleep period is hardly likely to produce typical behaviour. Therefore, nocturnal animals should be tested at night or more conveniently reared under reverse lighting conditions and tested under red light during daytime (Barlow and Sullivan 1975). Unfortunately this is very often not the case in developmental neurotoxicity studies and that may explain some of the variation between laboratories. Therefore, rats should normally be tested in their dark period, and control and treated animals should be tested alternately two to four at a time with no fixed rule.

Experiences of Offspring

84. The experience of offspring during infancy may affect their later behaviour. For example frequent handling of rats during infancy will alter their physiological response to stress (Levine *et al.*, 1967, Meaney *et al.*, 1988, 1991), and their behaviour in tests for emotionality and learning (Levine and Broadhurst 1963, Nunez *et al.*, 1995, Meaney *et al.*, 1988, 1991). Animals reared in enriched environment may not only show behavioural changes but also have a heavier cerebral cortex, increased cholinesterase activity and altered levels of brain amines (Rosenzweig and Bennett 1969).

85. In order to control for environmental experiences during infancy, the conditions under which the offspring are reared must be standardised within experiments with respect to variables such as temperature, humidity, noise level, lighting, cages, handling and cage cleaning (Barlow and Sullivan 1975).

Sex of Offspring

86. Sex differences in offspring behaviour are very common. It is therefore important that both sexes are tested and that the results from the two sexes are separated in the analysis of results from developmental neurotoxicity tests, otherwise effects may be masked.

Experimenters Influence on Results in Behavioural Tests

87. Studies have shown that the expectations of the experimenter in some cases may have a significant influence on the results obtained in behavioural studies (Rosenthal and Fode 1963). To rule out this influence totally, the experimenter should have no knowledge of which treatment group the animals belong to. The administration of this may cause some practical problems, for example in Good Laboratory Practices-studies

(GLP) or when the experimenters are involved in analysing and evaluating the data during the study period. However, testing with no indication of treatment group should be considered to avoid experimenter bias.

Test Automation

88. Automation of behavioural tests may reduce observer bias and the need for manpower, and increases the amount of information per experiment, allowing detailed data analysis for a better interpretation of results. Therefore, automation of the recording of the endpoints should have a high priority. Automated methods are required for the assessment of activity in both the OECD and the US EPA Test Guideline for the developmental neurotoxicity (draft TG 426). Commercial equipment is available for a number of behavioural tests.

Groups for Behavioural Assessment

89. The assessment of postnatal development and behaviour in the draft TG 426 includes the same groups as the one- and two-generation study, i.e. three treatment groups and a control group. However, in groups where clear-cut developmental toxicity effects (*e.g.*, markedly decreased litter size or increased postnatal death) are observed, behavioural testing may not be possible due to lack of pups or because the health of surviving pups are to compromised to perform behavioural testing. In addition, a meaningful interpretation of neurotoxicity may not be possible. In such cases, behavioural testing may not be considered necessary for the risk assessment.

Physical and Functional Developmental Landmarks

90. The physical development of the offspring is normally followed by recording body weight several times during the pre-weaning period and once or twice per months after weaning.

91. Other physical and functional parameters are followed by recording so-called developmental milestones or developmental landmarks. These observations often show “when” rather than “if”, the various landmarks first appear and are used to assess delayed or accelerated developmental time courses for the specific parameters being studied (Lochry *et al.*, 1986). These tests evaluate the presence or absence of each parameter, usually over a period of successive days, beginning prior to, or approximately on the day of expected development.

92. Examples of frequently suggested *physical developmental landmarks* are ear unfolding, first coat, upper and lower incisor eruption, eye opening, full coat and onset of puberty. The reliability of observing six physical landmarks (ear unfolding, first coat, upper incisor eruption, lower incisor eruption, eye opening, and full coat) among different observers was assessed in a study by Hughes and Palmer (1986). As a result of this evaluation, the three reliable physical parameters, *i.e.*, ear unfolding, upper incisor eruption, and eye opening were selected for use. Registration of these physical landmarks is not required in TG 416 or TG 426, but should be considered when appropriate.

93. In the TG 416, *anogenital distance* (AGD) should be measured at postnatal day zero in F2 pups if triggered by alterations in F1 sex ratio or timing of sexual maturation. The AGD is longer in males than in females. Several studies have shown that hormones and suspected endocrine disrupters may change the AGD (*e.g.*, Gray *et al.*, 2001, McIntyre *et al.*, 2001). AGD can for example be measured using a slide gauge or a stereomicroscope with measuring scale. It is import to establish clear criteria for the measurement for example from the centre of anus to the centre of the genital bud. In addition, the handling of the animals during the measure should be careful to avoid variation in the measure by for example stretching the region in some animals more than in others.

94. Assessment of *nipple or areola retention* in male rat offspring is not included at present as a parameter in OECD Test Guidelines, however, the measure are sensitive to exposure to hormones and some hormonal disrupters (especially anti-androgens) **(**Gray et al., 2001, McIntyre et al., 2001). In female offspring 12 areolas are normally visible around postnatal day 13 while very few or none are visible in male offspring. As the development of fur in the animals makes it difficult (impossible) to see the areolas a few days later it is important to establish the correct time for the assessment in the animals used for the study. Often only the presence or absence is registered in the males instead of the numbers per male. In some rat strains the frequency in control male rats are low, *e.g.*, below 2% and in such strains assessment of presence or absence may be of sufficient sensitivity. However, in other rat strains control values up to around 30% have been reported (*e.g.*, Long Evans rats, Hellwig *et al.*, 2000) and in such cases assessment of presence only will be rather insensitive. It is also possible to use the number of areolas in each male for the assessment and this is recommended, especially when the control values are high.

95. Assessment of *sexual maturation* is included in TG 416 and draft TG 426. Assessment of the onset of puberty in females is done by inspection of vaginal opening. In rats, this occurs around postnatal day 30-35. In males, testicle decent or balano-preputial separation may be used as indications of puberty. Observation of testicle descent relies on how the animal is handled during inspection and is rather difficult to assess. Balano-preputial separation corresponds to puberty in male rats (Korenbrot *et al.*, 1977) and is the endpoint included in the draft TG 426 and the TG 416. In the rat, this occurs around postnatal days 40-45. Both the age and the body weight of the animal at sexual maturation should be registered. Registration of sexual maturity in males and females requires some training in order to ensure that the animals are scored similarly each time. For that reason it is also preferable that the assessment is performed by one person (or a few persons) using similar criteria each time.

96. The *functional development* can be assessed for example by registration of the time of emergence of the surface righting reflex, negative geotaxis reflex, auditory startle reflex and air righting reflex (table 3). The draft TG 426 requires that two measures are to be registered equally spaced over the pre-weaning period. Registration of functional endpoints requires some training in order to ensure that the animals are scored similarly each time. For that reason it is also preferable that the assessment is performed by one person (or a few persons) using similar criteria each time.

97. Early handling during inspection of physical or functional landmark may influence the behaviour of the animals later in life. Therefore, litters from all groups should be investigated similarly as it may introduce bias if animals are only investigated until positive.

Table 3. Measures of reflex development and neuromotor abilities

Test	Description	Endpoints	Age at testing	Ref.
Surface righting reflex	Placed in supine position on flat surface. Time to right or yes/no	Neuromotor abilities	2-4 days	1
Negative geotaxis reflex	Placed head downward on inclined plane. Time to turn 180°	Reflex development Neuromotor abilities	7-10 days	2
Homing reflex	Placed between home bedding and clean bedding. Time to choose and the choice	Reflex development Neuromotor abilities, Sensory function (odour)	6-10 days	2

Air righting reflex	Dropped from supine position. Land on four feet, yes or no	Reflex development Neuromotor abilities	12-17 days	1,2
Auditory startle reflex	Presented to sudden noise. Startle response, yes or no	Reflex development Neuromotor abilities	10-14 days	3
Rotarod	Placed on rotating rod at fixed velocity. Time on rod and yes/no to fixed period of time	Neuromotor abilities	Usually after weaning	1,2, 9
Accelerod	Placed on rotating rod with increasing velocity. Time on rod	Neuromotor abilities	Usually after weaning	4
Grip strength	Duration of hanging on thin wire or special apparatus. Fore- and hind-limb grip strength	Neuromotor abilities	13 days or later	2,5
Swimming ontogeny	Scoring different stages of deve-lopment. Body position, limb usage, direction of locomotion	Reflex development Neuromotor abilities	Usually between day 6 and 25 days	6,7
Hind-limb splay	Dropped from 30 cm above a sheet of paper. Distance between prints of hind paws	Neuromotor abilities	Usually after weaning	8

1) Altman and Sudarshan 1975, 2) Adams 1986, 3) Buelke-Sam and Kimmel 1979, 4) Jones and Roberts 1968, 5) Meyer *et al.*, 1989, 6) Schapiro *et al.*, 1970, 7) Vorhees 1983, 8) Gerber and O'Shaughnessy 1986, 9) Kaplan and Murphy 1972

Measures of Motor Activity

98. According to the draft TG 426, motor activity should be monitored using automated recording apparatus at least once for each of the pre-weaning, post-weaning, and young adult periods.

99. There are basically three categories of activity measurements:
1. Short term (2-10 min);
2. Longer-term activity; and,
3. Circadian activity.

The open field and the hole-board have generally been used to measure short term activity, while automated devices (photo cells) such as the figure-8 maze, radial arm maze, and cages similar to home cages have been used for measuring activity over longer time periods (table 4).

Table 4. Measures of motor activity

Test	Description	Endpoints	Ref.
Open field	Square or circular box. Latency to leave centre area. Locomotion, rearing	Short term activity	1,5,6

Hole board	Box with holes in floor. Sometimes objects underneath.	Short term activity Exploration	1
Radial arm maze	Eight arms extending radially from central hub	Locomotion Learning	1
Figure-8 maze	Several interconnected alleyways - forms a figure 8	Activity - usually 1 to 24 hours	1,2,3
Home cage	Normal cages within series of infrared beams	Spontaneous long term activity	4

1) Adams 1986, 2) Reiter and MacPhail 1982, 3) Crofton *et al.*, 1993, 4) Gårdlund *et al.*, 1991, 5) Denenberg 1969, 6) Schiorring 1979.

100. Figure 8-maze is a widely used system (Reiter and MacPhail 1982). It has a good inter-laboratory reliability and it is sensitive enough to detect a 40% change in activity between groups with a group size of 10. It can identify reference compounds (d-amphetamine, chlorpromazine) producing both increased and decreased activity.

101. The recommendations given in the draft TG 426 relates mainly to the testing of longer-term activity, for example during 30 minutes and should, if followed, provide data that allows evaluating the potential for effects on motor activity and possibly habituation.

Motor and Sensory Functions

102. According to the draft TG 426, motor and sensory function should be examined in detail at least once for the adolescent period and once during the young adult period.

103. Neuromotor abilities are often evaluated in regard to the ontogeny of particular reflexes or co-ordinated movements. Measures of reflex and motor development are the most widespread of all functional endpoints assessed in behavioural teratology studies (Adams 1986; Buelke-Sam and Kimmel 1979). The procedures for most of the commonly used measurements have been described in several reviews (Barlow and Sullivan 1975; Adams 1986), therefore only a brief description is given in table 3.

104. Until recently, measures of sensory function in developmental toxicity studies have been quite gross, measuring the presence or absence of response, rather than the magnitude of the response. During the last decade, however, more sophisticated automated behavioural techniques allowing quantitative assessment of function have been developed (table 5).

Table 5. Measures of sensory function

Test	Endpoints	Comment(s)	References
Startle response	Hearing ability,	Qualitative measure, low sensitivity	1
Negative geotaxis reflex	Vestibular system Reflex development	Quantitative measure, large variability	2
Air righting	Vestibular system	Qualitative measure	1,3

	Reflex development		
Rotarod	Vestibular system	Quantitative measure	4
Pre-pulse modification of startle reflex	Hearing ability Vision Tactile	Quantitative measure, reflex modification	1,5
Eye-blink conditioning	Hearing ability Vision Tactile	Quantitative measure, reflex modification	6
Conditioned avoidance response	Hearing ability Vision	Quantitative measure, requires learning period	7
Discriminative avoidance learning	Hearing ability	Quantitative measure, requires learning period	8,9
Operant conditioning	Hearing ability Tactile-kinaesthetic	Quantitative measure, requires learning period	10

1) Adams 1986, 2) Pryor *et al.*, 1983a, 3) Altman and Sudarshan 1975, 4) Bogo *et al.*, 1981, 5) Crofton *et al.*, 1990, 1994a, 1994b, 6) Stanton and Freeman 1994, 7) Pryor *et al.*, 1983b, 8) Gabriel *et al.*, 1991, 9) Bushnell *et al.*, 1994, 10) Elsner 1991.

105. The presence of a *startle reflex* is often used as a gross measure of hearing ability in laboratory animals. The whole body startle response is a characteristic sequence of reflexive muscle movements elicited by a sudden intense sensory stimulus. In the rodent, this is typically a loud sound or an air puff. In the rat, an acoustic signal must reach an intensity of at least 90 dB within 12 msec of its onset to elicit the startle response and therefore the sensitivity of the methods is limited. Absence of startle may not mean that the animal cannot hear since lesion of the area in the brain which is thought to be the startle instigating centre reduces or abolishes acoustic startle even though the animal can still hear as judged by other criteria.

106. Performance in the *negative geotaxis test* or investigations of the *air righting reflex* may give some indication of vestibular dysfunction. In negative geotaxis test, the rat is placed facing up in the middle of a screen, which has an angle of 60^{o} with the base, and the screen is rotated slowly until the rat is facing down. The time to reorient to within 45^{o} facing up is measured. Although this test works well for assessing sensorimotor function in young animals, it appears to have limited usefulness in tests of adults. The main problem is its very large variability (Pryor *et al.*, 1983a). When investigating the air righting reflex, the animal is held supine and dropped from a height of 30-40 cm. Normal rats (age>16 days) are successful in righting themselves in the air while deficits in the vestibular system can lead to back or side landing.

107. Balance and co-ordinated movements can be evaluated in the *rotarod test* (Bogo *et al.*, 1981) which is a task performed on a motor-driven, rotating rod. The rat is placed on the rod and is required to maintain balance for a certain trial length. The test is most often performed by accelerating the rotarod gradually from zero until the subject falls off the rod. This approach yields a quantitative measure of performance capability, and is believed to be more sensitive than the non-accelerating version. However, motor disturbance must be ruled out before impairment can be ascribed to the sensory system.

108. Quantitative behavioural measurements of sensory function mainly use reflex modulation and conditioned behaviour. By using an olfactory, auditory, or visual stimulus as stimulus it is possible to

examine whether the rat is able to perceive the presence of it. If the response is unchanged after a certain stimulus has been presented, it is interpreted as if the rat did not perceive that stimulus. The stimulus can be varied in quality and intensity, thus allowing the determination of a detection threshold.

109. The *startle reflex modification* procedure makes use of the fact that the startle reflex can be inhibited by an event ("pre-pulse") shortly preceding the eliciting stimulus (Crofton and Sheets 1989). This phenomenon is referred to as "pre-pulse inhibition". The pre-pulse must be too weak to produce measurable startle by itself, and is presented 30-100 msec before the startle-eliciting stimulus. The amplitude of the startle response is reduced, while the latency remains unchanged or increased. A stimulus that produces good inhibition does not necessarily produce good startle, *i.e.*, a visual stimulus will inhibit auditory startle in the rat, but does not produce startle in this species. The startle response and pre-pulse inhibition phenomenon exists in all mammalian species as does the principle that the inhibitory efficacy of the pre-pulse is directly proportional to its intensity.

110. *Eye-blink conditioning* involves pairing of conditioning stimulus (typically a pure tone) and an unconditional stimulus (typically a brief air puff to the eye). The animals have to learn to respond to the conditional stimulus by blinking (see Stanton and Freeman, 1994 for review). Methods for studying Eye-blink conditioning exist in rats, rabbits and humans. Comparisons of studies involving humans and animal models indicate that the effects of a number of biological variables are similar in different species and the evidence that is available thus far suggests that the biological mechanisms are similar across species as well. The Eye-blink conditioning model has mainly been used for studying learning effects. However, by using conditioned stimuli such as tones with variable frequency or light in different intensity it possible to assess effects on sensory systems. This approach needs to be further investigated.

111. By operant conditioning techniques animals are trained to act in certain ways in response to different stimuli. Some of the most sensitive methods for testing sensory functions use operant conditioning techniques, which require discrimination of light intensity and wavelength, tone intensity, or pitch.

112. *Conditioned avoidance response (CAR)* involves training of rats to climb or pull a pole to avoid a foot shock, which is preceded by a warning signal (conditioned stimuli) (Pryor *et al.*, 1983a, b). The conditioned stimuli can be a tone where the frequency is variable, or a change in the intensity of light. If the rats climb the pole during the warning signal the trial is terminated. The pole-climbing response after presentation of a particular warning signal is an indication that the rats were able to perceive the signal. This test has revealed that toluene-exposed rats had a hearing deficit of frequencies above 8 kHz. Experiments with positive control substances have indicated that the CAR procedure is sensitive and capable of detecting specific sensory loss. Further development is needed to demonstrate the utility as a rapid and cost-effective procedure.

113. *Discriminative avoidance* learning has also been investigated in a rabbit model (Gabriel *et al.*, 1991) where the animals were trained to respond to a tone conditional stimulus in order to avoid a response-terminated foot-shock. In a model for testing of vigilance designed by Bushnell *et al.*, (1994), adult rats were trained to press a lever when they detected an auditory signal. By using auditory, visual, tactile, or other stimuli as the conditioned stimulus in these two models sensory effects can be assessed.

114. An operant conditioning technique was developed for assessment of rat's tactile-kinaesthetic performance by Elsner (1991). A force sensitive lever was used for training rats to press within a force band without delivering any additional feedback information than the rats' own tactile kinaesthetic perception. In the beginning the force band was wider than later on. The tactile-kinaesthetic system of rats exposed to methyl mercury appeared to be deficient, including a tendency to coarse motoricity, as is observed in children suffering from minimal brain damage.

Measures of Learning and Memory

115. Learning and memory function should be tested twice during the lifetime. In the draft TG, testing of associate learning and memory is required post-weaning and for young animals. The use of different tests or different animals is recommended because repeated testing in the same test may decrease the sensitivity of the test. The time of post-weaning testing could be shortly after the end of exposure, *i.e.*, at days 24-28 and just before termination of the study around day 60-80.

116. The test for cognitive function, *i.e.*, learning and memory, should be based on associative learning and should include the possibility to assess changes across repeated learning trials as well as memory function. The radial arm maze and the Morris maze have been used to demonstrate the effects of several positive control substances. Schedule controlled operant conditioning in Skinner boxes may detect very subtle effects, but may also be too time-consuming for the initial testing of learning and memory.

117. A number of different tests have been used for assessing effects on learning and memory (table 6). Many of them are based on the animals moving, using their senses etc. and therefore an impaired performance in a learning test may reflect other behavioural effects than on the learning abilities. It is also important to keep in mind that a measure of memory cannot be interpreted in the absence of a measure of learning obtained in the same test (US EPA 1990).

Table 6. Measures of learning and memory abilities

Test	Description	Endpoints	Reference
Auditory startle habituation	Loud tone presented at regular intervals. Amplitude of startle response.	Habituation	1
Eye-blink conditioning	Responding to conditional stimulus by blinking.	Habituation	2
Head dipping in hole board	Box with holes in floor. Sometimes objects underneath	Habituation Exploratory behaviour	3
Passive avoidance	Two-compartment shuttle-box, one light and one dark compartment (brief shock). Latency to enter dark compartment.	Learning and memory	3,4,8
Taste or odour aversion	A taste or scent is paired with an aversive stimuli such as the effect of a LiCl injection (nausea)	Learning and memory	3
Operant conditioning (Skinner box)	A particular response leads to a rein-forcement (food, water)	Learning and memory	3,5
Active avoidance	A particular response turns off or delays onset of aversive stimuli.	Learning and memory	3
Radial arm maze	Eight arms extending radially from central hub. Reinforcement in each arm.	Learning and memory	3
Morris maze	Use spatial cues to find platform hidden	Learning and	6,7

	under the surface in large circular pool.	memory	
T-maze	Select the arm with reinforcement.	Learning and memory	4,8

1) Crofton *et al.*, 1993, 2) Stanton and Freeman 1994, 3) Adams 1986, 4) Wier *et al.*, 1989, 5) Hass *et al.*, 1994a, 6) Kelly *et al.*, 1988, 7) Hass *et al.*, 1995, 8) Lochry and Riley 1980.

118. Most learning paradigms currently in use belong to one of three basic categories: appetitive, escape or avoidance (Lochry *et al.*, 1986). Appetitive learning usually involves food or water deprivation and is frequently conducted in mazes, while escape learning uses for example water or shock as the aversive stimulus.

119. Basically the experimental procedures can be described as habituation, classical conditioning and instrumental or operant conditioning.

120. At present, two measures of habituation that have been used are *auditory startle habituation* and habituation of *head dipping in a hole board*. Measurements of behavioural habituation is a classical technique in psychopharmacological research, but has only during the last decades been used in toxicology. A standardised, quantitative test was included in the Collaborative Behavioural Teratology Study conducted by the National Center for Toxicological Research and demonstrated effects of prenatal exposure to methyl mercury (Kimmel and Buelke-Sam 1985).

121. The classical conditioning paradigm is used in *passive-avoidance*, *taste-aversion* and *odour-aversion* learning. Passive avoidance procedures are technically easy and rapid, however, they have come under criticism in recent years because they tend to produce variable results and are sensitive to a number of variables such as activity level. Instrumental conditioning can be assessed for example in the *active avoidance task*, in mazes, and in the Skinner box (*operant conditioning*).

122. Many water mazes are simple swimming versions of previously developed appetitive mazes, while others have been developed as water tasks (review from Vorhees *et al.*, 1991). Water mazes as compared to dry mazes have valuable features. They are free of appetitive or taste confounds and, within limits, also free of influence produced by differences in body mass (Vorhees *et al.*, 1991). Water mazes also show no evidence of being significantly influenced by olfactory cues probably because olfactory cues cannot carry over from one trial to the next. Water mazes have some limitations as well. They rely upon aversive rather than positive reinforcement to motivate performance and aversively motivated behaviours may induce stress. However, appositively motivated procedures generally rely upon prior food and water deprivation, which themselves may induce more stress.

123. Spatial learning processes have been strongly linked, in the rat, to the hippocampus and related structures (Olton *et al.*, 1979, Morris *et al.*, 1982). Spatial navigation in the *Morris water maze* requires rats to locate, on the basis of distal, extra maze cues, a small platform hidden under the surface of the water in a large tank (Morris 1981, Morris 1984). The task has been shown to be sensitive to a variety of experimental insults to relevant brain structures and to age related impairments. It has been used to demonstrate effects of for example prenatal or early postnatal exposure to ethanol, technical xylene, N-methylpyrrolidone, and malnutrition in rodents (Blanchard *et al.*, 1987, Kelly *et al.*, 1988; Hass *et al.*, 1994a, b, Hass *et al.*, 1995).

124. Investigations of spatial learning in *radial arm maze* have demonstrated effects of chemicals such as ethanol and trimethyltin (Miller *et al.*, 1982), of maternal phenylketonuria and prenatal phenytoin (Weisenburger *et al.*, 1990), and of methylnitrosurea-induced micro-cephaly (Akaike *et al.*, 1988). However, known neurotoxic substances such as triethyllead and chlordimeform have been found not to impair performance in this model (Walsh and Chrobak 1987).

Data interpretation

Relationship of Maternal and Offspring End Points

125. There may be a relationship between maternal and offspring end points, especially during the lactation period where the offspring are depending on the maternal care. The relationship can be documented by the use of cross fostering of treated offspring to untreated mothers and untreated offspring to treated mothers. In the absence of documentation for relationship between maternal effects and developmental effects the effects of the offspring should be considered as treatment-induced and consequently used for the setting of NOAEL for developmental toxicity.

Litter Size, Sex and Mean Pup Body Weight

126. The effect of litter size on pup body weight is well established. In general, pup body weight is inversely related with litter size. This relationship may disappear after weaning and not be present in adult animals. In the statistical analysis, pup body weight should be statistically adjusted for the size of litter, for example by using co-variance analysis technique. However, it should be kept in mind that it is not appropriate to use litter size as a co-variate if there is a significant interaction between treatment and litter size. The sex of the pup also has influence on the body weight. Males are generally slightly heavier than females in the pre-weaning period and the difference increases markedly around and after sexual maturation. If both male and female offspring is included in the statistical analysis sex should be included in the analysis. In general, statistically significant changes in offspring body weights are considered as adverse.

Physical and Functional Developmental Landmarks

127. Delays of physical developmental and reflex ontogeny in the lactation period often occur in the presence of decreased body weight of the offspring and may as such be signs of developmental delay. In cases where changes in physical or functional development are observed in exposed pups without a corresponding change in growth, the changes may reflect specific developmental effects, and not a general delay in development.

a. AGD may be influenced by the size of the animal and this should be taken into account when evaluating the data. The size or length of the pups is normally not measured, but body weights are measured. In some cases, the anogenital index, *i.e.*, AGD divided by body weight, is used. However, body weights of pups may be quite variable leading to a large variation in the anogenital index. This could mask eventual effects on AGD and is therefore not recommended. Instead account of size of the animals should be taken by using including a covariant. Body weight can be used, but this parameter is in three dimensions, while AGD is in one dimension. Consequently, the optimal co-variate seems to be the cube root of the body weight. A statistically significant change in AGD that cannot be explained by the size of the animal indicates effects of the exposure and should be used for setting the NOAEL.

b. Consideration should be given to the possible effect of the treatment paradigm on behavioural testing. The timing of testing relative to the timing of daily test substance administration can have profound effects on the outcome, often dependent upon the pharmacokinetic and pharmacodynamic profile of the chemical. Additionally, it may not be possible to determine whether alterations in neurobehavioral parameters that are observed during the period of treatment are systemic or developmental in nature. Generally, such distinctions can only be made with confidence following the conduct of

additional experiments, *e.g.*, pharmacodynamic, mode of action, phased dosing, or cross-fostering studies. This information is typically not considered crucial for the determination of potential risk to susceptible developing populations; however, it may be important within the context of hazard-based chemical labelling.

c. In the standard developmental neurotoxicity study design, treatment is stopped following post-natal day (PND) 10 or PND 21, while neurobehavioral testing is conducted around the time of weaning, during the time of puberty and young adult growth (functional observational battery only), and again just before termination at approximately PND 60. Apparent reversibility in adult offspring, of effects observed early in life, may be related to compensatory developmental or behavioural processes, and not represent a true recovery. Likewise, findings observed in adult offspring that had not been previously observed in young test subjects should not be discounted for lack of concordance, since they may represent the latent expression of early alterations in neurological development.

Behavioural End Points

128. Validity refers to what a test measures (relevance) and how well it is measured (reliability). Validation of behavioural test batteries using a potential positive control, *i.e.*, a behavioural teratogen has been done several times. Methyl-mercury and ethanol were investigated in the European Inter-laboratory Study group on Behavioural Teratology with methyl-mercury giving the clearest results (Elsner *et al.*, 1986, Elsner *et al.*, 1988). In a study where propoxyphene, chlorpromazine, and vitamin A were investigated because they produce different behavioural teratogenic profiles, vitamin A was judged as the best positive control in the test battery used (Saillenfait and Vannier 1988).

129. A single positive control cannot, however, be expected to cover all types of behaviour (Saillenfait and Vannier 1988). Therefore, positive control data that demonstrate the sensitivity of each of the tests in a battery can be used and the positive control data do not have to be from studies using prenatal exposures (US EPA 1990).

130. The behavioural tests included in test batteries should be individually validated tests and there are several examples of positive control substances in the literature. For example, 2,5-hexanedione can be used as a positive control substance causing effects in Rotarod (Ladefoged *et al.*, 1989). Effects of triethyltin, acrylamide, 2,5-hexanedione, and lacking effect of insulin, carbon tetrachloride, and haloperidol have been demonstrated in the grip strength apparatus (Gerber and O'Shaughnessy 1986). This shows that the method is sensitive to known neurotoxicants, and will distinguish these from effects on plasma glucose levels, liver function, or CNS dopamine blockade.

131. An important aspect of test validity refers to the relevance of the test , *i.e.*, whether a test measures what it is designed to measure. For example, does a change in a learning test always only indicate effects on learning abilities? The answer is no, since amongst others motivation and activity level may influence the performance in a learning test. Therefore, the influence of these factors on the results must be assessed and evaluated before conclusions concerning learning abilities as such can be drawn.

132. The severity and nature of the behavioural effect should be considered. Generally, a pattern of effects (*e.g.*, impaired learning during several consecutive trials) is more persuasive evidence of developmental neurotoxicity than one or a few unrelated changes.

133. The reversibility of effects should also be considered. Irreversible effects are clearly serious, while reversible effects may be of concern if exposure is continuous. However, it is often not possible to

determine whether an effect is truly reversible. The nervous system possesses reserve capacity, which may compensate for damage, but the resulting reduction in reserve capacity should be regarded as an adverse effect. If developmental neurotoxicity is observed only during some time of the life-span compensation should be suspected. Also, effects observed for example during the beginning of a learning task but not at the end should not be interpreted as reversible effects. Rather the results indicate that the speed of learning is decreased.

134. The experience of offspring especially during infancy may affect their later behaviour. For example, frequent handling of rats during infancy may alter the physiological response to stress and the behaviour in tests for emotionality and learning. In order to control for environmental experiences, the conditions under which the offspring are reared should be standardised within experiments with respect to variables such as noise level, handling and cage cleaning. The performance of the animals during the behavioural testing may be influenced by *e.g.*, the time of day, and the stress level of the animals. Therefore, the most reliable data are obtained in studies where control and treated animals are tested alternately and environmental conditions are standardised.

135. A large attempt to validate a standardised protocol of a test battery for intra- and inter-laboratory reliability, and to evaluate experimental design and analysis issues was performed in the Collaborative Behavioral Teratology Study (CBTS) in the USA (Kimmel 1988). The results of the CBTS demonstrated that carefully conducted standardised procedures can result in reproducibility of patterns of behavioural response across laboratories.

136. The sensitivity of tests is influenced by the biological variation across animals and this is important in behavioural studies. However, many behavioural endpoints are measured more than once for the same animal, *i.e.*, a repeated measures design (Tamura and Buelke-Sam 1992), and when analysing such data with repeated measures analysis of variance, the variation across animals and time is taken into consideration and the sensitivity of the test is increased.

137. Investigations indicate that many behavioural tests, such as Morris maze, figure-8 maze, Open field and Rotarod, are able to detect differences around 30-50% from the control value (Hass 1993; Reiter and MacPhail 1982).

138. The CBTS, where six laboratories used a standardised behavioural test battery for investigating the effects of methyl mercury and amphetamine, gave valuable data for estimations of the sensitivity of behavioural tests. With a group size of 16 animals, the coefficient of detection, i.e. the percent change required to detect a significant change using a given alpha (*e.g.*, 0.05)**,** was 7-23% for auditory startle habituation, 2-26% for the operant conditioning technique, and 13-20% for a complex water maze procedure (Vorhees 1985). These results indicate that under standardised conditions behavioural measures are as precise as other scientific measures performed on living animals (Annau and Cuomo, 1988)

139. Four spatial maze learning tests with methyl-nitrosourea-induced micro-cephaly rats were compared in a study by Akaike *et al.*, (1994). In a single T-maze no learning defects were detected, while Biel water maze and Morris water maze showed learning effects at single dose of 5 mg/kg and learning effects were found in radial eight-arm maze after dosing with 3 mg/kg and 5 mg/kg. These results demonstrate that more complex mazes are more sensitive to learning impairment than simple tests. The radial arms maze showed the highest sensitivity for this dosing regime. Studies using other positive control substances are needed to investigate this further.

140. The biological sensitivity of a test relates to the ability to pick up and reflect changes in the function under evaluation. When evaluating results of behavioural tests it is important to give considerations to the complexity of the test in relation to the capabilities of the laboratory animal. An underestimation of the

capabilities of the animals means that the test was too easy and had a low sensitivity, *i.e.*, only animals with major defects would not be able to fulfil the tests. Overestimation means that the control group performed badly and therefore exposed animals could hardly be worse. For example, it was found in a study of pups pre-natally exposed to xylene that the biological sensitivity of Morris maze was satisfying for standard-housed offspring but it was too low (the test was too easy) for enriched-housed animals (Hass *et al.*, 1995). In rats pre-natally exposed to tributyltin and investigated in two spatial learning tasks, *i.e.*, radial arm maze and Morris maze, a clearly retarded acquisition was seen in the radial arm maze while no differences were obtained in the swim maze (Gårdlund *et al.*, 1991). The schedule for testing in Morris maze was, however, very reduced using the same start position for every trial and only one platform position and therefore the sensitivity may have been low.

141. Cross-species validity refers to how well a test conducted in animal species predicts neurotoxicity in humans. These predictions are very important but good data are, however, not often available because if screening is done successfully human exposure should be very limited (Vorhees 1985).

142. At a workshop in the US in 1989, the degree of qualitative and quantitative comparability between human and experimental data was assessed for lead, methyl-mercury, selected agents of abuse, phenytoin, PCB, ethanol, and ionizing radiation (Francis, 1990). For the qualitative comparison the following functions were evaluated: motor development and function, cognitive function, motivational/arousal behaviour, sensory function, and social behaviour. Although a number of limitations were identified with cross-species comparability, the degree of comparability was considered remarkable (Francis *et al.*, 1990).

143. It was not considered possible to make definitive quantitative comparisons at this time, based upon the relationship of endpoints to dose, but there were indications that for some of the agents discussed, cognitive function appeared to be the most sensitive category.

144. Dose-response data were often limited, especially for humans. Comparisons of administered effective doses revealed a wide range of differences across species (up to 10.000-fold difference), while comparisons using internal measurements of dose (*e.g.*, blood or brain levels) showed a remarkable correlation (generally, a 1-2-fold difference).

145. The findings at this workshop supported the assumption that, as for other endpoints of developmental toxicity, functional changes in animal studies indicate that an agent has the potential to alter development in humans.

Limitations in Assessment of Postnatal End Points

146. The available guidelines including assessment of postnatal effects after developmental exposure are mainly focused on growth, developmental landmarks and functional/behavioural endpoints while functional changes in e.g. the immune system is not covered. The behavioural testing includes assessment of the individual animal for a number of relevant behavioural functions, but none of the tests assess two or more animals together. This means that some behavioural endpoints of potential relevance as *e.g.*, sexual behaviour, play behaviour, social interaction among animals and aggression is not assessed using the current Test Guidelines.

147. The reproduction/developmental screening tests (TG 421, 422) do not provide complete information on all aspects of reproduction and development. In particular, it offers limited means of detecting postnatal manifestations of prenatal exposure or effect induced during postnatal exposure.

148. The value of a negative study is more limited than data from generation and teratology studies due to the lower number of animals per group, the shorter period of exposure as well as the limited number of

endpoints measured.

IV. REPRODUCTIVE TOXICITY ENDPOINTS IN ADULTS

Overview of Endpoints Assessed

149. A number of Test Guidelines are available that include an assessment of reproductive endpoints including the "Reproduction/Developmental Toxicity Screening Study" (TG 421), the combined "Repeated Dose Toxicity Study and Reproduction/Developmental Toxicity Screening Study" (TG 422), the "One-Generation Reproductive Toxicity Study" (TG 415) and the "Two-Generation Reproductive Toxicity Study" (TG 416). All of these studies provide some information on the adult reproductive system following exposures to mature animals. In addition, the two-generation reproductive toxicity study provides information on the reproductive system following exposures during all phases of development including exposures *in utero,* during lactation, and continuing through sexual maturation.

Methodological Issues

Examination of Male Reproductive Organs

150. A macroscopic evaluation of the male reproductive organs is included in most reproductive toxicity Test Guidelines. Particular emphasis should be placed on the evaluation for hypospadia. Absolute and relative (*i.e.,* adjusted for body weight) weights of the testes, epididymides, seminal vesicles, prostate and pituitary are recorded. A number of reviews are available outlining the proper techniques for the histological examination of the testes (Russell *et al.,* 1990; U.S. EPA, 1996; Chapin and Conner, 1999). The method used to examine the histology of the testis and epididymis is dependent on the specific reproductive toxicity study protocol that was followed. In protocols that specify dosing of the male for an entire spermatogenic cycle combined with analysis of the sperm (*e.g.*, TG 416) less extensive histological examination is required than in situations where the dosing regimen is less than the spermatogenic cycle and no sperm analyses are conducted (*e.g.,* TG 421, 422). For the one-generation (TG 415) and two-generation reproductive toxicity study (TG 416), the testis and epididymis are immersion-fixed in Bouin's fixative, the tissue is embedded in paraffin, sectioned at 5 µm and stained with hematoxylin and eosin (HE). Evaluation should include information on the caput, corpus and cauda segments of the epididymis. Several aspects of testicular degeneration can be recognised in the HE sections including hypocellularity (decreased numbers of germ cells in the epithelium), vacuolation of the Sertoli cells, formation of multinucleated cells (giant cells, composed of spermatocytes and spermatids), cell death within the seminiferous tubules, and the presence of round spermatids and cellular debris in the epididymis.

151. For shorter-term studies (*e.g.,* TG 421, 422), it is necessary to examine potential injury to the testis at specific stages of the cycle of the seminiferous epithelium. For the stage-specific analysis, the sectioned tissue is immersion-fixed in Bouin's fixative, paraffin-embedded, sectioned at 5 µm and stained with the periodic acid-Schiff technique (PAS). Better resolution is achieved on testes fixed by vascular perfusion and embedded in a water-soluble plastic such as glycol methacrylate. The most commonly used system for classifying the 14 stages in the rat was described by Leblond and Clermont (1952) based on the development of the acrosome and the shape of the elongate spermatid nucleus. PAS stains the developing acrosome on the round spermatids, and 6 of the 14 stages in the rat epithelium can only be distinguished with PAS. For standard toxicology studies, it is adequate to simply assess whether all stages of spermatogenesis are normal, categorising the stage of each tubular cross section of the testis is not necessary.

Sperm Parameters

152. The parameters that are included in the two-generation reproductive toxicity study (TG 416) are sperm number, sperm morphology, and sperm motility. A number of reviews are available outlining the proper techniques for sperm analyses (Russell *et al.*, 1990; US EPA, 1996; Seed *et al.*, 1996; Chapin and Conner, 1999). Typically, sperm motility is measured prior to morphology or count, because motility is critically dependent on temperature. For analysis of sperm motility, samples can be collected from the cauda epididymidis or the vas deferens. The samples can be collected anywhere from room temperature to 37ºC, and can be stored in any physiologic buffered saline solution for up to one hour. For sampling from the cauda, the tissue is placed in a dish with an aliquot of buffer and the cauda is nicked in a few sites with a blade. The sperm then diffuse into the medium and the tissue is removed. For sampling from the vas deferens, a small amount of tissue is placed in a dish with buffer and the sperm diffuse into the medium. In sampling from either tissue, care should be taken to avoid any unnecessary manipulation of the tissue.

153. Sperm motility can be assessed manually or by computer-assisted sperm analysis (CASA) systems. When assessing sperm motility, the samples should be at a temperature of 34-37ºC. The depth of the chamber is critical for an accurate assessment of motility; chambers greater than 20 µm are preferable for rodents. There is a 95% probability of detecting a change of 6% and 4.2% in a sperm motion parameter with a group size of 10 and 20 rats, respectively. In general, 200 sperm should be analysed. A minimum value of 70% motility is acceptable in controls. For manual assessments, the number of motile sperm is counted with a hema-cytometer. One easy method is to count the number of stationary sperm, fix the sample and then count the number of total sperm. For CASA analysis, a sperm is considered motile if the average path velocity is greater than a user-defined threshold, the threshold is determined for a given laboratory.

154. In addition, the percentage of progressively motile sperm is assessed. This measure distinguishes sperm that are simply twitching in place from those that are making forward progress. This can be done during manual assessments. For CASA analysis, progressive motility is defined as the percentage of motile sperm that have a linear index greater than a user-defined threshold, which is selected to distinguish sperm with relatively straight paths from those with more circular paths.

155. Sperm morphology can be assessed from samples from the cauda epipdidymidis or the vas deferens. Generally, sperm morphology is assessed from a sample that has not been collected for the assessment of sperm motility as the bovine serum albumin that is present in the buffer can interfere with the stains that are used to assess morphology. For the assessment of morphology, a small sample is placed on a slide and can be viewed either as a wet preparation or the slide can be air-dried. Drying can cause kinking of the tails so some laboratories prefer viewing wet samples. Typically the samples are stained with Eosin Y, but a variety of stains are acceptable as long as they allow the appropriate viewing of the sperm. The samples are viewed with a light microscope typically at a magnification of 400X.

156. There is no universal classification scheme for sperm morphology. The sperm abnormalities are generally described in terms of abnormalities of the head and tail, and can include a head that has too little or too much hook and a tail that is frayed or coiled, multiple tails, misplaced mitochondria, or there can be a residual drop of cytoplasm remaining on the tail. In addition, the neck at the junction of the head and tail can be a weak spot. In rodents, there is a high proportion of normal sperm so usually 200-400 sperm per rat are evaluated. The statistical power is high and can easily detect a change when comparisons are made between normal and abnormal sperm.

157. In the US EPA guideline OPPTS870.3800 samples of sperm from the distal cauda epididymis (or the proximal vas deferens) shall be collected for sperm-measurement (*e.g.*, the evaluation of the percentage of progressively motile sperm and sperm morphology). The entire cauda epididymis shall be minced in saline to

enumerate the total number of sperm. Histopathology of sex organs and oestrus cycle is evaluated and milestones for sexual maturation recorded.

158. In the two-generation reproductive toxicity study (TG 416), cauda epididymal and testicular sperm counts are recorded. The cauda sperm count is a measure of sperm storage. Testicular sperm counts generally refer to enumeration of detergent and homogenisation-resistant spermatid heads. During spermiogenesis, the nuclei of the spermatids become highly condensed and the nuclear material becomes cross-linked. The mature spermatid nuclei, unlike less mature spermatids, are very resistant to homogenisation. Thus, this technique provides a reliable estimate of the number of spermatids in the later stages of spermiogenesis. Sperm counts can be determined at the time of necropsy or the tissue can be frozen for later analysis. For both cauda epididymal and testicular counts, the tissue is homogenised in the presence of detergent such as Triton X-100®. Sperm counts can be obtained by use of a hema-cytometer or CASA. Each laboratory should determine the variability associated with the counts and the level of statistical power.

Examination of Female Reproductive Organs

159. The female reproductive organs are examined macroscopically in all reproductive Test Guidelines (TG's 421, 422, 415 and 416). The number of implantation sites in the uterus is recorded. The absolute and relative (*i.e.*, adjusted for body weight) weights of the ovary, uterus, vagina and pituitary are recorded, and these organs are examined for histopathology. Detailed histological examination of the ovaries should cover the follicular, luteal, and interstitial compartments of the ovary, as well as the epithelial capsule and ovarian stroma.

Oocyte Quantitation

160. In the two generation reproductive toxicity study, a quantitative evaluation of the primordial follicles is conducted for the F1 females. The methods for the quantitative analysis have recently been reviewed by Heindel (1999). For this analysis, the ovaries are fixed in Bouin's solution for 12-24 hours, embedded in paraffin in a longitudinal orientation, serially sectioned at 6 µm, and stained with HE. The ovaries can also be fixed in Kahles fixative, embedded in Paraplast®, and stained with Weigerts hematoxylin followed by picric acid-methylene blue counter stain.

161. The follicles can be categorised into three classes.

- Small follicles are defined as an isolated oocyte or an oocyte that is surrounded by a partial or broken granulosa cell.
- Growing follicles are defined as an oocyte that is surrounded by a multi-layered, solid mantle of granulosa cells, and,
- antral follicles are defined as an oocyte that is central and a fluid-filled space bordered by hundreds of layered granulosa cells.

162. For the TG 416, the small and growing follicles can be combined. A standard procedure is to evaluate 10 females in a treatment group, and to count the follicles in every tenth sequential section beginning with the section when follicles are first noted. However, some researchers have shown that the same statistical power can be achieved by counting fewer sections and/or by evaluating fewer animals. Each laboratory should determine the appropriate number of animals and number of sections to evaluate.

Vaginal Cytology

163. Vaginal cytology is evaluated to determine the length and normality of the oestrus cycle in the P and F1 females in the two-generation reproductive toxicity study. A recent review of the methods has been

conducted by Cooper and Goldman (1999). Vaginal smears must be collected for at least two weeks for an accurate determination of cycle length. There are several methods available for collection of the smears. One method involves collection by lavaging the vagina with an eyedropper containing approximately 0.25 mL water or physiological saline and drawing the fluid back into the eyedropper. Only the tip of the eyedropper is inserted into the vagina; stimulation of the cervix should be avoided to ensure that pseudo-pregnancy does not occur. The fluid is expelled evenly onto a microscope slide. The smears can be viewed immediately at approximately 100X magnification, stained for subsequent analysis or viewed wet and stored for subsequent staining. Vaginal smears are typically stained (*e.g.*, 1% aqueous toluidine blue O, methylene blue).

164. The entire smear should be examined as the cell distribution may be uneven. The smears are evaluated based on the types of cells present. In the rat, the oestrus cycle lasts for 4-5 days. Each cycle has three distinct phases including met-oestrus/di-oestrus (2-3 days), pro-oestrus (1 day), and oestrus (1 day). During di-oestrus, the smear contains a mixture of cell types; leukocytes are the predominant cell type with a varied number of cornified epithelial cells. During pro-oestrus, the smear contains a predominance of clumps of round, nucleated epithelial cells, and during oestrus, the smear contains mainly cornified cells. Different laboratories may classify the smears according to different criteria, and as such it is important that the cell types are recorded consistently on a daily basis and that the females are monitored over an extended time for comparison of the smears with the cycling status of the females. In addition, it is important that the smears are collected at the same time each day to reduce variability and enable identification of cycling patterns.

Reproductive Performance

165. Reproductive performance is the ability of male and female animals to successfully mate and produce viable offspring, and is assessed in the TG 421, 422, 415 and 416. Information on reproductive performance is generally expressed as indices that are ratios derived from the data collected in the studies. The major indices are described in the following table.

Index	**Calculation**	**Definition**
Male Mating Index	$\frac{\text{No. of males with confirmed mating}}{\text{Total No. of males cohabited}} \times 100$	Measure of male's ability to mate
Female Mating Index	$\frac{\text{No. of sperm-positive females}}{\text{Total No. of females cohabited}} \times 100$	Measure of female's ability to mate
Male Fertility Index	$\frac{\text{C of males impregnating a female}}{\text{Total No. of males cohabited}} \times 100$	Measure of male's ability to produce sperm that can fertilise eggs
Female Fertility Index	$\frac{\text{No. of pregnant females}}{\text{No. of sperm-positive females}} \times 100$	Measure of female's ability to become pregnant
Gestation Index	$\frac{\text{No. of females with live born pups}}{\text{No. of pregnant females}} \times 100$	Measure of pregnancy that provides at least one live pup
Survival Index	$\frac{\text{No. of live pups (at designated time)}}{\text{No. of pups born}} \times 100$	Measure of pup survival which is calculated at several times during lactation

Data Interpretation

Male Reproductive Organs

166. Absolute and relative weights of the male reproductive organs should be considered as a decrease in absolute weight may occur that is not necessarily related to a reduction in body weight gain. Since there is low inter-animal variability in testis weight, a significant change in absolute testis weight (increase or decrease) can indicate an adverse effect. Changes in testis weight can be due to damage to the seminiferous tubules or a variety of other cases including oedema, inflammation, cellular infiltration, Leydig cell hyper-plasia or fluid accumulation due to blocked efferent ducts. The weights of the prostate and seminal vesicles are androgen dependent and therefore may reflect changes in testicular function or endocrine status. Pituitary gland weight can also be an indicator of reproductive status. However, the pituitary contains many cell types that are responsible for the regulation of several physiologic functions including some that are separate from reproduction. Therefore, gonadotrophin-specific histopathological evaluation may be useful to determine which cell types are affected.

167. Histopathological evaluations of the reproductive tissues are relatively sensitive indicators of damage and are valuable for the assessment of male reproductive toxicity. Histopathological findings are generally classified according to qualitative criteria and the data are presented as the number of animals affected within a dose group. There is no standardised method of further quantifying the extent of the damage. In general, histopathological findings should be considered in light of the specific test protocol and in conjunction with information on sperm parameters and reproductive performance. Depending on the test protocol that was used, and the extent of the damage, there may not be an obvious relationship between the histopathological findings and fertility. For short-term studies (*i.e.*, TG 421 and 422) in which the animals are treated for less than the duration of the spermatogenic cycle, an effect on spermatogenesis may not have had adequate time to become evident as reduced sperm counts that affect fertility. In general, any dose related significant histopathological change indicates that there is a potential for similar effect in humans.

Sperm Parameters.

168. Two measures of sperm motility are usually calculated. The percentage of motile sperm is defined as the number of motile sperm/total number of sperm X100, and the percentage of progressively motile sperm is defined as the number of progressively motile sperm/total number of sperm X100. Studies have shown that there is a relationship between sperm motility and fertility, but there is no generally accepted standard of how much of a change in motility should be considered adverse. The specific protocol should be considered, as well as information from the other sperm parameters and histopathology. Since sperm motility is dependent on the testis and the epididymis, histological changes in spermatogenesis will likely lead to changes in sperm motility. However, the absence of a histological lesion does not necessarily mean that a change in motility should be discounted. A dose-response trend and a statistically significant change in sperm motility would generally be interpreted as indicating a potential effect on fertility in humans.

169. Sperm morphology is related to fertility, but there is no generally accepted standard of how much of a change in morphology should be considered adverse. Information on sperm motility and count, as well as histopathology should be considered in the interpretation of sperm morphology. Histological lesions of sufficient magnitude can impact sperm morphology. However, normal sperm morphology is dependent on numerous factors including the correct assembly of protamines, so changes in sperm morphology should not be discounted in the absence of histological lesions.

170. Sperm count data are usually expressed in two ways. Information regarding the number of sperm available for ejaculation is provided by the number of sperm per cauda and the number of homogenisation-resistant spermatids per testis. Information regarding the efficiency of the tissue is provided by the number

of sperm per milligram cauda and the number of homogenisation-resistant spermatids per milligram testis. Studies have demonstrated a strong relationship between sperm counts and fertility in all species that have been examined. Again, information on the other sperm parameters and histopatholgy should be considered in the overall interpretation. Testicular lesions of sufficient magnitude will be reflected in the sperm counts, but changes in sperm counts should not be discounted in the absence of histological lesions. Similarly, a reduction in sperm count may not result in reduced fertility, particularly in rodent studies. This is due to the fact that rats have a tremendous excess of spermatozoa in their ejaculates, and as such sperm counts have to be reduced by as much as 90% to affect fertility. It is important to note that human males do not have the large excess of spermatozoa typical of rodents, and therefore a small decrease in sperm count may affect fertility. In general, a statistically significant change in sperm count would be indicative of a potential effect on human fertility.

Female Reproductive Organs

171. The information on the weights and histopathology of the female reproductive organs as well as reproductive performance should be evaluated together, with consideration of the specific test protocol. Uterine weight fluctuates 3- to 4-fold during the oestrus cycle, peaking at pro-oestrus when it is filled with fluid in response to increased oestrogen secretion. Thus, compounds that inhibit steroidogenesis and cyclicity can cause the uterus to become atrophic; conversely compounds that are estrogenic can cause the uterus to become quite large. Uterine weights taken from cycling animals have a high variance, and only large compound-related effects will be demonstrated unless the data are interpreted relative to the animal's oestrus cycle stage. Similarly, the histologic appearance of the uterus varies with the stage of the oestrus cycle and pregnancy. The uterine endometrium, which is sensitive to estrogens and progestogens, will show hypertrophy and hyperplasia in response to these kinds of compounds, and will show hypoplasia and atrophy in response to compounds that inhibit steroidogenesis. Effects induced during development can result in delayed puberty and persistence of infantile genitalia. Information on the number of implantation sites should be interpreted along with the information on live pups and resorptions as described in paragraph 64.

172. A limited amount of information regarding ovarian histopathology can be obtained from female rodents that have been treated according to Test Guidelines 421 or 422 as the dams are sacrificed on lactation day 4, and therefore are not actively cycling. In the rat, ovarian weight does not fluctuate during the oestrus cycle, and any changes should be considered adverse. The function of the ovary shifts during the oestrus cycle so histopathology can reveal a variety of effects including oocyte and follicle depletion, persistent polycystic ovaries, inhibition of corpus luteum formation and luteal cyst development. It is important to note that not all histological alterations will affect ovarian weight, so the lack of an effect on ovarian weight does not preclude the need for histological evaluation. Similarly, the nature and the magnitude of the histological lesion will determine whether there is a concomitant effect on reproductive performance. In general, any dose related significant histological finding would be considered to indicate a potential effect in humans.

173. Many studies have shown a relationship between follicle number and fertility. There is still debate about the exact relationship between the number of follicles and the onset of menopause. There is no generally accepted standard of how much of a change in follicle counts should be considered adverse. The information on follicle counts should be examined in conjunction with the histological information and reproductive performance. Often a change in follicle number will be apparent prior to a change in organ weight or histopathology. The magnitude of the reduction in the number of follicles will determine whether there is an effect on reproductive performance. A decrease in follicle counts could indicate either direct oocyte toxicity, or an effect on the granulosa or thecal cells that alters the paracrine control of oocytes. A dose-response trend and a statistically significant change in follicle number would indicate a potential effect in humans.

174. Vaginal weight changes should parallel those seen in the uterus during the oestrus cycle, although the magnitude of the change will be smaller. The vaginal smear data collected in the two-generation reproductive toxicity study can provide information on cycle length, persistence of oestrus, persistence of di-oestrus, incidence of pseudo-pregnancy. Cycle length is determined by selecting a stage in the cycle and counting the number of days until that stage reoccurs. For statistical analysis, comparisons can be made among animals exhibiting normal 4-5 day cycles and abnormal cycles. Alternatively, altered cyclicity may be reflected by the percentage of time spent in one stage of the cycle, and comparisons can be made based on the number of days spent in di-oestrus or oestrus. However, this kind of analysis must take into account individual cycling patterns as it is possible that more than one stage may be affected; analyses that focus on just one stage may mask such effects. An effect on the oestrus cycle can also impact reproductive performance, but this will depend on the nature and magnitude of the oestrus cycle effect. In general, a statistically significant change in the length of the cycle or prolonged oestrus or di-oestrus would be considered adverse. Effects induced during development can lead to agenesis, hypoplasia, and dysgenesis of the vagina. Hypoplasia of the vagina can be concomitant with hyperplasia of the external genitalia, and alter AGD. The opening of the vaginal orifice at puberty is a simple and useful developmental marker.

175. Alterations in pituitary weight in females are indications of an adverse effect. Increased pituitary weight due to exposure to estrogenic compounds often precedes tumour formation, and may be accompanied by hyper-prolactinemia and a persistent vaginal cornified smear pattern. Decreased pituitary weight due to decreased estrogenic stimulation is less common. The discussion on pituitary weight and histopathology in males is also applicable to females.

Reproductive Performance

176. The data on reproductive performance has to be interpreted with reference to the specific protocol that was used. The data will not necessarily be similar for studies in which there is more than one mating per generation and more than one breeding generation. In addition, in the TG 421, 422, 415 and 416 both sexes are treated, and therefore it is usually not possible to determine whether an effect on some aspect of reproductive performance is due to the female or the male. As noted in the preceding sections, it is also important to evaluate the various indices of reproductive performance along with other available information including histopathology, sperm parameters, follicle numbers, and oestrus cycling.

177. The mating index provides information on the integrated function of the neuroendocrine-gonadal axis. It can be affected by many factors including libido, hormonal imbalance and oestrus cycle disruptions. In most protocols, it will not be possible to determine the cause of an effect on mating since the target may originate in the nervous system or have behavioural or hormonal causes.

178. The fertility index provides information on the ability of the male and female to achieve a pregnancy. The interpretation of the fertility data should consider the protocol that was used and in particular the duration of treatment of the males prior to mating. In protocols in which the males are treated for less than the duration of the spermatogenic cycle (*i.e.*, TG 421 and 422), it is unlikely that a reproductive effect would be manifested in the fertility index. When there is an effect on the fertility index, it may be difficult to determine the affected sex in studies where both sexes are dosed. The fertility information should be considered in conjunction with the available information on histopathology, sperm parameters, follicle numbers and oestrus cycling. As noted above, the male rodent has a large excess of spermatazoa and therefore it takes a large reduction in sperm number to be reflected as a change in the fertility index. Thus, the fertility index alone can be a rather insensitive endpoint.

179. The gestation index should be treated cautiously. It is not a particularly sensitive endpoint since all litters, regardless of size, are treated equally. Therefore, the gestation index will not provide information on an increased incidence of resorptions.

180. The length of gestation should be evaluated in conjunction with information on the birth weights and pup survival. A significant reduction in gestation length may also result in reduced birth weights and pup survival. A significant increase in gestation length can also result in a difficult delivery (dystocia) that may impact the health of the dam and/or pups. In addition, the birth weights of the pups may be higher than normal.

181. Litter size is an important indicator of overall reproductive performance. A decrease in litter size can result from several factors including the number of oocytes ovulated, failure to fertilise, an increase in pre- or post-implantation loss, and reduced sperm counts, motility or abnormal morphology. Thus, the litter size information should also be evaluated in conjunction with the other available information on reproductive endpoints.

182. Changes in the sex ratio index can result from a number of factors including selective impairment on the production of male or female embryos (*i.e.*, on X or Y sperm), sex-linked germ cell abnormalities, selective loss of male or female foetuses, and hormonal changes. Thus, the sex ratio information should be evaluated in conjunction with information on embryonic and foetal loss, as well as information from genetic toxicity studies.

183. The survival index reflects the ability of the pups to survive and is an important endpoint. The pups will begin to consume food and water on lactation day 14 and therefore may also directly consume the test substance. Reduced survival can result from a number of factors including developmental effects in the pups from either prenatal or lactational exposures, maternal neglect, or insufficient milk. Interpretation of the information on survival should also consider the information on pre- and post-implantation loss, number of stillborn pups, total number of dead pups, the number of affected pups per litter, and the pattern of mortality.

V. TESTING STRATEGIES

Regulatory issues

184. The United Nations (UN) Globally Harmonized System of Classification and Labelling of Chemicals (GHS) (UN 2003) provides a classification system for reproductive toxicity where chemical substances are allocated to one of two categories. Effects on reproductive ability or capacity, and on development, are considered as separate issues. In addition, effects on lactation are allocated to a separate hazard category. The two categories are:

- Category 1: Known or presumed human reproductive toxicant.
- Category 1A: Known to have produced an adverse effect on reproductive ability or capacity or on development in human. (*The placing of the substance is largely based on evidence from humans)*
- Category 1B: Presumed to produce an adverse effect on reproductive ability or capacity or on development in humans. (*The placing of a substance in this category is largely based on evidence from experimental animals)*
- Category 2: Suspected human reproductive or developmental toxicant (*This category includes substances for which there is some evidence from humans or experimental animals, possibly supplemented with other information of an adverse effect on reproductive ability or capacity, or on development, in the absence of other toxic effects, or if occurring together with other toxic effects*

the adverse effect on reproduction is considered not to be a secondary non-specific consequence of the other toxic effects, and where the evidence is not sufficiently convincing to place the substance in category 1)

185. The hazard-based classification system used by all EU Member Countries is defined in Annex VI to Commission Directive 67/548/EEC (Anon 1967), last updated by Commission Directive 2001/59/EC (Anon 2001). In the EU system reproductive toxicity is divided into effects on fertility and developmental toxicity. In addition, substances which may interfere with lactation or be present in the breast milk in significant amounts can be classified as hazardous to breast fed babies. There are three categories:

- Category 1: This category includes: (i), substances known to impair fertility in humans; and (ii), substances known to cause developmental toxicity in humans. They are placed in this category if there is sufficient evidence to establish a causal relationship between human exposure to the substance and impaired fertility and if there is sufficient evidence to establish a causal relationship between human exposure to the substance and subsequent developmental toxic effects in the progeny.

- Category 2: This category includes: (i) substances which should be regarded as if they impair fertility in humans, and (ii), substances which should be regarded as if they cause developmental toxicity in humans. They are assigned to this category if there is sufficient evidence to provide a strong presumption that human exposure to the substance may result in (a), impaired fertility on the basis of: clear evidence in animal studies of impaired fertility in the absence of toxic effects, or, evidence of impaired fertility occurring at around the same dose levels as other toxic effects but which is not a secondary non-specific consequence of the other toxic effects; other relevant information and (b), developmental toxicity, generally on the basis of: clear results in appropriate animal studies where effects have been observed in the absence of signs of marked maternal toxicity, or at around the same dose levels as other toxic effects but which are not a secondary non-specific consequence of the other toxic effects; other relevant information.

- Category 3: This category includes: (i), substances which cause concern for human fertility; and (ii), substances which cause concern for humans owing to possible developmental toxic effects. They are assigned to this category generally on the basis of results in appropriate animal studies which provide sufficient evidence to cause a strong suspicion of (a), impaired fertility in the absence of toxic effects, or evidence of impaired fertility occurring at around the same dose levels as other toxic effects, but which is not a secondary non-specific consequence of the other toxic effects, but where the evidence is insufficient to place the substance in category 2, and/or other relevant information and (b), developmental toxicity in the absence of marked maternal toxicity, or at around the same dose levels as other toxic effects but which are not a secondary non-specific consequence of the opther toxic effects, but where the evidence is insufficient to place the substance in category and/or other relevant information.

186. Essentially all European countries which operate a classification system for reproductive toxicity are currently using or are likely to adopt the EU system in the near future. Guidance for hazard classification can be found in the EU Technical Guidance Document on Risk Assessment (EU 2003b).

Conceptual Framework of Testing and Assessment

187. Current regulatory data requirements that are in place to assess hazards to the reproductive system vary considerably depending on the chemical category and its estimated or expected exposure. Data requirements are most stringent for chemicals with deliberate exposure and for those that are designed to

be bioactive. These include pharmaceuticals, food additives and pesticides/biocides. For other chemical categories the level of data requirements increases progressively with increasing production volumes or when other non-reproductive studies reveal indications of possible reproductive effects.

188. Regulatory hazard / risk assessment deals with the question of 'what are acceptable risks of chemical exposure' and has been described as a process that is made up of two distinct concepts: 'science' and 'art'. The science is considered the part that deals with the characterisation of adverse effects by collecting and producing sufficient information. The art has been defined as the evaluation of acquired information, the decisions for more, or different, information and the overall assessment of the hazard / risk.

189. Focussing on the science, currently a number of the key phases and events of the reproductive cycle are covered in standardised animal tests used for regulatory purposes. These tests, however, do not provide specific information on the different aspects of the reproductive cycle. Instead, they are largely designed as apical tests where, for instance, observations of "litter size" stand for and cover: fertilisation, implantation and prenatal development. Currently available standardised tests are:

OECD 407 (draft) :	Enhanced 28-Day Repeated Dose Oral Toxicity
OECD TG 414	Developmental Toxicity Study
OECD TG 415 :	One-Generation Reproduction Toxicity Study
OECD TG 416 :	Two-Generation Reproduction Toxicity Study
OECD TG 421 :	Reproduction/Developmental Toxicity Screening Study
OECD TG 422 :	Combined Repeated Dose Toxicity Study with the Reproduction / Developmental Toxicity Screening Study
OECD TG 426 (draft) :	Development Neurotoxicity Study
OECD Draft :	Uterotrophic Bioassay for (anti) estrogenic effects
OECD Draft :	Hershberger Bioassay for (anti) androgenic effects
ICH S5A :	Detection of Toxicity to Reproduction for Medicinal Products (Segment I, II and III studies) (ICH 1992)
ICH S5B(M) :	Reproductive Toxicology : Male Fertility Studies (ICH 2000)

Directive 67/548/EEC B.31 (= OECD TG 414) Teratogenicity test – rodent and non-rodent, additional

Annex VIII species for chemicals produced more than 1000t/a (Level 2 testing)

Testing, Study for peri-natal and post natal effects for level 2 testing

B.34 (=OECD TG 415) : one-generation reproduction toxicity test

B.35 (= OECD TG 416) : two-generation reproduction toxicity test

Extended B.35

Directive 67/548/EEC	B.22 (= OECD TG 478) : rodent dominant lethal test
(Annex VII A, B, C)	B.23 (= OECD TG 483) : mammalian spermatogonial chromosome aberration test

190. The 'science' is constantly progressing by the development of new animal tests and, more recently, by alternative, non-animal tests (*e.g.,* the embryo stem cell test). The latter tests are not yet considered for regulatory assessment in the area of reproductive toxicity. Currently, regulatory assessments are usually made once the required overall set of data has been obtained. However, in addition, rather than to continue the search for new alternatives to the animal tests (which would imply that the currently used animal studies together would provide sufficient information for regulatory assessment), one should focus more on the "art" of assessment. This would mean a shift from aiming at bringing together a predetermined package of information (*in vivo* and increasingly *in vitro* tests) to the incorporation of harmonised assessment/evaluation steps to evaluate acquired data each time new information is obtained and to make informed (harmonised) decisions on the need for additional information based on these interim assessments.

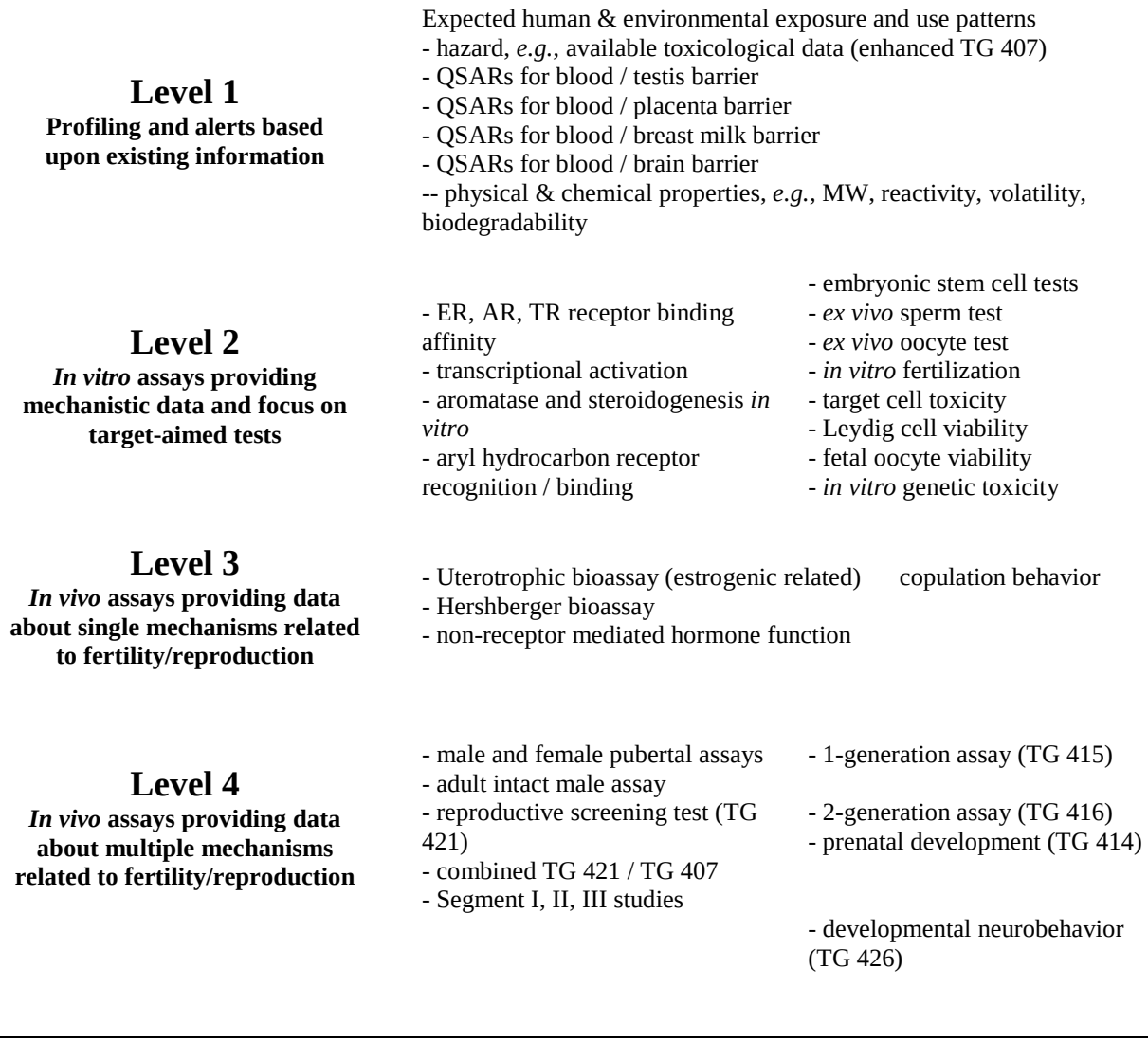
191. Ways to achieve the formal (regulatory) integration of the "art" in the process of hazard/risk assessment (rather than being the last step in the process) include the development of a conceptual framework that incorporates the evaluation of (more or less) each test/study immediately following the test/study and provides the rationale for the selection of a next test/study or for waiving the need for a particular test/study. Such a framework cannot be developed overnight and first experience needs to be gained from the consequences of the various interim decisions for more, or less, studies based on given study results.

192. Furthermore, the framework for the assessment of reproductive effects should not be a sequential testing scheme that has to be rigidly followed. These more rigid schemes only work for well-defined, simple endpoints such as skin or eve irritation/corrosion. Moreover, there is a tendency to "skip steps" in rigid sequential testing strategies to gain time as the animal test at the bottom of the strategy is often considered as anyway covering all that the alternative tests above it would provide. Instead, the framework is an intelligent decision-making process where each component provides a unique piece of information. After each test there is a need to decide what the next step should be. This could be a tailored animal test, designed to study a specific aspect of the reproductive system, or an in vitro approach, or a QSAR.

193. Although these interim assessments may be perceived as the opposite of harmonised testing and assessment, in fact they arise not: all "tools" in the toolbox (the collection of available tests) would be internationally harmonised and well defined in terms of the specific information they generate. The selection of the next test/study ("tool") would be dependent of: (i) the outcome of the previous one and (ii) the level of detail required for the chemical (category) under study. The latter would remain to be internationally harmonised and the former would indeed be subject to expert judgement but the choice should usually be obvious.

194. The proposed Conceptual Framework will be designed as a framework based on increasing levels of information provided (fig. 1). Again, the framework is based on the principle of tailor-made assessments. In this case the most logical start would be at Level I. The assessment of the outcome of this box would determine where to continue in Level 2. Within this level the assessment of one or more tests would determine the next step: (i) continue with Level 2 tests, stop testing or move on to another level.

Figure 1 : Conceptual Framework for the Regulatory Hazard Assessment of Chemicals With respect to Mammalian Reproduction, based on increasing levels of information provided



Sequence of testing

195. The present strategy deals with evidence for reproductive toxic effects, *i.e.*, hazard assessment. Potency is not an issue in the present sequence of testing, but should be considered at a later stage. Further studies may be needed for risk assessment, especially studies using appropriate routes of administration, *i.e.*,

inhalation or dermal.

196. Evidence for potential reproductive toxicity of a chemical can come from epidemiological studies or experimental animal studies. For new substances, human data will be scarce or lacking. For chemicals already in use, human data will be of importance in assessing hazard for reproductive toxicity. Data originating from *in vivo* experiments using laboratory animals, *in vitro* studies using cell-cultures, organ-cultures etc., or structure-activity-relationship (SAR) considerations should all be considered. Data derived from the use of new technologies like genomics or proteomics will maybe in the future provide useful information for the evaluation of reproductive toxicity potential of chemicals (Mirkes *et al.,* 2003).

197. The objective is to give guidance on a stepwise approach to hazard identification with regard to reproductive toxicity, so that sufficient data may be obtained for appropriate classification and labelling. The strategy seeks to ensure that these data are obtained in the most efficient and humane manner so that animal usage and costs are minimised.

198. In order to facilitate the systematic testing of substances, taking into account any existing data available, it is necessary to adopt a strategic approach towards testing. The following sets out a logical series of steps that may provide guidance for the testing of substances for reproductive toxicity.

199. The route of exposure in experimental animal studies should when ever possible mimic the expected exposure of human beings. For workplace chemicals, the relevant routes of exposure are inhalation and dermal exposure.

200. The current knowledge concerning the structure-activity-relationship (SAR) in reproductive toxicity is limited, and the absence of structural indications is not a valid reason for not testing a substance. However, when there is some structural relationship between the chemical and other chemicals shown to cause toxic effect on reproduction, such information should be included.

201. When a certain study is suggested in the present proposal, it is implicit that information from earlier studies is considered. Since most toxicological data are derived from experimental work, this section primarily focuses upon laboratory animal experiments. When recommending experiments using laboratory animals, it is mandatory to consider the welfare of the experimental animals. Thus, it must be clear whether the use of animals is necessary to answer the particular question and if so minimise pain and distress involved (OECD 2000).

202. Data from a test conducted using the TG 421 may identify a substance as being toxic to reproduction (*i.e.*, the test gives a clear positive result). However, this approach does not cover all relevant parameters (*e.g.*, mating behaviour, functional effects, postnatal development) and is therefore not a sufficient basis for a comprehensive risk assessment. In addition, because of the study design (*e.g.*, relatively small numbers of animals per dose level or relatively short study duration), the method will not provide evidence for claims of no effects. Similar criteria apply when evaluating data obtained using the method described in the TG 422. Therefore, it is strongly recommended that where exposure considerations or structural alerts suggest that the comprehensive risk assessment will ultimately require a more in-depth investigation of reproductive endpoints, then the TG 421 and 422 can be omitted.

203. The minimum data requirement in order to have an assessment of the various endpoints covered by the term reproductive toxicity is two-generation study (TG 416) and developmental toxicity studies in two species (TG 414), the need for the second species being dependent on the outcome of the study in the first species. For example, where there is a clear positive result in the first species, a second species may not be

required for classification, whereas a negative or equivocal result in the first species will require a further test in a second species. The need for a developmental neurotoxicity investigation must also be considered.

204. The testing strategy operates with two main sets of data, namely those deriving from humans and those from experimental studies. The levels considered concerning human data are: Sufficient data or limited data. The levels considered for experimental data are: A full toxicological data set, some data, and virtually no data.

205. The tests suggested in testing strategy are those for which an accepted guideline exists, and other tests evaluated as scientifically adequate (see chapters 2-4).

Human data

206. It is often very difficult unequivocally to associate human exposure to a specific substance with adverse effects on reproduction unless the adverse effect is a rare birth defect and the exposure is very well characterised. This is partly because it is difficult to identify a cause-effect relationship for reproductive effects which have a high "natural" incidence (*e.g.*, spontaneous abortion) and which may predispose to unreliable results due to recall bias. Also, some effects can be very subjective (*e.g.*, effects on libido). As with much human data, there may be mixed exposures and/or lifestyle related confounding factors. Nevertheless, well designed and well reported epidemiological studies in which both reproductive and relevant non- reproductive effects are described will contribute to the weight of evidence for whether or not the substance is toxic to reproduction.

207. If there is clear evidence of reproductive toxic effects in humans, the chemical should be considered toxic to reproduction and no further studies are needed for hazard identification and classification, however for risk characterisation further studies may be needed.

208. If there is some evidence of reproductive toxic effects in humans, experimental studies to investigate the causal relationship between exposure to a chemical and the effect are recommended.

209. If the data indicate that the effects are caused in a mixed exposure situation, the experimental studies should, ideally, also include studies into combined effects of the chemicals involved. Information regarding SAR and general toxicity should be included in planning such studies.

210. When there is some evidence of developmental toxicity in humans, a prenatal developmental toxicity study or developmental neurotoxicity study is needed, depending on the effect seen in humans. If the effects observed in humans indicate endocrine disruption or effects on development of sex organs, a two-generation reproductive toxicity study could provide further relevant information.

211. When there is some evidence of impaired fertility or indications of reproductive toxicity in humans, such as increases in sperm anomalies and menstrual disturbances, a one- or multi-generation study is needed.

Toxicological data in experimental animals

212. In the following, suggestions will be given for possible further testing for hazard identification based upon the data from OECD reproductive toxicity tests or other information including data from other toxicity tests. A variety of toxicological data for chemicals may be available including studies on acute, sub-chronic and chronic toxicity, genotoxicity studies, studies on toxicokinetics and studies on reproductive toxicity tests *i.e.*, one or two-generation studies and a prenatal developmental toxicity study in two species of laboratory animals. In addition an extensive toxicological data set sometimes includes data on neurotoxicity. For some

chemicals, an extensive toxicological data set is available for the authorities. However, for the prevailing chemical substances including workplace chemicals, such a comprehensive data set will only be available in few cases.

Reproductive Toxicity Tests

Results from One or Two-Generation Studies

213. When there is clear evidence of effects on fertility in animals the chemical should be considered toxic to reproduction and no further studies are needed for hazard assessment. However, further studies on toxicodynamics (and maybe toxicokinetics) can provide information for evaluation of cause-effects relationship and the relevance for humans.

214. Equivocal results could suggest the necessity of further mechanistic studies or maybe a repetition of the study including possible design adjustments. Special attention should be given to identify whether a possible effect is mediated via the male or the female or caused by general toxicity. In some cases specific studies *e.g.*, a continuous breeding study could be relevant.

215. If there is no evidence of effects on male or female fertility, sperm quality or oestrus cyclicity further studies of fertility effects are normally not needed for the hazard assessment of the chemical. However, if there is some, but limited evidence of fertility effects in humans further studies are needed *e.g*., in another laboratory animal species or including of additional endpoints.

216. Clear evidence of developmental effects *e.g.,* decreased litter size, stillbirth, anomalies, postnatal death or functional anomalies are sufficient for considering the chemical as a developmental toxicant. However, further studies on toxicodynamics (and maybe toxicokinetics) can provide information for evaluation of cause-effects relationship and the relevance for humans.

217. Equivocal evidence of developmental toxicity should be followed by a prenatal developmental toxicity study or a developmental neurotoxicity study.

218. The studies have limitations in detecting developmental toxicity with a sufficient sensitivity. Thus, the lack of evidence of any developmental toxicity is not sufficient to exclude the possibility of developmental toxicity. (For example, if there is evidence of neurotoxic effects in adult animals, a developmental neurotoxicity study should be considered).

Results from Prenatal Developmental Toxicity Studies

219. When there is clear evidence of developmental toxicity in animals the chemical should be considered toxic to reproduction and no further studies are needed for hazard assessment. However, further studies on toxicodynamics (and maybe toxicokinetics) can provide information for a better evaluation of cause-effects relationship and the relevance for humans.

220. Equivocal results could suggest the necessity of further mechanistic studies or a repetition of the study including possible design adjustments or a study using another species. If developmental toxicity has only been observed at doses causing marked toxic effects on the mother, studies using lower/less spaced dose levels and looking for specific endpoints are recommended. In some cases specific studies of developmental neurotoxicity could be relevant, *e.g.,* if there are anomalies in the brain or if other information indicates a neurotoxic potential of the chemical in question.

221. If there is no evidence of developmental toxicity of the chemical further studies are normally not

needed for the hazard assessment. However, if there is some, but limited evidence of developmental toxicity effects in humans, further studies are needed including studies in another laboratory animal species. In some cases specific studies of developmental neurotoxicity could be relevant, *e.g.*, if other information indicates a neurotoxic potential of the chemical in question.

Results from Screening Test (TG 421 and 422)

222. Clear evidence of reproductive toxicity is sufficient for considering the chemical as toxic to reproduction. However, these studies have a limited sensitivity and consequently either one- or two-generation study or prenatal developmental toxicity study should be considered to establish the dose-response curve.

223. Equivocal and negative studies are insufficient to exclude a clear or possible effect respectively. For further testing, the suggestions given in the sections on one- or two-generation studies, prenatal developmental toxicity studies and developmental neurotoxicity study are valid.

Results from Developmental Neurotoxicity Study (Draft TG 426)

224. When there is clear evidence of developmental neurotoxicity in animals the chemical should be considered toxic to reproduction and no further studies are needed for hazard assessment. However, further studies on toxicodynamics (and maybe toxicokinetics) can provide information for evaluation of cause-effects relationship and the relevance for humans. Substances giving rise to less clear manifestations of developmental neurotoxicity (such as retardation of postnatal reflex ontogeny) should be considered for further investigations so that the potential for induction of serious functional effects can be evaluated

225. Equivocal results (*i.e.*, the results of the study are such that no firm conclusion can be drawn from the study about the developmental neurotoxicity of the substance) could be investigated further by additional testing, normally in the same species. It will usually be necessary to vary the study design in order to obtain useful results, for example larger group sizes and/or special developmental neurotoxicity investigations could be considered.

226. If there is no evidence of developmental neurotoxicity of the chemical further studies are normally not needed for the hazard assessment of developmental neurotoxicity. However, a second study in another species and/or examination of developmental neurotoxicity end points not covered yet should be considered when there are potential widespread exposure of women of childbearing age and indications of developmental neurotoxicity in humans.

Other Data in Experimental Animals

Results from Repeated Dose Toxicity Studies

227. Data from repeated dose toxicity studies in which there are marked adverse effects on the reproductive organs can be used to identify a substance as being toxic to reproduction. Further studies are necessary to investigate the consequences of findings of testicular or ovarian effects, or gain some knowledge concerning reproductive effects on selected chemical substances. Data from repeated dose toxicity studies are not sufficient to exclude a possible effect on reproduction, since fertility may be impaired in the absence of any histological damage to the gonads.

228. If repeated dose toxicity studies have indicated testicular effects, a one- or two-generation study focusing on the development of the male sexual organs and fertility could provide useful further information. An additional group dosing males only could be included in the one- or two-generation study. If SAR

indicates an effect on fertility, a reduced generation study including sperm analysis might be considered as sufficient for the hazard evaluation. In some cases, clear toxic effects on the gonads may be detected making it unnecessary to perform fertility testing.

229. If repeated dose toxicity studies have indicated ovarian effects, a one- or two-generation study focusing on parameters on female reproductive toxicity including oestrus cyclicity is recommended. The inclusion of a group only dosing females in the one- or two-generation study would provide additional information. Information from SAR can further identify relevant parameters.

230. If there is evidence of neurotoxic effect in adult animals, a developmental neurotoxicity study should be considered. Developmental neurotoxicity testing should be conducted to further characterise neurological effects observed in other studies, and should be considered if the substance has been shown to cause neuropathology or neurotoxicity in adults, to be a hormonally-active material *in vivo* (*e.g.*, pituitary, thyroid, sex hormones), or to cause other types of toxicity, suggestive of nervous system involvement at a developmental stage

Other Data (In Vitro Tests, Kinetics/Dynamics)

231. If an *in vitro* teratogen study has indicated a developmental effect, and SAR and kinetic data support the relevance of this effect, further studies may not be needed for an initial hazard assessment. However, a prenatal developmental toxicity study would provide useful information.

232. If the chemical may cause e.g. accumulation in the foetus, competitive metabolism, lowered blood pressure/disturbed circulation, lowered oxygen tension and formation of carboxyhemoglobin, genotoxic effects (cancer) or is a vitamin or hormone analogue, studies on developmental toxicity are indicated.

233. If for example the chemical is a hormone analogue or has shown hormonal effects in specific tests (*e.g.,* Uterotrophic test, Hershberger test, receptor binding) or accumulates in sex organs, a one- or two-generation study is needed.

VI. DATA GAPS

234. Test Guidelines have been designed as screening tests for hazard identification and dose response assessment, and as such, they provide a valuable tool for risk assessment. The studies described in this document facilitate the broad screening of chemicals for potential developmental and reproductive toxicity. However, the sensitivity of the studies in the detection of effects is inherently limited by the protocols and how they are utilized. Some of these limitations, described below, may be important to consider when interpreting study results.

Endpoints

235. The screening level approach used in guideline reproduction and developmental toxicity testing does not incorporate chemical-specific information, such as mode of action or known target organ specificity, into the design of the studies or the endpoints assessed.

236. An understanding of the pharmacokinetic and pharmacodynamic profile of a test substance in the developing system and of the complexities of direct and indirect developmental exposures during pregnancy, lactation, and to neonates by various routes of exposure is critical to study design, dose

selection, and the interpretation and extrapolation of reproductive toxicity data. Current Test Guidelines do not address the collection or use of these data.

237. Current developmental and reproductive toxicity Test Guidelines provide only limited focus on alterations in the ontogeny of organ system function (US EPA, 2002). While reproductive system and nervous system functions are extensively assessed in the multi-generation reproduction study and the developmental neurotoxicity study, respectively, other functional endpoints are not included. For example, the development of normal functional capabilities of the cardiac, urinary, respiratory, immune, and endocrine systems is not assessed.

238. No specific standardised Test Guidelines exist for second tier testing on chemicals that are suspected to enhance carcinogenic response following perinatal exposure.

239. The behavioural functions assessed cover many important aspects of the nervous system, however, some functions of relevance for *e.g.*, endocrine disruption such as social interaction and mating behaviour are not included in guidelines at present. Other potentially relevant postnatal endpoints as *e.g.*, kidney function, liver function and immunotoxicity, are not included in these guidelines.

240. Neurobehavioral and neuropathological assessments included in the developmental neurotoxicity guideline provide a broad screen for apical effects on multiple functional domains. However, primarily as a function of the animal model typically used, the study does not address some aspects of behaviour (*e.g.*, social behaviour), and cannot provide information on higher level cortical functions that have serious implications in humans. Additionally, specialised follow up developmental neurotoxicity testing is not addressed. These would include experiments in which relatively sensitive measures of sensory and/or cognitive function are evaluated in the offspring of animals exposed to chemicals during pregnancy and/or postnatally.

Critical windows of exposure and effect

241. While current guideline studies provide opportunities for exposure across multiple periods of development, the studies are not designed to determine the critical window of exposure. Such information could be useful in targeting a risk assessment to various susceptible population subgroups. With the exception of the developmental neurotoxicity study, which includes a period (approximately 40 days) without treatment prior to neurobehavioral and neuropathological evaluations of offspring at study termination, the issue of longer-term consequences of developmental exposures is not examined. Additionally, none of the current guidelines address the potential for effects of developmental exposures in ageing individuals.

Exposure

242. In Test Guidelines that include reproductive phases, extensive variations in exposure to the test subjects, over various life stages, are typical; exposure of the offspring to the test substance is not characterized. In order to ensure adequate exposure of the offspring during critical stages of development (*e.g.*, when assessing effects on the developing nervous or immune system) and/or to quantify exposure to pups, the use of direct dosing of pre-weaning test subjects may be considered necessary.

VII. REFERENCES

Adams, J. (1986) Methods in behavioral teratology. Handbook of behavioral teratology, Riley EP, Vorhees CV (eds): New York: Plenum Press, 67-97.

Akaike, M., Tanaka, K., Goto, M. and T. Sakaguchi, (1988) Impaired Biel and Radial arm maze learning in rats with methylnitrosurea-induced microcephaly. Neurotoxicol. Teratol. 10: 327-332.

Akaike, M., Ohno, H., Tsutsumi, S. and M. Omusu, (1994) Comparison of four spatial maze learning tests with methylnitrosourea-induced microcephaly rats. Teratology 49: 83-89.

Altman, J. and K. Sudarshan, (1975) Postnatal development of locomotion in the laboratory rat. Anim. Behav. 23: 896-920.

Annau, Z. and V. Cuomo, (1988) Mechanisms of neurotoxicity and their relationship to behavioral changes. Toxicology 49: 219-225.

Anon (1967) European Commission Directive 67/548/EEC (Annex VI).

Anon (2001) European Commission Directive 67/548/EEC, updated by Commission Directive 2001/59/EC.

ATLA (2002) Alternative non-animal methods for chemical testing: current status and future prospects. 10. Reproductive toxicity. (Eds.) Worth, A.P., Balls, M., 30 (Supplement 1), 95-102.

Barlow, S. M. and F. M. Sullivan, (1975) 6. Behavioural teratology. Teratology trends and applications, Berry C. L., Poswillo D. E. (eds): Springer Verlag, 103-120.

Barlow, S. M. and F. M. Sullivan (1982) Reproductive toxicity testing in animals in Reproductive Hazards of Industrial Chemicals. An evaluation of animal and human data. Academic Press INC. (London LTD)).

Blanchard, B. A., Riley, E. P. and J. H. Hannigan, (1987) Deficits on a Spatial Navigation Task Following Prenatal Exposure to Ethanol. Neurotoxicol Teratol. 9: 253-258.

Bogo, V., Hill, T. A. and R. W. Young, (1981) Comparison of accelerod and rotarod sensitivity in detecting ethanol- and acrylamide-induced performance decrements in rats: Review of experimental considerations of rotating rod systems. Neurotoxicol. 2: 765-787.

Buelke-Sam, J. and C. A. Kimmel, (1979) Development and standardization of screening methods for behavioural teratology. Teratology, 20: 17-29.

Bushnell, P. J., Kelly, K. L. and K. M. Crofton, (1994) Effects of toluene inhalation on detection of auditory signals in rats. Neurotox Teratol. 16: 149-160.

Calabrese E. J. (1983) Animal Extrapolation, chapter 14 in Principles of Animal Extrapolation, John Wiley and Sons.

Chapin, R. E., Sloane, R. A., and J. K. Haseman. (1997) The relationships among reproductive endpoints in Swiss mice, using the reproductive assessment by Continous Breeding database. Fundam Appl Toxicol. 38(2):129-42.

Chapin, R. E. and M. W. Conner, (1999) Testicular histology and sperm parameters. In: An evaluation and interpretation of reproductive endpoints for human health risk assessment. Eds: G. Daston and C. Kimmel, International Life Sciences Institute Press.

Clemens, G. R., Petrere, J. A., and K. Oberholtzer. (1994) Historical control database survey (HCDS) Phase II (P2): External and visceral malformations in the Sprague-Dawley rat and the New Zealand White rabbit. Teratology 49:388-9.

Cooper, R. L. and J. M. Goldman. (1999). Vaginal cytology. In: An evaluation and interpretation of reproductive endpoints for human health risk assessment. Eds: G. Daston and C. Kimmel, International Life Sciences Institute Press.

Crofton, K. M. and L. P. Sheets (1989) Evaluation of sensory system function using reflex modification of the startle response. J. Am. Coll. Toxicol. 8: 199-211.

Crofton K. M., Dean K. F., Menache M. G., and R. Janssen. (1990) Trimethyltin effects on auditory function and cochlear morphology. Toxicol. Appl. Pharmacol. 105:123-132.

Crofton K. M., Peele D.B., and M.E. Stanton. (1993) Developmental neurotoxicity following neonatal exposure to 3,3'-iminodipropionitrile in the rat. Neurotoxicol. Teratol. 15:117-129.

Crofton K. M., Janssen R., Prazma J., Pulver S., and Jr. S. Barone. (1994a) The ototoxicity of 3,3'-iminopropionitrile: Functional and morphological evidence of cochlear damage. Hearing Research 80: 129-140.

Crofton K. M., Lassiter T. L., and C. S. Rebert. (1994b) Solvent-induced ototoxicity in rats: An atypical selective mid-frequency hearing deficit. Hearing Research; 80: 25-30.

CSTEE (Scientific Committee on Toxicity, Ecotoxicity and the Environment) (1999) Opinion on Human and Wildlife Health Effects of Endocrine Disrupting Chemicals, with Emphasis on Wildlife and on Ecotoxicity Test Methods. Expressed at the 8th CSTEE plenary meeting, Brussels, 4 March 1999. DG XXIV, Consumer Policy and Consumer Health Protection, European Commission.

Denenberg V.H. (1969) Open-field Behavior in the rat: What does it mean? Ann. N.Y. Acad. Sci. 159(3): 852-859.

DEPA (Danish Environmental Protection Agency) (1995) Male Reproductive Health and Environmental Chemicals with Estrogenic Effects. Ministry of Environment and Energy, Denmark, Miljøprojekt nr. 290, 1995 (ISSN 0105-3094/ISBN 87-7810-345-2) and IPCS-WHO 2001,

ECETOC (1983) Monograph No. 5. Identification and Assessment of the Effects of Chemicals on Reproduction and Development (reproductive Toxicology). Brussels December 1983) ECETOC. Guidance on Evaluation of Reproductive Toxicity Data. Monograph No. 31, Brussels, February 2002 (ISSN-0773-6347-31).

ECETOC (1992) (European Centre for Ecotoxicology and Toxicology of Chemicals (former E. Chemical Industry Ecology and Toxicology Centre) Technical Report No. 47. EC 7th Amendment: “Toxic to Reproduction” Guidance on Classification, Brussels, August 1992 (ISSN-0773-8072-47).

ECETOC (2002) Guidance on Evaluation of Reproductive Toxicity Data. Monograph No. 31, Brussels, February 2002 (ISSN-0773-6347-31).

EHC (1984) Environmental Health Criteria 30. Principles for Evaluating Health Risks to Progeny Associated with Exposure to Chemicals During Pregnancy, WHO Geneva (ISBN 92 4 154090 7).

Elsner J. (1991) The tactile-kinesthetic system of rats as an animal model for minimal brain dysfunction. Arch. Toxicol. 65: 465-473.

Elsner, J., Suter, K., Ulbrich, B., and G. Schreiner. (1986) Testing Strategies in Behavioral Teratology: IV. Review and General Conclusions. Neurobehav. Toxicol. Teratol. 8: 585-590.

Elsner, J., Hodel, B., Suter, K., Oelke, D., Ulbrich, B., Schreiner, G., Cuomo, V., Gagiano, R., Rosengren, L.E., Karlsson, J.E., and K.G. Haglid. (1988) Detection Limits of Different Approaches in Behavioral Teratology, and Correlation of Effects With Neurochemical Parameters. Neurobehav. Toxicol. Teratol. 10: 155-167.

EU (2003a) REACH (Registration, Evaluation and Authorisation of Chemicals). Proposal for a Directive of the European Parliament and of the Council amending Council Directive 67/548/EEC in order to adapt it to Regulation (EC) of the European Parliament and of the Council concerning the registration, evaluation, authorisation and restriction of chemicals.

EU (2003b) Technical Guidance Document in Support of Commission Directive 93/67/EEC on Risk Assessment for New Notified Substances and Commission Regulation (EC) No. 1488/94 on Risk Assessment for Existing Substances. Available at: [http://ecb.jrc.it/]

Francis, E.Z. (1990) Developmental neurotoxicity study and response to public and SAP comment, Washington D.C., US EPA, 1-20.

Francis, E. Z., Kimmel, C. A., and D. C. Rees. (1990) Workshop on the qualitative and quantitative comparability of human and animal developmental neurotoxicity: Summary and implications. Neurotoxicol. Teratol. 12: 285-292.

Gabriel, M., Kubota, Y., Sparenborg, S., Straube, K., and B. A. Vogt. (1991) Effects of cingulate cortical lesions on avoidance learning and training-induced unit activity in rabbits. Exp. Brain. Res. 86: 585-600.

Gerber, G. J., and D. O'Shaughnessy. (1986) Comparison of the behavioral effects of neurotoxic and systemically toxic agents: How discriminatory are behavioral tests of neurotoxicity. Neurobehav. Toxicol. Teratol. 8: 1-12.

Goodlett, C. R., Kelly, S. J., and J. R. West. (1987) Early postnatal alcohol exposure that procedures high blood alcohol levels impairs development of spatial navigation learning. Psychobiology 15(1): 64-74.

Gray, L. E., Ostby, J., Furr, J., Wolf, C. J., Lambright, C., Park, L., Veeramachaneni, D. N., Wilson, V., Price, M., Hotchiss, A., Orlando, E., and L. Guillette. (2001) Effects of environmental antiandrogens on reproductive develeopment in experimental animals. Human Reproduction Update 7: 248-264.

Gårdlund, A. T., Archer, T., Danielsson, K., Danielsson, B., Frederiksson, A., Lindquist, N.-G., Lindström, H., and J. Luthman. (1991) Effects of prenatal exposure to tributyltin and trihexyltin on behavior in rats. Neurotoxicol. Teratol. 13: 99-105.

Hass, U. (1993) Neurobehavioural teratology of industrial chemicals. Effects of prenatal exposure to organic solvents on postnatal development and behaviour - validation and use of a screening test battery in laboratory rats, Ph.D-thesis. Copenhagen:National Institute of Occupational Health, 1-108.

Hass, U., Hansen, E. V., and G Østergaard. (1994a) Experimental studies in laboratory animals. In Hass U et al., (eds) Occcupational reproductive toxicity - Methods and testing strategy for hazard assessment of workplace chemicals. Nordic Council of Ministers and National Institute of Occupational Health, Denmark.

Hass, U., Lund, S. P., and J. Elsner. (1994b) Effects of prenatal exposure to N-methylpyrrolidone on postnatal development and behavior in rats. Neurotoxicol. Teratol. 16: 241-249.

Hass, U., Lund, S. P., Simonsen, L., and A. S. Fries. (1995). Effects of prenatal exposure to xylene on postnatal development and behavior in rats. Neurotoxicol. Teratol. 17: 341-349.

Heindel, J. J. (1999) Oocyte quantitation and ovarian histology. In: An evaluation and interpretation of reproductive endpoints for human health risk assessment. Eds: G. Daston and C. Kimmel, International Life Sciences Institute Press.

Hellwig, J., van Ravenzwaay, B., Mayer, M., and C. Gembardi. (2000) Pre-and postnatal oral toxicity of vinclozolin in Wistar and Long-Evans rats. Reg. Tox. 32: 42-50.

Hughes, E. (1986) Evolution of a behavioural screen, Huntingdon Research Center.

Hughes, E. W., and A. K Palmer. (1986) An assessment of pre-weaning development and behaviour in safety evaluation studies. Department of Reproductive Toxicology HCR (Abstract).

ICH (1993). International Conference on Harmonisation of Technical Requirements for Registration of Pharaceuticals for Human Use, ICH Harmonised Tripatriate Guideline. Detection of toxicity to reproduction for medicinal products, S5A. Recommended for Adoption at Step 4 of the ICH Processon 24 June 1993 by the ICH Steering Committee.

ICH (2000). International Conference on Harmonisation of Technical Requirements for Registration of Pharaceuticals for Human Use, ICH Harmonised Tripatriate Guideline. Maintenance of the ICH guideline on toxicity to male fertility, S5B(M). Recommended for Adoption at Step 4 of the ICH Process on 29 November 1995 and amended on 9 November 2000 by the ICH Steering Committee.

IPCS (International Programme on Chemical Safety) (2001). Principles for Evaluating Health Risks to Reproduction Associated with Exposure to Chemicals, Environmental Health Criteria 225, WHO Geneva, 2001.

Jakobsen B. M., and O. Meyer (1989) Extrapolation from in vivo/in vitro Experiments to Human Beings in Embryo-Foetal Damage and Chemical Substances, Working Party Report. National Food Agency of Denmark, Ministry of Health, publication No. 181, May 1989 (ISBN 87-503-7881-3/ISSN 0903-9783).

Jones, B. J., and D. J. Roberts (1968) The quantitative measurement of motor incoordination in naive mice using an accelerating rotarod. Pharm. Pharmac. 20: 302-304.

Kaplan, M., and S. D. Murphy. (1972) Effect of acrylamide on rotarod performance and sciatic nerve beta-glucuronidase of rats. Toxicol. Appl. Pharmacol. 22: 259-268.

Kelly, S. J., Goodlett, C. R., Hulsether, S. A., and J. R. West. (1988) Impaired spatial navigation in adult female but not adult male rats exposed to alcohol during the brain growth spurt. Behav. Brain Res. 27: 247-257.

Kimmel C. A. (1988) Current status of behavioral teratology: science and regulation. Crit. Rev. Toxicol. 19: 1-10.

Kimmel, C. A., and J. Buelke-Sam. (1985) Collaborative Behavioral Teratology Study: Background and overview. Neurobehav. Toxicol. Teratol. 7: 541-546.

Korenbrot, C. C., Huhtaniemi, I. T., and R. I. Weiner. (1977) Preputial separation as an external sign of pubertal development in the male rat. Biol. Reprod. 17: 298-303.

Ladefoged, O., Hass, U., and L. Simonsen. (1989) Neurophysiological and behavioural effects of combined exposure to 2,5-hexanedione and acetone or ethanol in rats. Pharmacol. Toxicol. 65: 372-375.

Lamb, J.C. (1985) Reproductive toxicity testing: evaluating and developing new testing system. J. Am. Coll. Toxicol. 4(2).

Leblond, C. P. and Y. Clermont (1952). Definition of the stages of the cycle of the seminiferous epithelium in the rat. Ann. N. Y. Acad. Sci. 55: 548-573.

Levine, S., and P. L. Broadhurst, (1963) Genetic and ontogenetic determinants of behaviour. II. Effects of infantile stimulation on adult emotionality and learning in selected strains of rats. J. Comp. Physiol. Psychol. : 56-65.

Levine, S., Haltmeyer, G. C., Karas, G. G., and V. H. Denenberg. (1967) Physiological and behavioural effects of infantile stimulation. Physiol. Behav. 2:32-37.

Lochry, E. A., and E. P. Riley. (1980) Retention of passive avoidance and T-maze escape in rats exposed to alcohol prenatally. Neurobehav. Toxicol. 2.

Lochry, E. A., Hoberman, A. M., and M. S. Christian. (1986) Standardization and application of behavioral teratology screens. Safety evaluation and regulation of chemicals 3, Homburger (ed): Basel, Karger, 49-61.

Lore, R., and H., Avis. (1970) Effects of auditory stimulation and litter size upon subsequent emotional behavior in the rat. Develop. Psychobiol. 2(4):212-215.

LST (1989) Embryo-Foetal Damage and Chemical Substances, Working Party Report. National Food Agency of Denmark, Ministry of Health, publication No. 181, May 1989 (ISBN 87-503-7881-3/ISSN 0903-9783).

MARTA (Middle Atlantic Reproduction and Teratology Association) and MTA (Midwest Teratology Association) (1995) Historical Control Data (1992-1994) for Developmental and Reproductive Toxicity Studies Using the Crl:CD@BR Rat. Charles River Laboratories, Inc.

MARTA (Middle Atlantic Reproduction and Teratology Association) and MTA (Midwest Teratology Association) (1996) Historical Control Data (1992-1994) for Developmental and Reproductive l'oxicity Studies Using the New Zealand White Rabbit. Hazleton Research Products, Inc., Denver, P A.

McIntyre, B. S., Barlow, N. J., and P. M. D. Foster. (2001) Androgen-mediated development in male rat offspring exposed to flutamide in utero: performance and correlation of early postnatal changes in anogenital distance and nipple retention with malformations in androgen-dependent tissue. Toxicol. Sci. 62: 236-249.

Meaney, M. J., Aitken, D. H., Bhatnager S., Van Berkel, C., and R. M. Sapolsky. (1988) Postnatal handling attenuates neuroendocrine, anatomical, and cognitive impairments related to the aged hippocampus. Science 238: 766-768.

Meaney, M. J., Aitken, D. H., Bhatnager, S., sand R. M. Sapolsky. (1991) Postnatal handling attenuates certain neuroendocrine, anatomical, and cognitive dysfunctions associated with aging in female rats. Neurobiol. Aging 12: 31-38.

Meyer, O., Jakobsen, B. M., and E. V. Hansen. (1989) Identification of embryo-foetal toxicity by means of animal studies. In: Andreasen, P. B., Brandt, N. J., Cohr, K.-H., Hansen, E. V., Hass, U., Hauge, M., Jakobsen, B.M., Knudsen, I., Lauritsen, J. G., Melchior, J. C., Meldgaard, L., Meyer, O., Olsen, J. H., Palludan, B., Poulsen, E. Embryofoetal damage and chemical substances. Working party report, Copenhagen: Levnedsmiddelstyrelsen, 1989, 1-127. (Translated version of Danish report from 1986).

Meyer, O and T. Svendsen (2003) Animal Models in Pharmacology and Toxicology. In: Handbook of Laboratory Animal Science, Second Edition, Vol II, Animal Models (Eds. J. Hau and G. L. Van Hoosier, Jr), CRC Press2003.

Miller, D. B., Eckerman, D. A., Krigman, M. R., and L. D. Grant. (1982) Chronic neonatal organotin exposure alters radial-arm maze performance in adult rats. Neurobehav. Toxicol. Teratol. 4:185-190.

Mirkes, P., McClure, M. E., Heindel, J. J., and Sander, M. (2003) Developmental toxicology in the 21st century: Multidisciplinary approaches using model organisms and genomics. Birth Defects Research (Part A): Clinical and Molecular Teratology 67(1):21-34.

Morris, R. (1981) Spatial Localization Does Not Require the Presence of Local Cues. Learning and Motivation 12: 329-260.

Morris, R. (1984) Developments of a water-maze procedure for studying spatial learning in the rat. J. Neuro. Sci. Meth. 11: 47-60.

Morris, R., Garrud, P., Rawlins, J. N. P., and J. O'Keefe. (1982) Place navigation impaired in rats with hippocampal lesions. Nature 297: 681-683.

Mylchrest, E., Sar, M., Catley, R. C., and P. M. D. Foster. (1999) Disruption of androgen-regulated reproductive development by di(n-butyl) phthalate during late gestation in rats is different from flutamide. Toxicol. Appl. Pharmacol. 156: 81-95.

NIOH (AMI) (1994) Occupational reproductive toxicity. Methods and testing strategies for hazard assessment of workplace chemicals. Nordic Council of Ministers and National Institute of Occupational Health, Denmark. Copenhagen 1994 (ISBN: 87-7534-455-6).

Nunez, J. F., Ferré, P., García, E., Escorihuele, R. M., Fernández-Teruel, A., and A. Tobena. (1995) Postnatal handling reduces emotionality ratings and accelerates two-way active avoidance in female rats. Physiol. Behav. 831-835.

OECD (1983) Test Guideline 415. OECD Guideline for Testing of Chemicals. One-generation reproduction toxicity study. Available: [http://www.oecd.org/document/22/0,2340,en_2649_34377_1916054_1_1_1_1,00.html].

OECD (1984) Test Guideline 478. OECD Guideline for Testing of Chemicals. Genetic toxicology: rodent dominant lethal test. Available: [http://www.oecd.org/document/22/0,2340,en_2649_34377_1916054_1_1_1_1,00.html].

OECD (1995). Test Guideline 407. OECD Guideline for Testing of Chemicals. Repeated dose 28-day oral toxicity study in rodents. Available: [http://www.oecd.org/document/22/0,2340,en_2649_34377_1916054_1_1_1_1,00.html].

OECD (1995). Test Guideline 421. OECD Guideline for Testing of Chemicals. Reproduction/developmental toxicity screening test. Available: [http://www.oecd.org/document/22/0,2340,en_2649_34377_1916054_1_1_1_1,00.html].

OECD (1996). Test Guideline 422. OECD Guideline for Testing of Chemicals. Combine repeated dose toxicity study with the reproduction/developmental toxicity screening test. Available: [http://www.oecd.org/document/22/0,2340,en_2649_34377_1916054_1_1_1_1,00.html].

OECD (1998). Test Guideline 408. OECD Guideline for Testing of Chemicals. Repeated dose 90-day oral toxicity study in rodents. Available: [http://www.oecd.org/document/22/0,2340,en_2649_34377_1916054_1_1_1_1,00.html].

OECD (1998). Test Guideline 409. OECD Guideline for Testing of Chemicals. Repeated dose 90-day oral toxicity study in non-rodents. Available: [http://www.oecd.org/document/22/0,2340,en_2649_34377_1916054_1_1_1_1,00.html].

OECD (2000). Guidance Document on the Recognition, Assessment, and Use of Clinical Signs as Humane Endpoints for Experimental Animals Used in Safety Evaluations. OECD Environment, Health and Safety Publications: Series on Testing and Assessment, No. 19. Available: [http://www.oecd.org/document/22/0,2340,en_2649_34377_1916054_1_1_1_1,00.html].

OECD (2001). Test Guideline 414. OECD Guideline for Testing of Chemicals. Prenatal developmental toxicity study. Available: [http://www.oecd.org/document/22/0,2340,en_2649_34377_1916054_1_1_1_1,00.html].

OECD (2001). Test Guideline 416. OECD Guideline for Testing of Chemicals. Two-generation reproduction toxicity study. Available: [http://www.oecd.org/document/22/0,2340,en_2649_34377_1916054_1_1_1_1,00.html].

OECD (2002a) IPCS (International Programme on Chemical Safety). Principles for Evaluating Health Risks to Reproduction Associated with Exposure to Chemicals, Environmental Health Criteria 225, WHO Geneva, 2001.

OECD (2002b). Detailed Review Paper: Appraisal of Test Methods for Sex Hormone Disrupting Chemicals. OECD Environment, Health and Safety Publications: Series on Testing and Assessment, No. 21. Available: [http://www.oecd.org/document/22/0,2340,en_2649_34377_1916054_1_1_1_1,00.html].

OECD. Draft Test Guideline 426. OECD Guideline for Testing of Chemicals. Developmental Neurotoxicity Study. Available: [http://www.oecd.org/document/22/0,2340,en_2649_34377_1916054_1_1_1_1,00.html].

OECD. Draft Test Guideline. OECD Guideline for Testing of Chemicals. The Uterotrophic Bioassay.

OECD. Draft Test Guideline. OECD Guideline for Testing of Chemicals. The Hershberger assay: adult castrated animal model.

OECD. Draft Test Guideline. OECD Guideline for Testing of Chemicals. The Enhanced TG407.

Olton, D. S., Becker, J. T., and G. E. Handelmann. (1979) Hippocampus, space, and memory. Behav. Brain Sci. 2: 313-365.

OTA (Office of Technological Assessment) (1990). Neurotoxicity: Identification and controlling poisons of the nervous system. U.S. Congress, Office of Technological Assessment, OTA-BA-436, U.S. Government Printing Office, Washington, D.C., 7.

Palmer, A.K. (1981) Regulatory requirements for reproductive toxicology: theory and practice. In: Developmental toxicology, (Eds.) Kimmel, C. A. and Buelke-Sam, J., New York, Raven Press, 259-288.

Palmer, A. K. (1986) A simpler multi-generation study. International Congress of Pesticide Chemistry; 1-20 (Abstract).

Pryor, G. T., Dickinson, J., Howd, R. A., and C. S. Rebert. (1983a) Neurobehavioral effects of subchronic exposure of weanling rats to toluene and hexane. Neurobehav. Toxicol. Teratol. 5: 47-52.

Pryor, G. T., Dickinson, J., Howd, R. A., and C. S. Rebert. (1983b) Transient cognitive deficits and high-frequency hearing loss in weanling rats exposed to toluene. Neurobehav. Toxicol. Teratol. 5: 53-57.

Raimondo, S., and M. T. Draghetti. (1990) Influence of length of intra-uterine life on the appearance of developmental markers and neurobehavioural reflexes in rat. 18. Conference of the European Teratology Society 1990; (Abstract).

Reiter, L. W., and R. C. MacPhail. (1982) Factors influencing motor activity measurements in neurotoxicology. In Mitchell, C. T, (ed). Nervous system toxicology.

Rosenthal, R., and K. L. Fode. (1963) The effect of experimenter bias on the performance of the albino rat. Behav. Science 8: 183-189.

Rosenzweig, M. R., and E. L. Bennett. (1969) Effects of differential environment on brain weights and enzyme activities in gerbils, rats, and mice. Develop. Psychobiol. 2: 87-95.

Russel, L. D., Ettlin, R. A., Sinha Hikim, A. P., and E.D. Clegg. (1990) Histological and Histopathological Evaluation of the Testis. Cache River Press.

Saillenfait, A. M., and B. Vannier. (1988) Methodological proposal in behavioural teratogenicity testing: assessment of propoxyphene, chlorpromazine, and vitamin A as positive controls. Teratology 37: 185-199.

Schapiro, S., Salas, M., and B. A. Shaywitz. (1970) Hormonal effects on the ontogeny of swimming ability in the rat: Assessment of central nervous system development. Science 193: 146-151.

Schardein et al., (1985) Evaluation of Human Risk, section VII of Principles of Teratogenesis Applicable to Human Exposure to Drugs and Chemicals in Chemically induced Birth Defects, Marcel Dekker INC.

Schiorring, E. (1979) An open field study of stereotyped locomotor activity in amphetamine treated rats. Psychopharmacol. 66: 281.

Seed, J., Chapin, R. E., Clegg, E. D., Darney, S. P., Dostal, L., Foote, R. H., Hurtt, M. E., Klinefelter, G. R., Makris, S., Schrader, S., Seyler, D., Sprando, R., Treinen, K. A., Veeramachaneni, R. and L. D. Wise. (1996). Methods for assessing sperm motility, morphology, and counts in the rat, rabbit, and dog: A consensus report. Reprod. Toxicol. 10:237-244.

Selevan, S. G., Kimmel, C. A., and P. Mendola. (2000) Identifying critical windows of exposure for children's health. Environmental Health Perspectives 108(3):451-455.

Spark, C. and Dawson, A. B. (1928). The order and time of appearance of centers of ossification in the fore and hind limbs of the albino rat, with special reference to the possible influence of the sex factor. American Journal of Anatomy 41, 411-445.

Stanton, M. E. and J. H. Freeman. (1994) Eyeblink conditioning in the infant rat: An animal model of learning in developmental neurotoxicology. Environ. Health Perspect. 102: 131-139.

Staples, R. E. (1974) Detection of visceral alterations in mammalian fetuses. Teratology 9(3):A37- A38.

Tamura, R. N., and J. Buelke-Sam. (1992) The use of repeated measures analyses in developmental toxicology studies. Neurotoxicol. Teratol. 14: 205-210.

Toppari, J., Larsen, J., Christiansen, P., Giwercman, A., Grandjean, P., Guillette Jr. L. J., Jegou, B., Jensen, T. K., Jouannet, P., Keiding, N., Leffers, H., McLachlan, J. A., Meyer, O., Muller, J., Rajpert-De Meyts, E., Scheike, T., Sharpe, R., Sumpter, J., and N. E. Skakkebaek. (1996) Male Reproductive Health and Environmental Xenoestrogens. Environmental Health Perspectives 104(Suppl 4):741-803.

United Nations (UN) (2003) Globally Harmonized System of Classification and Labelling of Chemicals (GHS). ST/SG/AC.10/30, UN New York and Geneva. Available:
[http://www.unece.org/trans/danger/publi/ghs/officialtext.html]

US EPA. (1990) Developmental neurotoxicity study and response to public and SAP comment. Washington D.C.

US EPA (1991) Guidelines for developmental toxicity risk assessment. Federal Register 56(234):63798-63826.

US EPA, (1994). Principles of neurotoxicity risk assessment (final report). Federal register 59, 42360.

US EPA. (1996) Guidelines for Reproductive Toxicity Risk Assessment; October 1996. EPA/630/R-96/009.

US EPA (2002) A review of the reference dose and reference concentration processes. Risk Assessment Forum, EPA/630/P-02/002F, Washington, D.C., available at: http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=55365.

Vorhees, C. V. (1983) Foetal anticonvulsant syndrome in rats: Dose and period response relationships of prenatal diphenylhydantoin, trimethadione, and phenobarbital exposure on the structural and functional development of the offspring. Pharmacol. Exp. Ther. 227: 274-287.

Vorhees, C. V. (1985) Comparison of the collaborative behavioral teratology study and Cincinnati behavioral test batteries. Neurobehav. Toxicol. Teratol. 7: 625-633.

Vorhees, C. V., Weisenburger, W. P., Acuff-Smith, K. D., and R. D. Minck. (1991) An analysis of factors influencing complex watermaze learning in rats: effects of task complexity, path order and escape assistance on performance following prenatal exposure to phenytoin. Neurotoxicol. Teratol.; 13: 213-222.

Walsh, T. J., ad J. J. Chrobak. (1987) The use of radial arm maze in neurotoxicology. Physiol. Behav. 40: 799-803.

Weisenburger, W. P., Minck, D. R., Acuff, K. D., and C. V. Vorhees. (1990) Working memory and the Olton radial-arm maze: Effects of maternal phenylketonuria and prenatal phenytoin in rats. Teratol. 41: 615.

Wier, P. J., Guerriero, F. J., and R. F. Walker. (1989) Implementation of a primary screen for developmental neurotoxicity. Fundam. Appl. Toxicol. 13:118-136.

Wilson, J. D. (1965) Embryological Considerations in Teratology. In: Teratology: Principles and Techniques, Chapter 10. (Eds.) J. G. Wilson and J. Warkany, Chicago: The University of Chicago Press, 256.

Wise, L. D., Beck, S. L., Beltrame, D., Beyer, B. K., Chahoud, I., Clark, R. L., Clark, R., Druga, A. M., Fueston, M. H., Guittin, P., Henwood, S. M., Kimmel, C. A., Lindstrom, P., Palmer, A. K., Petrere, J. A., Solomon, H. M., Yasuda, M., and R. G. York. (1997) Terminology of developmental abnormalities in common laboratory mammals (version 1). Teratol.55: 249-292.

APPENDIX I

DEFINITIONS

Adverse effect: A treatment-related alteration from baseline that diminishes an organism's ability to survive, reproduce and adapt to the environment.

Anomalies: Marked deviation from the normal standard.

Embryo: The early or developing stage of any organism, especially the developing product of fertilization of an egg.

Fertility: The capacity to become pregnant or to impregnate. In humans, the capacity for producing viable offspring.

Fertilization: The fusion of the sperm and ovum resulting in the restoration of the diploid number of chromosomes.

Gestation: Length of time between conception and birth.

Hazard identification: The identification of the inherent capability of a substance to cause adverse effects.

Implantation: Attachment of the fertilized ovum (blastocyst) to the endometrium and its subsequent embedding in the compact layer.

No-Observed-Adverse-Effect-Level (NOAEL): Highest concentration or amount of a substance, found by experiment or observation, that causes no detectable adverse alteration of morphology, functional capacity, growth, development or life span of the target organism under defined conditions of exposure. Alterations of morphology, functional capacity, growth, development or life span of the target organisms of the same species and strain under the same defined conditions of exposure.

Perinatal: The period before, during and after the time of birth.

Pregnancy: The condition of having an implanted embryo or fetus in the body, after fusion of an ovum and spermatozoon.

Risk assessment: An empirical based paradigm that estimates the risk of adverse effect(s) from exposure of an individual or population to a chemical, physical or biological agent. It includes the components of hazard identification, assessment of dose-response relationships, exposure assessment and risk characterization.

Sexual maturation: Achievement of full development of sexual function and reproductive system.

Teratogenicity: Induction of structural abnormality.