Contents

Developmental neurotoxicity tests: Introduction

Background

The developing nervous system is especially vulnerable to certain chemicals (Bondy and Campbell, 2005; Rodier, 1995), and exposures may result in altered neural development with consequences that may be quite unlike the chemical's effects in an adult nervous system (Barone et al., 2000). For these reasons, regulatory agencies (OECD 2007; U.S.EPA 1998a) have promulgated testing guidelines for developmental neurotoxicity (DNT). DNT refers to any adverse effect of exposure to a toxic substance on the normal development of nervous system structures and/or functions (U.S.EPA 1998b). The basic purpose of DNT guideline testing is to act as an initial assessment and screen for the potential of chemicals to cause adverse neurodevelopmental outcomes.

Experimental Guidelines

The full history of the emergence of regulatory DNT testing can be found in Makris et al. (2009). Briefly, the basic design and test specifics of the US EPA test guidelines were developed at a workshop held in 1989, following which the specific guideline (US EPA, 1998a) were developed and eventually finalized in 1998. The OECD test guideline was based on that of the US EPA, but included enhancements developed through discussion and international agreement. The final guideline represents compromises in some areas of the need for and specific conduct of some tests (OECD, 2007).

Once put into practice there were some aspects of the guidelines that were modified to improve their sensitivity. The US EPA has issued guidance on specific aspects of testing, although the 1998 guideline has not been formally revised. These changes include: 1) increasing the dosing period to weaning, given the considerable nervous system development that is occurring up to that time; 2) increasing the sample size for neuropathology for better quantification; and 3) in certain cases, including acetylcholinesterase measurements in the dam and pups.

The OECD extended one-generation guideline was most recently developed, with the goal of allowing neurotoxicological assessments of a subset of offspring (neurotoxicity cohort) as part of a reproductive toxicity study. The need for including the neurotoxicity group is based on "existing knowledge" and "needs of various regulatory authorities", and its inclusion in all studies is doubtful. Since there are many fewer pups, the number of behavioral tests included is severely limited. It is not intended to provide a complete assessment of developmental neurotoxicity or a replacement for more detailed studies (OECD, 2011).

The guideline-specific requirements for each test measure are described in subsequent chapters. The following table presents aspects of general experimental design across these guidelines.

Table 1 – Guideline requirements for DNT and extended one-generation reproduction studies

	US EPA OPPTS 870.6300 (1998)	OECD 426 (2007)	OECD 443 (2011) Neurotoxicity cohort
Animal selection	Rat, do not use Fischer 344	Rat, justify other species	Rat, justify other species
Number of litters/ animals	20 litters/dose recommended	At least 20/dose	10 litters/dose out of 20/litter in larger study
Testing assignments	1 male or 1 female from each of 10 litters/dose for specific behavioral tests, allocated such that testing histories do not confound subsequent measures 6/sex/dose for neuropathology* *unofficial guidance to increase to 10/sex/dose	1 male and 1 female from each of 20 litters/dose for specific behavioral tests, allocated such that testing histories do not confound subsequent measures 10/sex/dose for neuropathology	1 male and 1 female from each of 10 litters/dose, all pups get all tests
Dosing period	GD6-LD10* *unofficial guidance to extend dosing to LD21 GD0=sperm positive	GD6-LD21 GD0=sperm positive	2 wk premating, 2 wk during mating, through LD21
Dose administration	Oral to dams	Most relevant route, usually oral, to dams Consider direct dosing to pups where warranted	Most relevant route, usually oral, to dams
Dose selection	High dose should produce some toxicity (decrease weight gain <20%) Low dose should not produce effects	High dose should produce some toxicity (decrease weight gain <10%) Low dose should not produce effects	Base on TK data
Culling of litters	PND4, aim for 4 male and 4 female	On or before PND4, aim for equal number males and females	PND4, aim for 5 males and 5 females/litter
Age nomenclature		PND0=day of birth but prefer post-coital to postnatal age	

Standard Evaluation Goals

Health Canada and the US EPA developed this document on the review and interpretation of submitted DNT data to provide guidance on how to evaluate the quality, the conduct, and resulting data derived from the behavioral methods employed in the OECD and/or EPA DNT Guidelines. The reviewer may be exposed to data collected under any of the current guidelines (OECD 2007; OECD 2011; US EPA 1998a), as guideline selection depends on when the studies were conducted, and whether the study was specific for the US EPA or other international organizations; this guidance is applicable to all. This guidance is for use by for regulatory agency scientists reviewing these DNT data, especially those who may not be experts in neurotoxicity or developmental neurotoxicity. The guidance was generated by an international collaboration between Health Canada and the US EPA. The overall goal of the guidance is to foster better and more consistent reviews of DNT behavioral data between these two countries. This guidance may also be useful for other international regulatory agencies or those interpreting data generated under the auspices of the OECD or US EPA DNT guidelines.

References

Barone S Jr, Das KP, Lassiter TL, White LD. (2000). Vulnerable processes of nervous system development: A review of markers and methods. *Neurotoxicology* 21:15-36.

Bondy SC, Campbell A (2005). Developmental neurotoxicology. *J. Neurosci. Res*. 81:605-612.

Makris SL, Raffaele K, Allen S, Bowers WJ, Hass U, Alleva E, Calamandrei G, Sheets L, Amcoff P, Delrue N, Crofton KM. (2009). A retrospective performance assessment of the developmental neurotoxicity study in support of the OECD test guideline 426. *Environ. Health Perspec.* 117:17-25

OECD Guideline for the Testing of Chemicals, Developmental Neurotoxicity Study No. 426, 2007.

OECD Guideline for the Testing of Chemicals, Extended One-Generation Reproductive Toxicity Study N. 443, 2011.

Rodier PM. (1995) Developing brain as a target of toxicity. *Environ. Health Perspect.* 103 (suppl 6):73-76.

U.S.EPA. (1998a). Health effects guidelines OPPTS 870.6300 Developmental Neurotoxicity Study, EPA/ 712/c-98/239.Office of Prevention Pesticides and Toxic Substances.

U.S.EPA. (1998b). Guidelines for neurotoxicity risk assessment. EPA/630/R-95/001F. Washington, DC:US EPA.

MODULE - A

EVALUATION OF OBSERVATIONS AND FUNCTIONAL TEST DATA

MODULE A - INTERPRETATION OF OBSERVATIONS AND FUNCTIONAL TEST DATA

1. Introduction

The use of observational and functional tests for screening in toxicology assessments gained popularity after several expert panels and scientists recommended such in the late 1970s and early 1980s. Their use is based on the premise that behavior represents the integrated sum of activities mediated by the nervous system, and is a sensitive marker of nervous system dysfunction. Taken from the Irwin screening battery that is widely used in drug development (Irwin, 1962, 1968), test batteries for behavioral evaluations in neurotoxicology, broadly known as functional observational batteries (FOB; Moser, 1989; McDaniel and Moser, 1993), were developed and validated in subsequent years. Based on the widespread and accepted use of such test batteries in adult neurotoxicity screening in rodents, observational and functional tests were included as part of DNT testing to evaluate overall neurological function in both the dam and offspring. However, FOB testing, *per se*, is not required in the EPA or OECD (426) DNT guidelines.

2. Test Description

When conducted optimally, observational test batteries provide a systematic and detailed evaluation of the animal's behavior and function. Measurements of complex movements, e.g., gait, reflect multiple neuronal functions, contribute to sensitivity of the approach but suffer with lack of specificity. On the other hand, tests of simple reflex behaviors, e.g., simple sensory responses, may be more specific but only be altered by a few neurotoxic agents. In general, screening batteries are based on noninvasive observations and manipulations, and rely heavily on careful evaluation of the individual animal. Some tests can include appropriate test devices (for example, automated motor activity chambers, strain gauges for grip strength) to provide quantitative, objective data. On the other hand, many measures are often subjective.

There is a range of approaches for these observations, from very simplified to more expanded clinical observations. These clinical observations vary widely across laboratories and there are no published protocols or guidelines. A lack of standardization in how and what data are collected negatively influences the ability to interpret or compare data across laboratories.

More specific screening batteries, such as the FOB, include a broad range of assessments. To accurately be considered an FOB, the protocol should focus on detailed observations as well as specific tests of reflexes, responses, and abilities. Using an FOB, several neurological functions must be assessed, including autonomic, neuromuscular, and sensory, as well as levels of activity and excitability. Numerous FOB protocols have been published, but even more are unpublished (e.g., those used in testing laboratories). As with clinical observations, there is a lack of standardization across laboratories.

3. Guideline requirements

All of the DNT test guidelines require close evaluations of both dams (during gestation and lactation) and offspring (from an early age to adulthood). The verbiage in the DNT guidelines

that describe the clinical observations (see Table 1) is, however, lifted from parts of the US EPA and OECD adult neurotoxicity test guidelines. There is a wide variety of test protocols being used in testing laboratories that generally fulfill the DNT guideline requirements.

It is important to understand that the EPA and OECD (426) DNT guidelines do not specify or even mention the use of the FOB. Unfortunately, many researchers and reviewers do not understand this distinction. On the other hand, the OECD extended one-generation study (443) does require observations in the main study, and does actually specify the FOB in the DNT cohort.

The test requirements are listed in Table 1. It is notable that the most recent guideline (OECD 443) actually has the most explicit description and requirements, whereas the earliest guideline (US EPA 870.6200) is the least specific.

Table 1. Requirements for observational and functional assessments in DNT test guidelines.

	US EPA 870.6300	OECD 426	OECD 443
Test subject and age	Dams at least twice during the gestational period, at least twice during lactation 10/dose group	Dams at least twice during the gestational period, at least twice during lactation At least 10/dose group	Dams on a weekly basis
	Offspring on PNDs 4, 11, 21, 35, 45, 60 10/sex/dose group	Offspring weekly preweaning, at least every two weeks in adolescence and as adults At least 1/sex/litter	All F1 animals on a weekly basis after weaning DNT cohort: 10/sex/dose between PND63-75
Test apparatus	Outside the home cage	Outside the home cage	Home cage and observation arena outside the cage
Specific observations	Autonomic function: lacrimation,	Changes in skin, fur, eyes, mucous	F1: Changes in skin, fur, eyes, mucous

	salivation, piloerection, exophthalmos, urination, defecation, pupillary function Convulsions, tremors, abnormal movements Posture, gait abnormalities Unusual or abnormal behaviors	membranes, secretions, autonomic activity Unusual responses in body position, activity, coordination, gait, posture Reactivity to handling, placing, other stimuli Clonic or tonic movements, convulsions, tremors Usual or abnormal behaviors	membranes, secretions, autonomic activity Unusual responses in gait, posture Reactivity to handling Clonic or tonic movements Usual or abnormal behaviors DNT cohort: Requires FOB, Appendix lists specific endpoints under categories of home cage and open field, manipulative, and physiologic (taken from several FOB publications)
Observers	Trained technicians, unaware of treatment. Demonstrate interobserver reliability if not same technician throughout	Trained technicians, unaware of treatment. Advisable to have same technician throughout	Trained technicians, unaware of treatment. Advisable to have same technician throughout
Protocols	Standardized procedures	Standardized procedures	Explicit operationally defined scales and scoring criteria Objective quantitative measures where possible

4. Observational Testing Procedures

4.1 Specific Protocols and Tests

There are differences in guideline requirements, but common to all guidelines are assessments of autonomic (e.g., salivation, lacrimation) and motor function (gait, posture), convulsive behaviors, as well as a catch-all phrase "unusual or abnormal behaviors". The OECD 426 adds mention of sensory responsiveness, and the OECD 443 specifically requires the more extensive FOB. It is likely, however, that individual testing laboratories use or will use the same protocol for all DNT studies.

In reality, testing laboratories often have one protocol of clinical observations that is used regardless of the specific requirements (i.e., for adult and DNT studies). The focus is on autonomic dysfunction, abnormal posture or movements, and unusual behaviors or appearance. In some protocols, measures of reactivity in response to handling or removal from cage are added. Responses to sensory stimuli, specific reflexes, or measures of grip strength are most often not included, and these clinical observations would not be classified as a valid FOB. Thus, even if a report states that a FOB was used, that is not the case unless the full battery of tests is used. The reviewer should understand these differences between clinical observations, which are most common, and a full FOB, which will likely not be used.

Regardless of which endpoints are included, clear defined protocols are critical to good experimentation. The sequence of tests should progress from the least (e.g., observations in the home cage and open field) to the most invasive (e.g., handling assessments) to minimize the influence of stress on subsequent measures. Since these are often subjective evaluations, explicitly defined scores and criteria should be used: anything less is subject to observer bias. Measures that are ranked provide more information on treatment effects than do binary, or all-or-nothing, measures. If used, binary measures should include a clear description of what constitutes "other than normal." With lethality, for example, "yes" and "no" are easily distinguished. However, the case for other evaluations such as activity is not so obvious, where "increased" vs "decreased", or "normal" vs "abnormal", must have associated with them operational definitions or else the measure is meaningless and large observer differences may occur. A ranking, or scale, describing different levels of activity would improve consistency across observers and would allow a reasonable evaluation of the data by the reviewer. The laboratory's standard operating procedure (SOP) should provide descriptions of each test measure, order of testing, and evaluation scales. The reviewer can request this protocol when necessary to understand better the submitted data.

4.2 Observers

Unlike the many automated and quantitative tests used in a DNT study, the observer is often the "instrument" collecting the data for these functional and observational measures. The guidelines highlight the need for trained technicians. Furthermore, the same technician throughout the study contributes to consistency, and information regarding comparability of technicians (inter-observer reliability) is required if the same technician is not used. Such information should be included in the main study report to assure that this was addressed.

Observational procedures are not difficult, but do require a significant level of comprehension and technique on the part of the observer. It is crucial that the observer be unaware ("blind") of the subject's treatment, in order to prevent deliberate or subconscious bias from affecting the data. The availability of a training video and manual for the FOB (available from VC Moser, US EPA) has done much to train observers in many types of settings, including academic, government, and commercial laboratories. The use of positive-control studies is also useful to demonstrate sensitivity of the protocol as well as expertise of the technicians. If conducted, these positive-control studies should encompass all test ages (as young as one week of age) with multiple behavioral outcomes.

4.3 Experimental Control

Good experimental practices of counterbalancing testing procedures, minimizing interfering factors, etc., are a necessary part of conducting behavioral tests. The response of the whole animal is measured; therefore, there is a multitude of possible influences from various sources. Extraneous factors (e.g., noise, light, odors, etc.) must be rigidly controlled as they may have a significant effect on the behaviors being monitored. Many of the specifics of conducting these tests have been summarized (Slikker et al., 2005). Since no training of the subject is involved, there is no explicit control over observed behaviors and variability may be high, suggesting a need for larger sample size. In general, behaviors that include the subject's innate responses (e.g., righting reflex, pain response) or a function that is important or critical to its survival (e.g., pupil response) tend to be less variable than spontaneous behaviors (e.g., activity, rearing). Strain and sex of the subjects may also influence the results (e.g., Moser, 1996).

4.4 Test Subject

The test guidelines include evaluation of both dams and offspring. While a single protocol may not be appropriate for evaluations at all ages, this seems to be a common practice in many testing laboratories. Explanations of modifications to adjust for these different situations are desirable. Such modifications should account for the immature reflexes and motor function of young animals, especially true at the youngest ages.

Testing pregnant dams requires care and minimal manipulations, especially late in gestation. Also, testing dams during lactation usually means separating her from the litter. The dams should only be removed for a short period of time (less than 20-30 min) or else the pups should be provided some means of thermoregulation. Likewise, when pups are removed for observations, thermoregulatory controls must be provided if they are away from the dam for very long. These situations can cause differences in behavior that are not the case for standard adult rodents, and therefore data may not be easily extrapolated.

The US EPA guidelines require these observations be conducted on at least 10 pups/sex/dose, and the OECD guidelines state 1 pup/sex/litter (which would provide 10/sex/dose if 20 litters/dose are used). As with all the DNT tests, the assignment of pups to observational testing should be one male and one female from each litter, or else one male or one female should be drawn from separate litters. For dams, at least 10/dose should be examined. It is desirable to use the same subject at each testing time, even though this requires identification of individual pups. There should be a random selection for which subjects are tested. Specific information for these issues should be included in the report.

5. Data Reporting

There are many dependent variables in these test batteries, making data summarization a challenge. At the minimum, tables of summary data across all endpoints should be provided. Where treatment effects are noted, individual animal data and/or more detailed information should be provided.

5.1 Data checks

There are specific checks that can be used to evaluate the adequacy of submitted observational data. Up front, there should be:

- Listing and description of what endpoints were included in the protocol and the scoring criteria or explicit descriptions used for each endpoint. Simply referencing a standard operating procedure is not appropriate
- Modifications made to adjust the observations at different stages of the study, including age of the test subject
- Training of the technician and whether the same technician was used throughout. If more than one technician was involved, inter-observer reliability should be described
- Procedures for assuring that the technician was unaware of the treatment of each test subject

Another data check is to evaluate the degree of variability. While this discussion is mostly applicable for controls, it should be noted that treatment may produce more or less variability and such information is also useful. While it may be counterintuitive, for some measures a lack of variability suggests inappropriate evaluations or scoring criteria. Behavioral measures that are not under experimenter control, e.g., reactivity or activity, will vary across animals. Scoring criteria without sufficient resolution to document these differences will not be sensitive to anything but the most obvious toxicity. In cases where all subjects receive the same score, across all treatments and test times, it is obvious that the criteria are not adequate. Often, this includes a phrase such as "normal" that does not explicitly define anything. Simply put, such evaluations have little value for detecting subtle to moderate behavioral changes. In addition, at some ages there may be more variability due to differences in the rate of complete nervous system maturation. Examples of measures for which some variability across animals should be expected are:

- Reactivity to handling, placing
- Activity and/or rearing in an open field or test arena
- Responses to sensory stimuli
- Gait, posture – to the extent that some rodents walk differently
- Urination, defecation

On the other hand, some functions are biologically controlled and show little variability. Any rare occurrence of these signs in control animals should be addressed. For example, lacrimation or salivation indicate autonomic dysfunction, and should not be seen in control animals.

Reflexes such as the righting response are stereotyped neurological actions and show very little variability; however, some of these reflexes are developing during the preweaning period and there could be differences in the subjects during this time. Specific observations reflecting poor health should not be seen in controls. Note that the wording of some required observations listed in the guidelines are vague, such as "changes in skin, fur, mucous membranes", and "unusual or abnormal behaviors", but these may be indicators of nonspecific illness. Some of the endpoints that should not vary in controls are listed below.

- Salivation, lacrimation
- Pupil response to light
- Piloerection, exophthalmos
- Convulsions, tremor – although young pups will display very fine "tremors"
- Major gait or postural changes
- Changes in righting reflex

5.2 Positive control data

Positive control data are useful for determining the range of neurological effects that can be characterized by the testing laboratory. Often, however, these studies are conducted in adult rodents with acute, high doses of neurotoxic chemicals that produce overt toxicity. These types of studies are less useful for assuring that the laboratory can detect subtle neurotoxicity, especially in very young pups. The DNT guidelines do not require positive control data with clinical observations.

6. Data analysis

The use of rating scales or descriptions makes statistical analyses of observational data more complex, since it is not appropriate to treat such data as continuous, parametric data. Transformations of the data may be necessary as well as consideration of non-parametric analyses. In addition, some data may show very low frequencies (e.g., 0, 1, or 2), making both analysis and interpretation more difficult. Typically, data with high variability should be examined for outliers or data errors; in such cases, evaluation of individual data is critical. For example, extreme values may be due to technician or instrument error, differences in individual sensitivity, or other factor. It is also important to note that numerous measures are collected on each individual subject. When littermates are tested, this should also be included in the analyses. Some appropriate approaches for analyzing behavioral data are described by Holson and colleagues (2008).

7. Interpretation

Interpretation is based on the information from individual endpoints and where possible, the profile or pattern of effects observed; for example, motor deficits would be evidenced as changes in gait, posture, righting, or other neuromuscular measures. When evaluating changes in specific endpoints, it is important to have an understanding of the variability, specificity, and influences on that measure. There is greater confidence in data collected quantitatively over observational data. The US EPA has published guidance for the evaluation of neurobehavioral data, including screening batteries (US EPA 1998). Neurotoxic adverse effects are defined as any change in the structure or function of the central and/or peripheral nervous system. This includes alterations in either direction (i.e., increases or decreases) from baseline or normal conditions as well as effects that are transient, occur only at specific times during development, or appear as changes in the ontogeny of developmental processes.

Observed changes in the dams on study indicate acute (or subchronic) toxicity of the chemical itself. It is important to note the time of testing in relation to chemical administration if it is given directly. For example with oral gavage dosing, testing after that day's dose could be

assessing acute but reversible effects from that dose. Time of testing is less important in studies where the chemical is administered via feed or water. Given that the dose ranges used in DNT studies often include a high dose with some systemic toxicity or effects of body weight, some significant changes might be expected. On the other hand, severe toxicity, e.g., convulsions, observed in the dam suggest that findings in the pups could be confounded by the maternal toxicity.

Effects observed in pups during the preweanling period may likewise reflect acute toxicity. This is especially true in studies where the pup is directly dosed with the test chemical. As with the dams, considerations regarding time of testing and severity of effects apply. For example, convulsions in pups indicate highly toxic doses and, if the pups survive, behavioral changes into adulthood may be expected. This should be considered when evaluating other DNT endpoints.

Most of the observations specified in DNT studies have been shown to be sensitive to direct chemical effects following acute or chronic dosing, and it is unlikely that such effects will be detected in adult offspring since by this age the animals are no longer receiving direct chemical exposures. For example, direct effects of chemicals that inhibit acetylcholinesterase include autonomic dysfunction, but this has never been seen in adult offspring that were developmentally exposed to those chemicals. The types of persistent effects that may occur as a result of exposure to developmentally neurotoxic chemicals, e.g., changes in reactivity, sensory responses, or neuromotor function, are mostly not assessed with the current guideline structure: the sole exception is the neurotoxicity cohort of the recent OECD 443 study. Consequently, the absence of findings in the adult offspring should not be used to discount effects in the young animal.

8. References

Holson RR, Freshwater L, Maurissen JPJ, Moser VC, and Phang W. (2008) Statistical issues and techniques appropriate for developmental neurotoxicity testing: A report from the ILSI Research Foundation/Risk Science Institute expert working group on neurodevelopmental endpoints. *Neurotoxicol. Teratol.* **30**, 326-348.

Irwin S. (1962) Drug screening and evaluative procedures. *Science* **136**, 123-128

Irwin S. (1968) Comprehensive observational assessment: Ia. A systematic, quantitative procedure for assessing the behavioral and physiologic state of the mouse. *Psychopharmacologia (Berl)* **13**, 222-257

McDaniel KL, Moser VC. (1993) Utility of a neurobehavioral screening battery for differentiating the effects of two pyrethroids, permethrin and cypermethrin. *Neurotoxicol. Teratol.* **15**, 71-83

Moser VC (1989) Screening approaches for neurotoxicity - a functional observational battery. *J. Am. Coll. Toxicol.* **8**, 85-93.

Slikker, Jr. W, Acuff K, Boyes W, Chelonis J, Crofton KM, Dearlove G, Li A, Moser VC, Newland C, Rossi J, Schantz S, Sette W, Sheets L, Stanton M, Tyl S, and Sobotka TJ. (2005). Behavioral test methods workshop: A summary. *Neurotoxicol. Teratol.* **27,** 417-427.

US EPA (1998) Guidelines for neurotoxicity risk assessment. *Federal Register* **63,** 26926-26954.

MODULE - B

EVALUATION OF MOTOR ACTIVITY DATA

MODULE B - EVALUATION OF MOTOR ACTIVITY DATA

1. Introduction

Overall study design plays an important role in determining the quality, reliability, and interpretation of motor activity data in DNT studies. Described herein are factors more specifically associated with design and conduct of motor activity testing in DNT studies and considerations in the interpretation of motor activity data. This document is not meant to be prescriptive but to provide background and the identification of design and test details for motor activity data that need to be considered in the interpretation of motor activity test data.

2. Test Description

As defined in the US EPA Test Guidelines (870.6200), motor activity is "any movement of the experimental animal". Although this is a very broad definition, motor activity is more typically considered to be locomotor movements in a horizontal direction (ambulation) as well as other directions (e.g., vertical, rearing). Smaller, fine movements (e.g., sniffing, grooming) may be included as motor activity but are not typically considered ambulatory, or locomotor, activity. Motor activity is an apical behavior that reflects a number of underlying processes including motor capacity, sensory functioning, emotional processing, non-associative learning (habituation), and integrative (cognitive) processes (Denenberg, 1969; Kelley, 1993; MacPhail et al., 1989; Maurissen and Mattsson, 1989; Ross, 2001). Decades of research in psychology, pharmacology, and toxicology have established the ability of motor activity measurements to provide insight in normal or altered nervous system function and development.

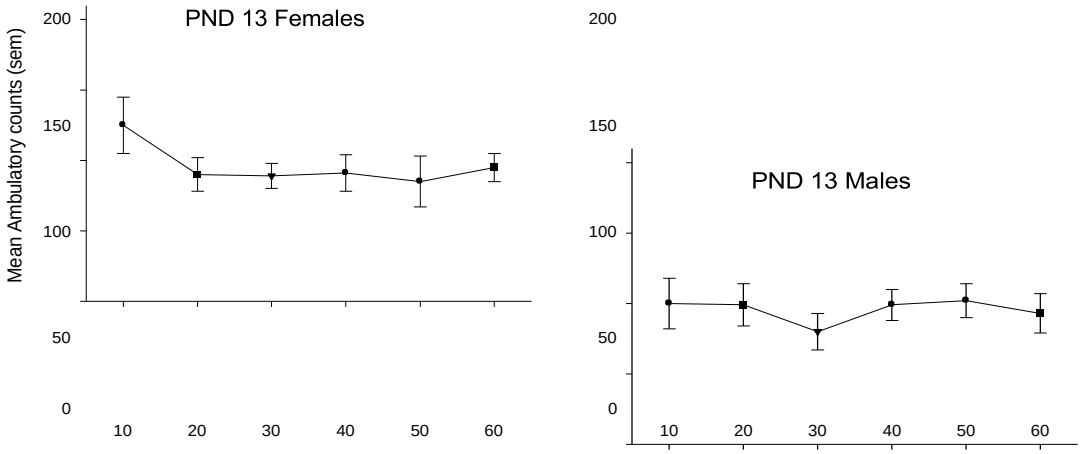
Motor activity tests are designed primarily to assess locomotor activity although modern automated test equipment provide a wide array of activity measures including ambulatory activity (movement between locations), exploratory movements (rearing or vertical movements) as well as a variety of fine motor movements (e.g., sniffing, scratching, grooming). It is important to be aware that automated equipment can provide measures of these actions separately or as an integrated measure of all activity movements (referred to as "total activity" in this document). An understanding of the measures that are operationally defined by the software or protocol is necessary to determine precisely what measures are actually reported. Since fine motor movements such as sniffing, scratching, etc., do not provide a measure of locomotor or ambulatory activity, care should be taken to ensure that activity test data clearly distinguish among these various activity measures and clearly indicate the type of the activity measures reported.

In DNT studies, motor activity measurements serve three main purposes. First, motor activity levels provide a sensitive measure of apical nervous system function. Second, the ontogeny of motor activity follows a developmental pattern that reflects the development and maturation of the nervous system. Third, motor activity studies provide a measure of non-associative learning (habituation), a basic form of learning essential to adaptive behavior and critical for normal interaction of animals with their environment. In the case of activity tests, habituation is normally measured as a decrease in locomotor movements over the course of the test session, although decreases in total motor activity (all measures of all motor movement) may also be used to assess habituation.

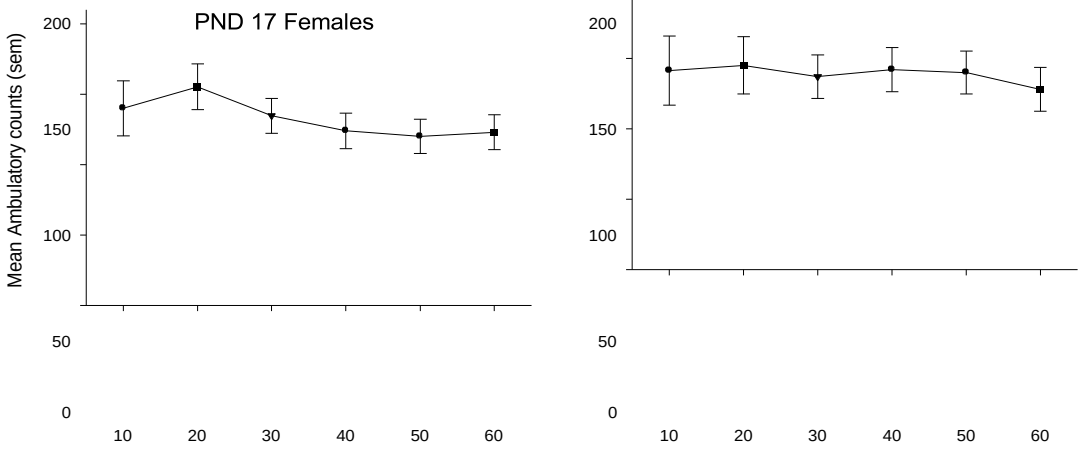
The ontogeny of motor function has been well investigated. The development of normal locomotor capacity in rodents begins around PND13 (Bolles and Woods, 1964; Bronstein, 1972; Shriner et al., 2009) although the required postural development is not fully developed at this age. Adult-form locomotion typically occurs after PND15-16 in rats but does not appear to be fully developed until around PND21 (Altman and Sudarshan, 1975; Muir, 2000). Indeed, both the neuronal and neuromuscular development required for adult-form locomotion occurs primarily in the third week of postnatal development in rodents (Clarac et al., 2004; Vinay et al., 2000, 2002).

Early studies on the development of motor activity show that in pre-weanling rodents, motor activity increases between PND12-13 to PND16-18 and decreases between PND20 and 23 (Campbell and Mabry, 1972; Moorcroft, 1971; Oakley and Plotkin, 1975; Shaywitz et al., 1979). In one laboratory, the peak period of spontaneous ambulation in Sprague–Dawley rats was between 19 and 22 days of age (Shaywitz et al., 1979). There are some results suggesting a role for environmental novelty in developmental patterns of motor activity. For example, Ruppert et al. (1985) reported that when the same Long-Evans rat pups tested repeatedly from PND13 to 21 they display an increase in activity from PND14 and 16. When pups were tested only on PND15, 18 or 21 this increase in activity was not observed (Ruppert et al. 1985); however, this study did not test PND13 and precluded the possibility of detecting the increase between PND13 and PND15-18. Others have suggested that peak ambulation between PND14 and 16 only occurs when the test environment is novel (Campbell and Raskin, 1978).

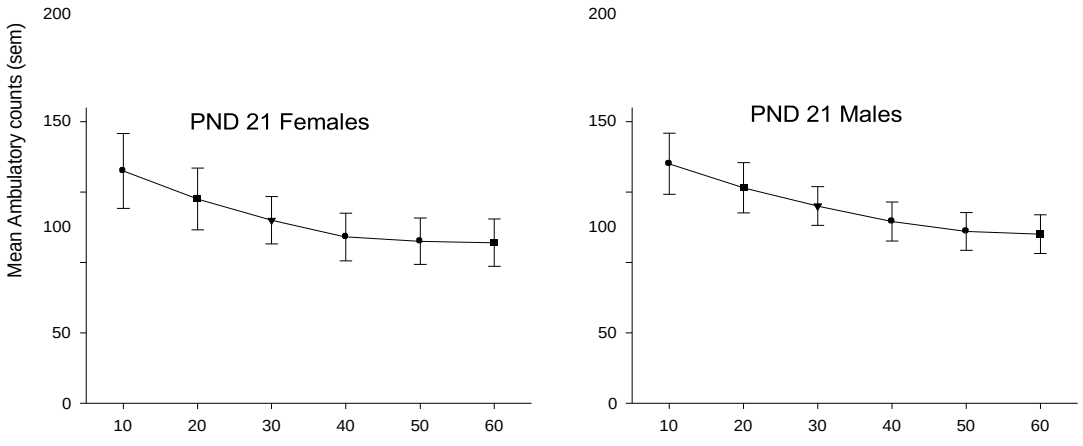
More recently, however, guideline studies from a number of laboratories have shown a pattern of increased locomotor activity at PND17 compared to PND13, and decreased or similar locomotion levels at PND21 compared to PND17 (Raffaele et al., 2003). An example of typical patterns of changing overall activity levels as well as development of habituation with age is shown in Figure 1. There have been some suggestions that the developmental profiles are relatively flat, while monotonically increasing patterns may also be evident, depending on the strain of animal, activity device, frequency of testing, and other experimental conditions; this has not been fully characterized. The EPA DNT guideline requirements imply that the same animals must be tested in activity tests in the pre-weaning period, which is important since experimental history and intra-session habituation which may influence subsequent activity levels. The OECD guidelines indicate that the same animals must be tested in the pre-weaning period if motor activity data are used to assess behavioral ontogeny.


PND 13 Females
Mean Ambulatory counts (sem)
200
150
100
50
0
10
20
30
40
50
60

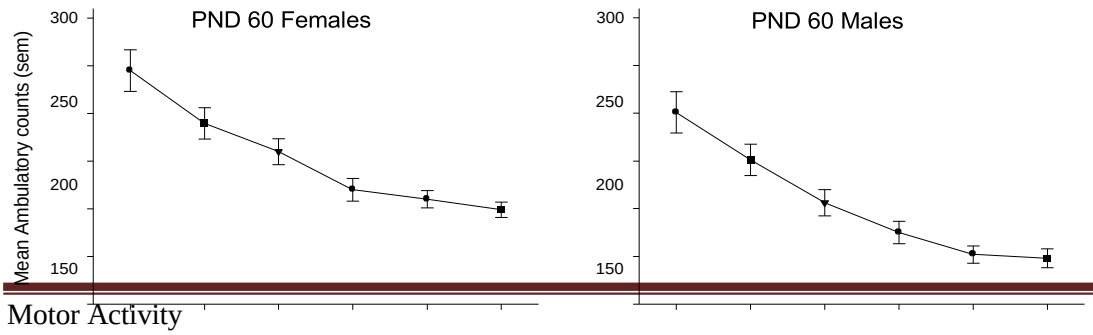
PND 13 Males
200
150
100
50
0
10
20
30
40
50
60


PND 17 Females
Mean Ambulatory counts (sem)
200
150
100
50
0
10
20
30
40
50
60

PND 17 Males
200
150
100
50
0
10
20
30
40
50
60


PND 21 Females
Mean Ambulatory counts (sem)
200
150
100
50
0
10
20
30
40
50
60

PND 21 Males
200
150
100
50
0
10
20
30
40
50
60


PND 60 Females
Mean Ambulatory counts (sem)
300
250
200
150

PND 60 Males
300
250
200
150

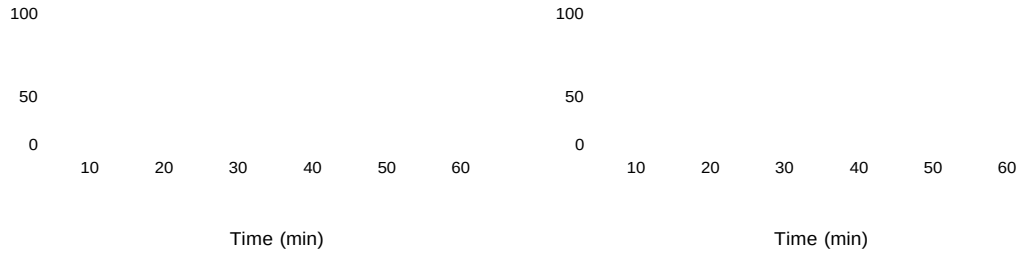


Figure 1. An example of motor activity counts across the sessions for a single group of male and female rats tested on PND13, 17, 21, and 60. Activity is very low at PND13 and higher at PND17 but habituation is not evident. At PND21 and 60, habituation is evident and in addition, female activity levels are often higher than males at PND60.

3. Guideline Requirements

While regulatory test requirements for the OECD and the US EPA are comparable for motor activity testing as part of DNT studies, there are some differences that should be considered. These are described in the table below. Note that the OECD TG 443 extended one-generation reproductive toxicity study that includes optional testing for motor activity employs the same test procedures as described below.

Table 1. Test requirements for OECD and US EPA DNT motor activity tests.

	US EPA OPPTS 870.6300	**OECD 426**
Age of testing	PND13, 17, 21, and 60 (±2 days)	1-3 times preweaning, once as young adults (PND60-70), Use of preweaning motor activity as measure of behavioral ontogeny "strongly recommended"[a]
Test duration	Duration sufficient to approach asymptotic activity levels "by the last 20% of the session" for controls All test sessions have same duration	Duration sufficient for detection of habituation in control animals
Data collection	Activity counts collected in blocks of ≤ 10 min	No mention
Study design	Testing of treatment groups should be counterbalanced over test device, time of day and each animal tested individually	

[a] Motor activity tests on PND13, 17, 21 fulfills OECD's requirement for behavioral ontogeny evaluation

4. Motor Activity Testing Systems

4.1. Test Chambers

Assessment of motor activity can be measured using a variety of techniques and equipment that will vary between laboratories. The shape, size, movement detection system, and data processing (software) may all affect the measured level and type of motor activity reported. It is important to be aware of these influences, especially in comparing motor activity results between studies and laboratories or even within laboratories when activity test equipment or procedures are changed. While there is a wide array of potential test arenas for assessing motor activity, most activity testing is conducted in one of four main types of chambers:

- Open-field type activity chambers (usually square or rectangular) of various sizes;
- Shaped chambers, such as figure-8 or doughnut configurations;

- Home cage activity monitors (usually photocell or toggle switch systems mounted onto home cages);
- Running wheels, usually in home cages.

4.2. Detection Systems

Both the OECD and the US EPA test guidelines require that motor activity testing be conducted in automated test chambers. Like the architecture of test chambers, a number of systems have been developed to automatically record motor activity. Most automated motor activity systems fall into one of the following categories for movement detection, although the last two types of systems are rarely, if ever, used in DNT studies:

- Photocell systems: these systems record the number of photocell beam breaks, with more complex systems recording the location and timing of photocell beam breaks;
- Video systems: these systems record and digitize the location of animals within the test arena and use software to characterize movement in the arena;
- Infrared sensors: these systems record changes in temperature associated with animals in the test arena and use software to characterize movement within the test arena;
- Jiggle or tilt-meters: these systems record the number of contacts of switches that detects changes in vertical force as animals move in the test arena and provide measures of gross body movements;
- Contact switches to measure rotations (for running wheels): these systems monitor the movement of a rotating wheel as animals run within a running wheel and are typically located in home cages of animals.

Because of these variations in the types of systems used to collect motor activity data, there are a number of potential motor activity metrics that may be reported. For photocell or contact switch systems, only counts are directly measured. Other parameters such as ambulatory distance and time are generated by software algorithms using photocell counts, changes in sequences of photocell counts, and timing of changes in photocell counts. Measures such as ambulatory counts, ambulatory distance, ambulatory time, etc., are generated with software algorithms that incorporate beam breaks, timing of beam breaks and location of beam breaks. For video-based systems, only distance moved and time in movement are actually detected. Software algorithms are used to generate measures such as ambulatory distance, time in ambulation, time immobile, etc. based on changes in location, speed, and nature of changes in location of the animals. In most automated activity measuring systems, ambulatory counts, ambulatory time and ambulatory distance are highly correlated. It should be noted that most photocell and video-based systems employ software settings to eliminate small movements (e.g., grooming, sniffing) from inclusion in ambulatory time estimates. Jiggle or tilt meters provide less detailed information on movement and typically provide information only on the number of switch contacts that indicate the number of gross movements.

4.3. Activity Measurements

Potential motor activity metrics include:

- Ambulatory counts: such as photocell counts, wheel rotations, tilt-meter contacts;

- Ambulatory distance: such as computed distance moved (e.g., based on software algorithm or based on number of photocells beams broken);
- Ambulatory time: such as amount of time animal is engaged in ambulatory movement; may vary depending on algorithm used to detect ambulatory episodes;
- Vertical movements (rearing): photocell based systems that incorporate a second set of photocell mounted at a height above the normal height of a prone animal will detect vertical movements of the animals. Video-based systems can also measure vertical movements based on changes in the size of the video image of the animal as it rears and thus changes size and shape. Infrared, jiggle or tilt-meters are not effective in measuring vertical movement of rodents;
- Small or fine motor movements: usually refers to discrete non-ambulatory movements such as sniffing, grooming, and stereotypy;
- Total activity: typically refers to the sum of all movements of the animal in the test chamber, and may include ambulatory, vertical (rearing) as well as small non-ambulatory movements (grooming, sniffing). Total activity should be clearly distinguished from total session activity which refers to a specific activity measure (e.g., ambulatory counts) summed or collapsed across the entire session. In this document, the terms total activity indicates all activity measured incorporated into one measure (e.g., ambulatory, vertical and small movements added together to generate one single activity measure), while total session activity refers to a single activity measures totaled over and time blocks in the test session (e.g., sum of ambulatory counts for all time blocks). The distinction between these metrics is important, even though some laboratory reports may not be clear on this.

Software settings for defining the type and threshold for activity units can be critical for computing measures of ambulatory activity. For example, some systems permit the setting of thresholds of the number of sequential photocell breaks required to trigger an ambulatory count. Because the size of the animal (e.g., young versus mature animals) and the distance between photocells will influence thresholds for recording ambulatory counts, it is important that these software threshold parameters be reported. For example, low thresholds for ambulatory triggers may result in high sensitivity for fine movements (e.g., twitching, sniffing, scratching, and head movement) that may be included in ambulatory counts and thus provide an inaccurate measure of true locomotor movement. Some software can adjust for the size of animals and can provide options for setting thresholds to trigger locomotor counts. Vertical positioning of photocell sensors can be critical for detecting vertical (rearing) movements, especially in young rodents since incorrect vertical location of photocell may preclude the possibility of detecting vertical movements in small animals. It should also be considered that in photocell systems, because of constant distance between photocells, breaking 4 photocells in sequence for a small animal (e.g., 30 gm) reflects greater locomotor distance than the same number of photocell breaks for an adult animal (e.g., 400 gm). In addition, some chambers may be more suited for specific ages, leading to the use of completely different systems for young and adult rats. For these many reasons, direct comparison of measures of activity between young and adult animals should be conducted with caution.

Similar to photocell-based systems, video-based activity systems also employ key parameters in software computed measures of ambulation, small movements, etc. Thresholds for both the amount of movement required as well as the speed of movement are employed in calculations of ambulatory distance and ambulatory time. As with photocell-based systems, differences in the size of young and mature animals will affect how these threshold criteria should be set since the

same criteria parameters will have different sensitivities for detecting ambulatory movements in small and large animals. In addition, video tracking systems are based on contrast, and are therefore prone to artifacts caused by reflections and other situations where the tracker does not accurately follow the rat. Reviewing video files would allow assessment of this, but this may not be feasible.

Consideration of measured parameters is also important for infrared systems where the size (weight) of the animal, distance from the sensor, as well as position relative to the sensor can all affect the signal. Chamber size may play a more important role in sensitivity for infrared systems than photocell or video systems.

In order to permit critical evaluation of motor activity data reported in DNT studies, reports should include sufficient detail to ensure that it is clear how activity data were collected and computed. This includes:

- Providing details of recording instrumentation (e.g., number and height of photocell beams, beam locations and distance between beams, or details of recording procedures for other detection methods such as video systems);
- Providing details on software versions and settings used for computing dependent variables. For example, photocell systems require software setting of thresholds for the number of photocell breaks required to trigger an ambulatory count, number of sequential photocell breaks (determined by size of animals). These settings will determine when an ambulatory count is recorded and should be set relative to size of animals. For video-based systems, software parameters for establishing thresholds to trigger ambulatory counts should be reported.

These settings affect the raw data generated, as well as the software-computed activity measures, and in turn will impact comparisons across studies and/or laboratories and comparisons across ages (e.g., preweaning, adult) within a study.

5. Test Procedures

Besides the multitude of environmental factors (e.g., lighting, noise) that may alter behavior in general (described in Slikker et al., 2005), there are a number of factors that specifically influence motor activity. Procedures to control these factors should be described in the protocol or experimental design and/or reported in the results. These include:

- Environmental factors such as sound levels, temperature in test area, lighting conditions, time of testing relative to light-dark cycle;
- Testing time in relation to daily dosing when either the dam or pup are being directly dosed (acute effects of dosing);
- Testing of treatment groups should be counterbalanced over test device, time of day and each animal tested individually;

- Standardized procedures for cleaning activity chambers, including procedures (chemicals and drying time) used between testing animals and between testing days;
- Identify how animals are loaded into activity chambers and how individual test sessions are started. If animals are stagger started, then identify precautions used to minimize

disturbance of animals already in the test chambers while other animals are loaded into test chambers. If started simultaneously, identify procedures used to place animals in test chambers and then start activity recording in all animals.

6. Data Reporting

The main results that are of interest from DNT motor activity testing are:

- The effects of chemicals on overall level of motor activity (total session activity);
- Effects of chemicals on habituation in motor activity;
- Effects of chemicals on the ontogeny of motor activity in pre-weaning animals;
- Sex differences in effects of chemicals on motor activity.

In order to evaluate the effects of chemicals on these parameters, it is essential that DNT motor activity reports provide clear information on study design, the measures (dependent variables), and the statistical procedures used to evaluate the main results. As noted in section 4, a number of possible measures can be generated from automated motor activity systems depending on the specific test equipment and software parameter settings employed. In the reporting and analysis of motor activity data, it should be clearly stated whether the total of all activity measures have been aggregated into a single measure (e.g., sum of ambulatory counts, vertical counts, and small movements). Likewise, if motor activity data are reported as only ambulatory counts, then this too should be clearly stated.

6.1 Dependent Variables

As described in section 4.2, there are a number of dependent measures that can be used depending on equipment setup and software parameters, and it is important that the dependent variables recorded and analyzed are unambiguously described. For most test systems, dependent measures reported may be one or more of the following:

- Ambulatory counts;
- Ambulatory distance;
- Ambulatory time;
- Average speed of movement
- Vertical movements (rearing);
- Small or fine motor movements;
- Total activity (some aggregation of multiple separate activity measures);
- Habituation.

Assessment of treatment effects on habituation of motor activity is one of the main results from DNT activity tests. However, habituation it is not a directly-measured dependent variable; it is actually an inferred result based on within-session decreases in activity measures.

While habituation can be assessed in a variety of ways, the main characteristic of habituation is a decrease in a response following repeated exposure to a stimulus that does not involve sensory adaptation/fatigue or motor fatigue (Grissom and Bhatnagar, 2009; Leussis and Bolivar, 2006; Rankin et al., 2009; Thompson, 2009; Thompson and Spencer, 1966). In the case of motor activity tests, habituation to the test environment is measured as a decrease in motor activity

over the course of a single test session. Test sessions should be of sufficient duration that asymptotic activity levels in control rats are reached before the end of the session (US EPA guidelines specify the last 20% of the session); the time necessary for habituation may vary across test arenas. The precise method used to determine habituation in DNT motor activity study results should clearly indicate how habituation is calculated. One measure is a calculation of the ratio of activity over the first and second halves of the test session. Another approach is testing for a significant effect of decreasing activity level over the time blocks for the duration of the test session (e.g., trend analysis). Other habituation metrics may also be used (e.g., first time block relative to the last time block). Regardless, the specific method to calculate habituation must be clearly described.

Many automated motor activity systems provide more than one measure of motor activity and may report them separately as well as include all measures in a single aggregated total activity measure. Since the total activity is correlated with the measures contained within it, these variables are not independent and therefore cannot be evaluated independently. In some cases, analyzing only total activity may be appropriate, but in other cases a more detailed analysis of separate measures (e.g., both horizontally and vertically-directed activity) may provide more detailed information. For example, it is possible that specific activity measures (e.g., vertical counts) may be affected by chemical exposure when total activity is not affected by chemical exposure. Justification for the choice of measures to report and/or analyze should be described.

6.2 Reported Data

In order to evaluate treatment effects, the following level of data detail should be reported:

- Activity measures should be described broken down by dose group, sex and time blocks at each test age. Note that the US EPA DNT guidelines require that each time block be no greater than 10 minutes;
- Activity measures should be reported as means ± standard deviations for each time block, treatment group and sex;
- Total session activity (i.e., collapsed across time blocks) should be reported as mean ± standard deviation for each measure of motor activity (e.g., ambulatory counts, vertical counts). Where total activity (i.e., all separate activity measures aggregated into a single activity measure) is reported, these should be reported by time block, sex and treatment group (means ± standard deviations) as well as collapsed across time blocks (means ± standard deviations).

Standard errors may also be reported as an estimate of population variability, but note that this measure may not be appropriate with very different sample sizes across groups.

7. Data Analysis

Measurement endpoints must be statistically analyzed using statistical methods appropriate for repeated-measures data. Statistical analyses specific for DNT data are described by Holson and colleagues (2008). As with all developmental data, the unit of analysis is the litter. Where a male and female are sampled from the same litter, the activity must be analyzed with sex nested within the litter, using litter as a random factor, or a hierarchical analysis with sex as a matching factor. While there are a number of approaches for the analysis of activity data, repeated-

measures ANOVA is the most generally accepted approach. Data analysis should include an overall ANOVA that includes treatment, sex and time blocks; one-way analyses of each sex can only be conducted in the face of significant interactions between sex and treatment. Furthermore, demonstrating within-session habituation requires repeated-measures analyses in order to show an effect of time block. Interval-by-interval analysis of dose effects at each time block is absolutely not appropriate in the absence of a statistically significant interaction between time and dose group from repeated-measures analyses. While most analyses assume normal distributions,

A number of relatively common problems are evident in the presentation and analysis of DNT data, and these limit the interpretation of study results and may produce misleading results. These include**:**

- Sex is often not included in the overall data analysis, with results for males and females analyzed separately. This approach not only precludes direct testing of sex effects and treatment-by-sex interactions, it also reduces statistical power to detect treatment effects;
- Motor activity results are analyzed separately at each time block. This approach precludes direct testing of a time-block effect, and any interactions associated with time blocks. Because a significant time-block effect is normally required to support demonstration of habituation, this approach precludes direct demonstration of habituation, or any treatment effects on habituation;
- Time block is not included in data analysis in cases where motor activity results are reported only for total session activity (i.e., collapsed across time blocks). Total session activity data reported as mean values collapsed across time blocks may mask treatment-related effects, as differences can exist at separate time blocks (i.e., treatment-by-time interaction).

In some laboratories, motor activity data can be highly variable with coefficients of variation (CV) ranging from 20-100% or more, depending on age, test conditions and the specific measures used. Since CV measures can provide an indication of the overall variability (and uncontrolled variation) associated with the test situation and the inherent variability of the behavior, it is not appropriate to use a fixed percent change in response (e.g., 20-25% change) in the absence of a statistically significant difference as an index of a treatment effect. With mixed-effects-modeling, a covariance matrix that best fits the data can be selected from many available types of covariance matrices. For analysis of results within a specific age group, factors include sex, dose and time blocks as the repeated-measure factor, keeping in mind that litter should also be included where there is more than one animal from each litter. Generally, interactions should be examined first, as significant interactions indicate that main effects are modified by a second factor. Interactions that could be examined include treatment-by-time blocks, treatment-by-sex, and treatment-by-dose-by-sex. Depending on the nature of the interactions (e.g., non-crossed interactions), main effects may also be examined.

7.1 Statistical Models

Below is a brief description of the types of statistical effects that may be obtained from analyses of DNT motor activity data:

- Main effects:

- Treatment: indicates that chemical exposure affects the overall level of activity collapsed across all other factors (i.e., sex, time blocks, and litter). This is effectively a direct statistical test of the effect of treatment on total session activity since time blocks and sex are ignored in this statistical test. A significant treatment effect indicates that the exposure affects total session activity;
- Sex: indicates a difference in response between males and females while all other factors (treatment, time block, litter) are held constant;
- Intersession Time Block: indicates that activity levels vary across time blocks in the test session. A significant effect of time block is required to demonstrate habituation. Note that a significant effect of time is necessary but not sufficient to demonstrate habituation. A significant effect of time may indicate an increase in activity levels over time blocks, while habituation requires a decrease in activity levels over time blocks.

- Interactions:
  - Treatment-by-time blocks: indicates only that the effect of time blocks varies between treatment groups (or alternately, the effect of treatment varies between time blocks). Normally this interaction is required to demonstrate that chemical exposure affects habituation. However, the precise nature of the interaction must be evaluated to determine how treatment groups differ over time blocks since; for example, it is possible to have a significant treatment-by-time interaction that does not indicate habituation (e.g., treatment increases activity levels over time in a test session but the size of the increase varies between treatment groups). Potential differences in patterns of habituation are illustrated in Figure 2. Acceptable approaches to assess the interaction include simple effects tests or specific contrasts, at each time block. Another approach is to treat time block as a continuous variable instead of as a categorical variable in the model, and the habituation is described by the slope (i.e., coefficient) of time block. A significant treatment-by-time block interaction indicates that habituation patterns are different between the dose groups;
  - Treatment-by-sex: indicates only that the effect of treatment differs between males and females. Note that the nature of the interaction must be evaluated, usually with simple effects or contrast interaction tests, or specific contrasts. It is possible to have a significant treatment-by-sex interaction that reflects that males and females are both significantly affected by the treatment in the same direction but the magnitude of the change differs between males and females;
  - Time blocks-by-sex: indicates that the pattern of activity across time blocks differs between males and females;
  - Treatment-by-time blocks-by-sex: indicates that the interaction between sex and treatment varies over time blocks (or, alternately, that the effect of treatment varies between males and females over time blocks). This interaction is normally required to demonstrate that the effect of treatment on habituation differs between sexes. As with other interactions, the nature of the interaction must be evaluated using contrast interaction tests, simple- or slice-effects, or specific contrast. It is necessary to show that the change in activity levels over the test session is influenced by treatment and that this change in activity levels is affected differently between sexes.

In addition to ANOVA results described above, the actual magnitude of change should be considered since effects of a large magnitude are sometimes not statistically significant due to high variability, but could be of substantial biological significance (also see discussion of

biological significance in section 8.5). This is especially useful where critical significance values (e.g., p-values) are inconsistent (e.g., results between age groups) due to variability in data, while effects size measures may be comparable. Lastly, actual significance values of statistical tests should be reported rather than critical p-values like $p<0.05$ since all modern statistical software programs report actual p-values. Reporting results of statistical testing using only critical p-values such as $p<0.05$ can result in qualitatively different interpretation of test results where $p=0.048$ and $p=0.051$ when there is no solid basis for treating these two statistical test results substantially different.

For the analysis and statistical treatment of motor activity ontogeny in the preweaning period (i.e., comparison of motor activity between preweaning test days), it is most useful to analyze total session activity rather than activity-by-time blocks, since habituation has not developed at younger ages (PND13/17). Thus, this analysis would involve a two-way ANOVA with total session activity as the dependent measure, and age and treatment as independent variables. As motor activity is typically not influenced by sex at these ages, sex may be excluded in the analysis. In some cases, separate statistical analyses for each age period may be justified. These include circumstances where there are significant differences in the test systems (equipment), relative measurement scales, data variability, non-linearity, etc. While the pattern of ontogeny may be evident, total-session activity data may not always achieve statistical significance due to the large variability seen at younger ages. Despite this, qualitative evaluation of these results should be conducted.

8. Interpretation of Motor Activity Data

8.1 Properties of motor activity control data

Regardless of the system used to measure motor activity, there are a number of behavioral properties in the measured data that should be evident and should serve as a reference point for evaluating reported results. Because of differences among the automated test systems commercially available, direct comparisons of absolute ambulatory counts between equipment types or laboratories are not likely to be useful. However, properties of behavioral measures of activity that should be present in control animals include:

- Control animals should demonstrate age-appropriate levels of activity;
- Control animals should show age-appropriate levels of variability in activity measures; for example, Figure 1 shows how the variability decreases with age;
- Habituation of ambulatory activity should be evident in animals tested at or after about three weeks of age.

The absence of these properties in control animals indicates that the normal activity pattern is not present in control animals and may indicate some unspecified problem in the test procedures.

Unambiguous interpretation of treatment effects on overall activity and habituation under these conditions is not possible since control animals are not exhibiting the behavioral profile that is being evaluated for a treatment effect. Historical control data can provide perspective regarding pattern of control activity at different ages based on equipment and laboratory procedures.

Motor activity positive control data should demonstrate the ability of the test equipment to detect

both increases and decreases in motor activity, and are therefore important for evaluating the relative proficiency of a testing laboratory in detecting chemical-induced changes. Positive control data help verify the ability of laboratory personnel in a realistic environment, and can help to characterize intra-laboratory and inter-laboratory reliability of a test method. The methods should be completely described, and must be the same as those used in the study being evaluated (e.g., the same equipment should be used, motor activity sessions should be of the same duration, the same number and strain of animals should be used). The data presentation should be sufficiently complete to permit evaluation of the sensitivity of the method, including individual data and measures of variability. The positive control data should have been collected within a reasonable time frame before the current study (e.g., the last few years). New positive control data should also be collected when personnel or other critical laboratory elements change. This new positive control data, while important for interpretation of treatment-related effects, can also be used to determine whether a shift has occurred in the control data collected by the testing laboratory. The lack of adequate positive control data can result in low confidence in negative data which in turn increases the possibility that negative results may in fact be a false negative finding.

8.2 Variability in motor activity data

Variability of control data is a key to understanding the laboratory's experimental control of the study. Historical control data from laboratories conducting the US EPA DNT guideline studies consistently demonstrate that motor activity is much more variable at PND13 and 17 than at later ages (Raffaele et al., 2003, 2008). Mean CV's for 14 laboratories was approximately 50–140% at PND13; 40–100% at PND17; 20–60% at PND21 and 18–30% at PND60 (Raffaele et al., 2008; data based on inspection of graphs). It is possible that some of this variability could be reduced by improving environmental experimental conditions (Raffaele et al., 2008). The high CV's in data collected at PND13 and 17 may be due, in part, to lower levels of motor activity as well as higher intrinsic variability during a period of rapid development of motor function when eyes are beginning to open. Other (non-DNT) studies in adult rats reported study means of 20-25% CV (Crofton et al., 1991), whereas Moser et al. (1997) reported mean CVs of 20-53% across studies from laboratories worldwide.

Excessive variation limits the ability to detect statistically significant treatment effects by increasing error variance. Where high levels of variability are reported, checking raw data values for unusual or extreme measures can help determine if treatment effects are masked by odd or unusual measures from individual animals that inflate variability and thus decrease power to detect treatment effects. Another potential reason for high CVs is that the data are not normally distributed, warranting evaluation of the underlying distribution or transformation of the data before analyzing. Review of the laboratory's historical control data (see section 8.1) will also provide a mechanism for assessing the variability in submitted studies. Increased variance in treated groups may also affect statistical power to detect significant differences but this increased variance may also be an important indicator of a treatment effect.

8.3 Evaluation of Treatment effects

A review of the data for each treatment group is necessary to look for unusual values, measurement errors, etc. Variability in treatment groups may be higher than in controls since there can be individual differences in sensitivity to the treatment, resulting in a range of responses. It is also possible that a single outlier value could reflect a chamber failure or some other non-treatment related factor. Thus, a survey of the actual values within each treatment group is an important first step in analyzing the data. For simplicity, it may be useful to examine the total session data before the habituation data.

Activity data should be appropriately analyzed across all treatment groups, and in the face of significant main effects or interactions, post-hoc tests may be conducted to either determine which groups are different from control (e.g., Dunnett's) or have multiple comparisons across all dose groups (e.g., Tukey-Kramer, or other methods that protect for alpha inflation from multiple comparisons). In studies where there are multiple levels of treatment (i.e., doses or concentration), changes in activity are more straightforward when there is a graded change (i.e., dose-response). There may be studies in which only the highest dose produces statistically significant changes, which could also be considered a form of dose-response. More difficult, however, are cases in which there are significant differences from control at only the low or middle doses, or some other non-linear response.

8.4 Treatment-related effects on ontogeny

A shift, or alteration, of the pattern of motor activity development may be an indication of treatment effects that is somewhat different from evaluating only habituation or total session data. For example, effects may not occur until a certain age, then persist, whereas other effects may only be evident early and then revert to control levels. An example of different patterns of changes was presented by Ruppert et al. (1985), where all three metals under study produced hyperactivity by PND21 but the patterns of change differed in timing, onset, and development of this effect. These specific changes are described in Table 2. Clear patterns such as these, however, are more difficult when only three days are sampled (as in current DNT protocols) compared to the daily sampling conducted in this study.

Table 2. Example of chemical-induced changes in the ontogeny of motor activity in pre-weaning rats.

Metal	Age at testing*		
	PND13-15	PND16-18	PND19-21
Cadmium	Decreased activity	Decreased activity but only on PND16; PND17 & 18 at control levels	Increased activity, peak at PND20
Triethyl tin	No change	Increased activity, peak at PND18	Increased activity
Trimethyl tin	No change	Increased activity, peak at PND16	Increased activity

* note that in this study, pups were tested daily from PND13 to 21. The data have been collapsed across 2-3 days at a time for presentation here.

8.5. Statistical versus biological significance

The conclusion of whether or not a chemical exposure has resulted in adverse outcomes on neurodevelopment should not rely solely on evidence of a statistically significant treatment-related effect. Holson and colleagues (2008) provide a well-balanced and in-depth review of appropriate statistical models for use with DNT studies. The use of appropriate statistical models for data analysis is critical to support conclusions. Beyond the discussion presented in section 7, a reviewer not conversant in statistical models may want to seek professional assistance before making any decisions about the nature of findings in a DNT study.

Issues to consider when assessing the biological significance of effects in DNT studies include variability in the reported data and the implications of this on the potential for false-negative (Type II error, more likely) or else false-positive (type I error) results. In addition, statistical significance of treatment effects does not necessarily mean that the statistically significant effect is biological relevant. Because of the large number of potential endpoints in DNT studies, care must be exercised in ensuring that experiment-wise errors rates are appropriately controlled to minimize false positives (Holson et al. 2008). However, guideline DNT studies are conducted to screen chemicals for possible adverse effects on the developing nervous system and are often the only study examining all of these endpoints. Thus, consideration of a higher false positive rate rather than a lower false negative rate may be a more conservative approach in some cases.

Most importantly, biologically significant, treatment-related findings may occur in the absence of statistical significance for a variety of reasons. Large variance in the measured behavior may occur due to poor control over testing environment or testing procedures, improper implementation of testing equipment, or inherent variability in the behavior. If reported variances in a given behavioral test compared with other reports from the same laboratory (historical controls) or with other laboratories suggest excessive variance within a study, the study itself may be flawed, at least for the parameter under consideration. This test would be considered in the context of the other measures in the study, at which point the particular parameter may not be used, or this could warrant a repeat of the study with better experimental control. It is critical to remember that statistical models are one of many tools that assist in data interpretation, not a sole replacement for expert judgment and careful scientific interpretation.

8.6. Dose-response results

Motor activity data typically show decreased levels as a result of treatment; however, both increased activity and inverted U-shape curves (increased activity only at a middle or lower dose) are not unusual. Increased activity in acute studies represents stimulatory effects on the nervous system. Persistent increases in activity may also reflect connectivity alterations in specific brain regions, but such associations have not been carefully characterized. In studies of acute effects (not typical in DNT studies), inverted U-shapes can represent decreases in inhibitory systems producing excitation, followed by generalized decreases (depression). Biphasic response data have also been reported in offspring following perinatal exposure to neurotoxicants (Gilbert et al., 1999; Rasmussen and Newland, 2001). Furthermore, it is not uncommon for only lower doses to be effective on behavioral measures, which may indicate that compensation or other mechanisms of toxicity may play a role higher doses (Robertson and

Grutsch Jr., 1987). Thus, while monotonic decreases in activity are common, increased activity as well as non-linear dose-response data may also be relevant and should be closely evaluated.

8.7. Habituation results

Demonstrating a significant treatment effect on habituation requires that control animals display normal habituation patterns observed around weaning or in adults. Generally, this would be indicated by a decrease in motor activity in control animals over the course of a test session; statistically this would be supported by a significant effect of time blocks in control animals. The demonstration of treatment-related effects on habituation would normally be reflected by alteration in the rate of change in motor activity levels over the course of the test session relative to the rate of change in motor activity levels in control animals. In Figure 2, different patterns are shown that may be observed as a function of dose of a single chemical; however, such patterns may also be produced by different chemicals. For example, both Dose 2 and Dose 3 show alterations in the rate of change in ambulation over the test session compared to control animals. Note that it is possible for treatments to either increase or decrease habituation. While habituation would be reduced to comparable levels in both Doses 2 and 3 (note parallel response curves), Dose 2 differs from Control in the last half of the test while Dose 3 differs from Control in the first half of the test. Dose 4 provides an example of potentiated or enhanced habituation, where ambulation counts are lower than Control levels in the second half of the test but not the first half of the test. Figure 2 also illustrates that reporting only total session counts may mask alterations in habituation. For instance, Dose 3 and 4 show reduced and enhanced habituation compared to Control, respectively, but total session counts for Dose 3 and 4 are not different from Control.

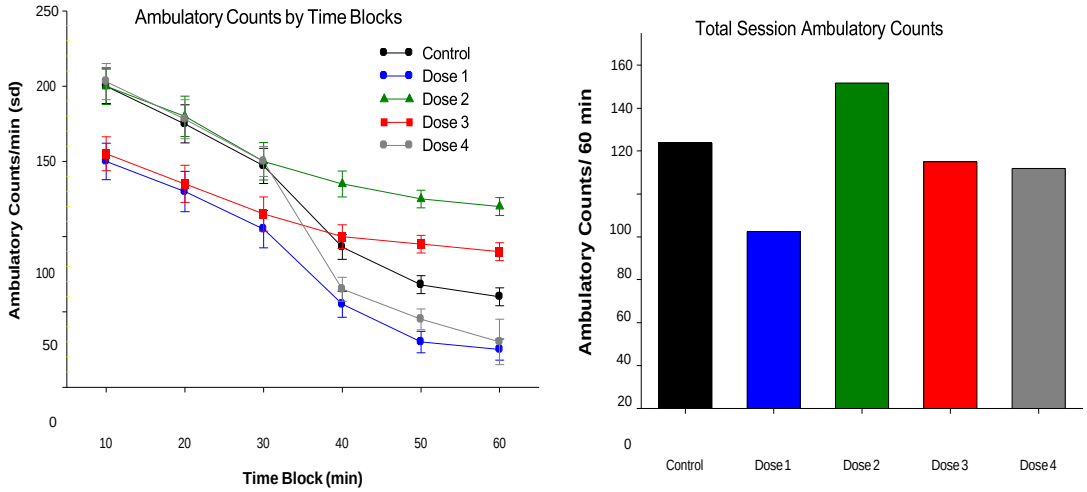


Figure 2: Examples of habituation data. Normal habituation (decreased ambulation over the test session; left panel) in adult rats is shown in Control and Dose 1, although Dose 1 shows overall lower ambulatory activity (right panel). Dose 2 and Dose 3 have comparable habituation patterns but have a reduced rate of habituation compared to Control and Dose 1. Dose 4 exhibits an enhanced rate of habituation compared to all other groups as illustrated by the more rapid decrease in ambulatory counts in the second half of the test session. The right panel illustrates total session activity collapsed across time blocks. As can be seen, without data presented as time blocks, it is not possible to evaluate habituation. In addition, Dose 3 and Dose 4 show comparable total session activity despite exhibiting opposite patterns of habituation. Similarly, Dose 2 and Dose 3 have different levels of total session activity despite having comparable patterns of habituation. In the absence of time block data, the conclusion that Dose 3 and Dose 4 are comparable (i.e., no treatment differences) would be erroneous. Similarly, while total session activity is comparable between Control and Dose 3 and Dose 4, using only total session activity would mask the fact that Dose 3 and 4 exhibit opposite alterations in habituation patterns and both differ from the Control habituation pattern.

9. General interpretation

Because motor activity is an apical behavior, non-specific or indirect effects on the nervous system from abnormal maternal behavior, malnutrition, growth retardation and/or other more general effects on development can also result in changes in motor activity (Li 2005). Alterations in motor activity may also reflect other effects such as altered sensory function and altered reactivity (emotionality) that result in behaviors that are inconsistent with motor activity (e.g., altered reactivity producing freezing). When supported by the appropriate statistical analyses, sex-specific alterations in activity (e.g., effects in one sex only) may suggest endocrine-mediated effects or differences in toxicokinetic parameters.

Although these alterations in motor activity may reflect neurotoxic effects, they may not necessarily reflect specific neurotoxic effects on motor function *per se*. While this is a disadvantage in specifying the nature of the neurotoxic effects, motor activity does have the advantage of monitoring a range of neurotoxic effects in addition to motor function. Because alterations in motor activity can be influenced by a variety of factors (e.g., effects on body weight), it is important for changes in motor activity to be evaluated in the context of other effects detected in DNT and other toxicity studies. It is also important to recognize that effects not specific to neurotoxicity could co-occur with treatment-related alterations in nervous system function or development. Changes in motor activity should not be dismissed or considered secondary to these non-specific toxicity findings.

10. References

Altman J and Sudarshan K. (1975) Postnatal development of locomotion in the laboratory rat. *Animal Behavior*, **23,** 896-920.

Bolles RC and Woods PJ. (1964) The ontogeny of behaviour in the albino rat. *Animal Behavior*, **12,** 427-441.

Bronstein PM. (1972) Open field behaviour of the rat as a function of age. *J. Comp. Physiol. Psychol.*, **80,** 335-341.

Campbell BA and Mabry PD. (1972) Ontogeny of behavioral arousal. A comparative study. *J. Comp. Physiol. Psychol.*, **81,** 371-379.

Campbell RA and Raskin LA. (1978) Ontogeny of behavioral arousal: the role of environmental stimuli. *J. Comp. Physiol. Psychol.*, **92,** 176-184.

Clarac F, Brocard F, and Vinay, L. (2004) The maturation of locomotor networks. In: *Progress in Brain Research: Brain Mechanisms for the Integration of Posture and Movement* (Ed. by D.G.S.Shigemi Mori), pp. 57-66. Elsevier.

Crofton KM, Howard JL, Moser VC, Gill MW, Reiter LW, Tilson HA, and MacPhail RC. (1991) Interlaboratory comparison of motor activity experiments: implications for neurotoxicological assessments. *Neurotoxicol. Teratol.*, **13,** 599-609.

Denenberg VH. (1969) Open-field behavior in the rat: what does it mean? *Ann. NY Acad. Sci.* **169,** 852-859.

Gilbert ME, Mack CM, and Lasley SM. (1999) Chronic developmental lead exposure and hippocampal long-term potentiation: biphasic dose-response relationship. *Neurotoxicol.* **20,** 71-82.

Grissom N and Bhatnagar S. (2009) Habituation to repeated stress: Get used to it. *Neurobiol. Learning Memory*, **92,** 215-224.

Holson RR, Freshwater L, Maurissen JPJ, Moser VC and Phang W. (2008) Statistical issues and techniques appropriate for developmental neurotoxicity testing: A report from the ILSI Research Foundation/Risk Science Institute expert working group on neurodevelopmental endpoints. *Neurotoxicol. Teratol.*, **30,** 326-348.

Kelley AE. (1993) Locomotor activity and exploration. In: *Techniques in Behavioral and Neural Sciences: Methods in Behavioral Pharmacology (vol 10)* (Ed. by F.van Haaren), pp. 499-518. Elsevier, New York.

Leussis MP and Bolivar VJ. (2006) Habituation in rodents: A review of behavior, neurobiology, and genetics. *Neurosci. Biobehav. Rev.*, **30,** 1045-1064.

Li AA (2005) Regulatory developmental neurotoxicology testing: data evaluation for risk assessment purposes. *Environ. Toxicol. Pharmacol.* **19**, 727-733.

MacPhail RC, Peele DB, and Crofton KM. (1989) Motor activity and screening for neurotoxicity. *J. Am. Coll. Toxicol.*, **8,** 117-125.

Maurissen JP and Mattsson JL. (1989) Critical assessment of motor activity as a screen for neurotoxicity. *Toxicol. Ind. Health*, **5,** 195-202.

Moorcroft WH. (1971) Ontogeny of forebrain inhibition of behavioral arousal in the rat. *Brain Res.*, **35,** 513-522.

Moser VC, Becking GC, Cuomo V, Frantik E, Kulig BM, MacPhail RC, Tilson HA, Winneke G, Brightwell WS, De-Salvia MA, Gill MW, Haggerty GC, Hornychova M, Lammers J, Larsen JJ, McDaniel KL, Nelson BK, and Ostergaard G. (1997) The IPCS Collaborative Study on Neurobehavioral Screening Methods: IV. Control data. *Neurotoxicol.*, **18,** 947-968.

Muir G. (2000) Early ontogeny of locomotor behaviour: a comparison between altricial and precocial animals. *Brain Res. Bull.* **53,** 719-726.

Oakley DA and Plotkin HC. (1975) Ontogeny of spontaneous locomotor activity in rabbit, rat, and guinea pig. *J. Comp. Physiol. Psychol.* **89,** 267-273.

Raffaele KC, Fisher, Jr JE, Hancock S, Hazelden K and Sobrian SK. (2008) Determining normal variability in a developmental neurotoxicity test: a report from the ILSI Research Foundation/Risk Science Institute expert working group on neurodevelopmental endpoints. *Neurotoxicol. Teratol.* **30,** 288-325.

Raffaele KC, Sette WF, Makris SL, Moser VC, Crofton KM. Motor activity in developmental neurotoxicity testing: A cross-laboratory comparison of control data. Presented at the annual Society of Toxicology meeting, 2003.

Rankin CH, Abrams T, Barry RJ, Bhatnagar S, Clayton DF, Colombo J, Coppola G, Geyer MA, Glanzman DL, Marsland S, McSweeney FK, Wilson DA, Wu CF, and Thompson RF. (2009) Habituation revisited: An updated and revised description of the behavioral characteristics of habituation. *Neurobiol. Learn. Mem.* **92,** 135-138.

Rasmussen EB and Newland MC. (2001) Developmental exposure to methylmercury alters behavioral sensitivity to D-amphetamine and pentobarbital in adult rats. *Neurotoxicol. Teratol.* **23,** 45-55.

Robertson AD and Grutsch, Jr JF. (1987) Biphasic responses, quantal signals and cellular behaviour. *J. Theor. Biol.*, **125,** 41-60.

Ross JF. (2001) Tier I neurological assessment in regulated animal safety studies. In: *Handbook of Neurotoxicology* (Ed. by E.J.Massarro), pp. 461-506. Humana Press, New Jersey.

Ruppert PH, Dean KF, and Reiter LW. (1985) Development of locomotor activity of rat pups exposed to heavy metals. *Toxicol. Appl. Pharmacol.* **78,** 69-77.

Shaywitz BA, Gordon JW, Klopper JH, Zelterman DA, and Irvine J. (1979) Ontogenesis of spontaneous activity and habituation of activity in the rat pup. *Dev. Psychobiol.*, **12,** 359-367.

Shriner AM, Drever FR, and Metz GA. (2009) The development of skilled walking in the rat. *Behav. Brain Res.* **205,** 426-435.

Slikker, Jr W, Acuff K, Boyes WK, Chelonis J, Crofton KM, Dearlove GE, Li A, Moser VC, and Newland C. (2005) Behavioral test methods workshop. *Neurotoxicol. Teratol.* **27**, 417-427.

Thompson RF and Spencer WA. (1966) Habituation: a model phenomenon for the study of neuronal substrates of behavior. *Psychol. Rev.*, **73,** 16-43.

Thompson RF (2009) Habituation: A history. *Neurobiol. Learn. Mem.* **92,** 127-134.

Vinay L, Brocard F, Clarac F, Norreel JC, Pearlstein E, and Pflieger JF. (2002) Development of posture and locomotion: an interplay of endogenously generated activities and neurotrophic actions by descending pathways. *Brain Res. Brain Res. Rev*, **40,** 118-129.

Vinay L, Brocard F, Pflieger J, Simeoni-Alias J, and Clarac F. (2000) Perinatal development of lumbar motoneurons and their inputs in the rat. *Brain Res. Bull.* **53,** 635-647.

MODULE C

EVALUATION OF AUDITORY STARTLE RESPONSE DATA

MODULE C - EVALUATION OF AUDITORY STARTLE RESPONSE DATA

1. Introduction

The design of the DNT study plays a critical role in determining the overall quality, reliability, and interpretation of startle data in DNT studies. Described herein are factors more specifically associated with the design and conduct of auditory, or acoustic, startle response testing in DNT studies as well as considerations in interpreting auditory startle data. This document is not intended to be prescriptive but to provide background and the identification of design and test details for auditory startle studies that need to be considered in the interpretation of test data.

2. Test Description

The US EPA Test Guidelines (TG) (870.6300) and the OECD TG 443 both require a test of auditory startle. The OECD TG 426 is less specific, only requiring testing of "motor and sensory testing". TG 426 allows flexibility in how to measure motor and sensory testing and provides the following examples: extensor thrust response, righting reflex, and auditory startle habituation.

The startle response is a characteristic set of reflexive muscle movements elicited by a sudden intense sensory stimulus (Davis and Eaton, 1984). In toxicology studies, acoustic stimuli are most commonly used to elicit the startle response, but visual and somatosensory stimuli have also been used (Ison 1984). The auditory startle response, elicited by sudden intense acoustic stimuli, is mediated by a simple neuronal circuit of four to six neurons between the cochlea, the brainstem, and spinal ventral horn motor neurons (Davis and Eaton, 1984; Koch, 1999). For a detailed review of the startle response the reader is directed to reviews of the physiology, psychophysiology and anatomy of the response, and its use in toxicology (Crofton, 1992; Davis, 1980; Davis et al., 1982; Fechter et al., 1986; Geyer and Swerdlow, 1998; Hoffman and Ison, 1980).

In DNT studies, the auditory startle response test is used to assess the integrity of a sensory-evoked motor response. In addition, the use of repeated test trials allows estimation of habituation to the startle stimulus (Geyer and Swerdlow, 1998). Habituation is a basic form of learning essential to adaptive behavior and critical for normal interaction of animals with their environment. Habituation is normally measured as a decrease in response amplitude after repeated startle stimulus presentation during the test session. Although not required or commonly used, some DNT studies will also include tests of reflex modification (Crofton, 1992; Ison and Hoffman, 1983; Li et al., 2009). Reflex modification is the change in a reflex response due to a perceptible and antecedent change in the sensory environment (Ison and Hoffman, 1983). Auditory pre-pulse inhibition of the startle response is a specific form of reflex modification. In DNT testing this typically involves the use of two auditory stimuli. One stimulus is a supra-threshold stimulus that reliably elicits a startle response (i.e., eliciting stimulus) and normally exceeds 110 decibels. The second stimulus (pre-pulse stimulus) is presented at a sound pressure level that is subthreshold (i.e., does not produce a startle response by itself).

For example, a prepulse stimulus of 20-50 msec duration may be presented 30-500 msec prior to the presentation of the startle stimulus (Swerdlow and Geyer, 1998). An effective prepulse stimulus will decrease the amplitude of the startle response, despite the fact that the prepulse stimulus will normally have no impact when presented alone. Reflex modification, while rarely used in regulatory studies, can provide valuable information to discriminate between a dysfunction in the sensory or motor component of the auditory startle response (Crofton, 1992; Young and Fechter, 1983)**.**

When used in guideline DNT studies, auditory startle testing is conducted at an early post-weaning age, as well as a young adult age near the end of the study. The ontogeny of auditory startle response as well as reflex modification has been well studied. The startle develops late in the second week of life in rats and mice. The startle response is tightly linked to maturation of the cochlea and opening of external auditory meatus, with the hearing and musculature with responses beginning within the second postnatal week (Brunjes and Alberts, 1981; Parisi and Ison, 1979; Sheets et al., 1988; Shnerson and Willott, 1980). Development of auditory reflex modification begins around 15 days of age and shows an adult pattern by PND21 (Parisi and Ison, 1979). Sheets et al. (1988) studied the ontogeny of both the amplitude and latency of the startle response, as well as the impact of background noise levels in rats from 13 to 21 days of age. This work demonstrated a decline in response latency as age increased, but the behavior had not completely reached adult status by PND21. Ontogeny of habituation of the response has not been as well studied and is thought to mature somewhere between PND18 and 25. Early work by Williams et al. (1975) found no habituation at PND17 and 'profound' decrement across repeated trials at PND36. Other reports have shown habituation occurring by PND20 or PND23 (Goldey et al., 1994; Wise et al., 1997), but not to the extent found in adults. It can therefore be concluded that the startle response and habituation thereof, while not fully mature, are measureable within days of weaning.

Startle testing is normally conducted using automated test equipment that can provide a wide array of response characteristics, the most important of which are measures of amplitude and latency (see section 4 below). Critical to proper interpretation of results in DNT studies is an understanding of the test system and the measured outputs that are operationally defined by the software used to process the response. This includes both the equipment used to produce, as well as calibrate, the response measuring device as well as the acoustic stimuli used to elicit the response.

3. Guideline Requirements

Neither the US EPA nor the OECD provide specific guidance on the types of test systems that should be used (Table 1). The EPA guideline specifies the use of 5 blocks of 10 trials per session, and provides a reference for specific details (Adams et al., 1985b) that outlines the use of an automated testing system with specific acoustic stimuli, number of trials, and data collection procedures. OECD 426 provides no specifics on testing methods for auditory startle, but does have a series of references for both auditory habituation and reflex modification procedures. OECD 443, similar to the EPA guideline, states that testing should include 50 trials, consisting of data blocked into 5 blocks of ten trials, but also indicates that test conditions should be optimized to produce intra-session habituation.

The remainder of this section is intended to provide a brief outline of the kinds of experimental equipment needed to test auditory startle response habituation. Since reflex modification is a more complex procedure, and is rarely used, the reader is directed to published reviews as stated above for more information.

Table 1. Auditory startle response testing requirements listed in the US EPA and the OECD guidelines.

	US EPA OPPTS 870.6300	**OECD 426**	**OECD 443**
Age of testing	"around the time of weaning and around day 60"	Recommended ages for adolescent testing are: learning and memory = PND25±2; motor and sensory function = PND25±2. Recommended ages for testing young adults is PND60-70. "the minimum number of times when measurements should be performed. Depending on the anticipated effects, and the results of the initial measurements, it may be advisable to add additional time points (*e.g.*, aged animals) or to perform the measurements in other developmental stages"	An auditory startle test should be performed on PND24 (±1 day) using animals in cohort 2A. The day of testing should be counterbalanced across treated and control groups.
Assignment of pups	"one male or one female from each litter (total of 10 males and 10 females per dose group)"	"20/sex (1/sex/litter)" "the same or separate pairs of male and female animals may be assigned to different behavioral tests."	"Ten male and 10 female cohort 2A animals and 10 male and 10 female cohort 2B animals, from each treatment group (for each cohort: 1 male or 1 female per litter; all litters represented by at least 1 pup; randomly selected) should be used for neurotoxicity assessments. Cohort 2A animals should be subjected to auditory startle…"

Test device	"Details on the conduct of this testing may be obtained in" (Adams et al., 1985a).	No specifics listed	No specifics listed
Study design	"The mean response amplitude on each block of 10 trials (5 blocks of 10 trials per session on each day of testing) should be made. While use of prepulse inhibition is not a requirement, it is highly recommended	"Motor and sensory function should be examined in detail at least once for the adolescent period and once during the young adult period (*e.g.*, PND60-70). "…examples of tests for motor and sensory function are … auditory startle habituation" Cited refs: (Adams, 1986, Crofton, 1992, Crofton et al., 1994, Crofton and Sheets, 1989, Davis and Eaton, 1984, Ison, 1984)	"Each session consists of 50 trials. In performing the auditory startle test, the mean response amplitude on each block of 10 trials (5 blocks of 10 trials) should be determined, with test conditions optimized to produce intra-session habituation. These procedures should be consistent with OECD TG 426 (35).
Data collection	No specifics listed	No specifics listed	No specifics listed
Data reporting	"Auditory startle response amplitude per session and intra-session amplitudes on each day measured." "The mean and standard deviation for each continuous endpoint at each observation time."	No specifics listed	No specifics listed

4. Auditory Startle Response Testing Systems

4.1 Testing Equipment

Auditory startle testing requires the following: 1) sound dampening chambers in which the animals are tested, 2) equipment and software used to generate the auditory stimuli, and 3) equipment and software needed to measure the response.

Most testing apparatuses include sound dampening test chambers. Sound dampening is important for startle testing because the level of background noise during startle testing is known to impact the amplitude of the startle response via a process called sensitization (Davis et al., 1984; Sheets et al., 1988). The level of background noise impacts the startle response in an inverted-U shape function, with increasing response levels tracking increasing background up to a point at which higher backgrounds actually decrease response levels (Ison and Russo, 1990). The critical aspect of this variable is to maintain background noise at a constant level throughout testing, and quantification of this background level should be included in the report. Commercially available testing systems do this by a combination of sound-dampening material and supplemental background white-noise, either from speakers or venting fans, or both. Note that the level of background noise must be below any acoustic stimuli used during testing.

The generation, presentation, and quality control of acoustic stimuli is a complex process. There are a number of important issues that should be kept in mind when reviewing DNT studies, and should be included in the reports, including:

- Study reports should clearly document how the sound levels of acoustic stimuli were calibrated. The appropriate scale for sound level is sound pressure level (SPL) measured in dB. Note that the use of sound calibration equipment with "A-scale" weighting is not recommended in most circumstances. This weighting scale was developed to mimic the human hearing range and it not appropriate for use with rodents that can perceive sounds at much higher frequencies than humans. A review of species specificity in hearing can be found in Fay (1988);
- Selection of speakers for use in presenting the acoustic stimuli should be capable of responding in the proper frequency range with the proper frequency characteristics (e.g., rise and fall times, lack of acoustic transients). The use of a broad band signal is preferable to pure tones;
- Test chambers should permit sufficient movement, i.e., minimal restraint, so that animals can generate startle responses. Note that the appropriate chamber size differs according to age of the animal. A holder too small will confine the animal and prevent a full startle response, and one too large could allow the animal to move away from the speaker, altering the delivered stimuli.

There are two main types of equipment used to record the startle response. The first is a load cell (or force transducer), and the second is an accelerometer. Because these sensors have different properties and record different information, it is important to understand the differences between these types of devices and the limitations of the data collected with them. Typically, both systems output an analog voltage from the sensor to an analog-to-digital converter that transforms the analog signal to a digital signal that is stored for analysis in a computer system. It should be noted that commercial startle equipment has improved over the years, but DNT data that are reviewed may have utilized either older or newer systems.

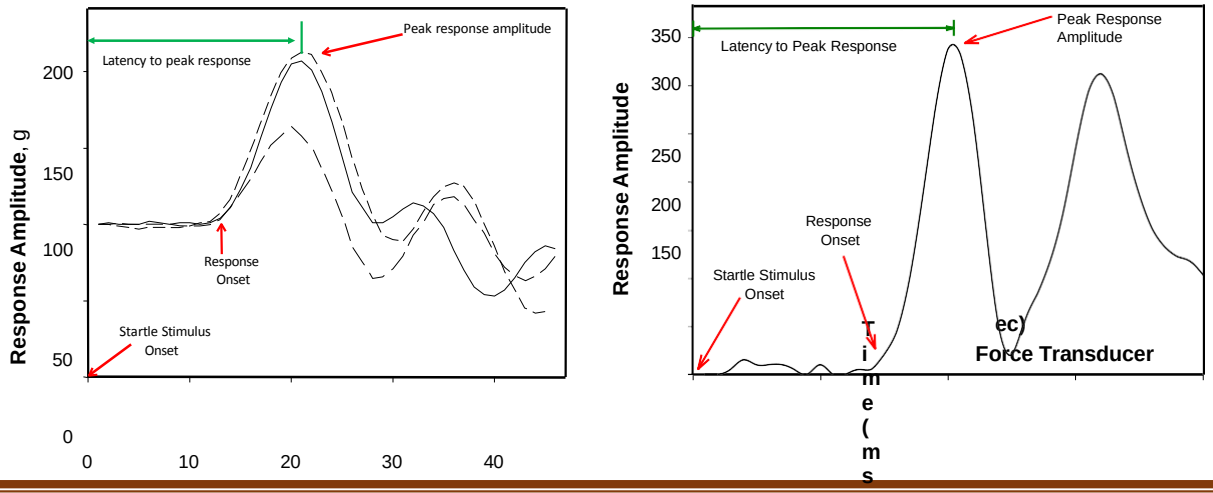
A load cell converts mechanical force into electrical signals that are output as voltage. It measures the force (mass and acceleration) applied to the testing platform by the animal. That force is influenced by the animal's body weight, thus weight impacts the starting, or resting, point, as well as the magnitude of startle response. Since the load cell and platform typically have very little movement,

the response is often presented as a mass measurement (e.g., grams, or force in grams). Most load cells are based on strain gauges that use minute deformations in electrical wires to alter resistance in, for example, a Wheatstone bridge circuit. With load cells a maximum voltage output is recorded at the time of maximum force applied to the platform. An advantage of load cells is the ability to calibrate output voltage to standardized weights, which allow voltage responses to be converted to grams. Some load cells are based on piezoelectric sensors. However, a disadvantage of piezoelectric sensors is that they have no voltage output under static, non-moving conditions, and thus lose the ability to calibrate with standardized weights.

An accelerometer is a device that measures acceleration forces. Early startle response testing systems used mono-phonograph cartridges resting on the testing platform. As the animal responds, the platform moves and this movement is detected by the change in resistance of the phono cartridge. Other versions of accelerometers are either piezoelectric or capacitance-based. All of these devices measure the change in motion, i.e., the acceleration, of the testing platform. With accelerometers, the maximum voltage output is recorded at the time that the platform is undergoing the largest change in movement. Accelerometers are usually calibrated with pulse calibrators ("thumpers") that provide a standardized movement to the platform. While some systems allow for conversion of voltage output to Newtons, the use of only one calibrator may not be adequate to determine linearity of the system needed for accurate conversion of output voltages to Newtons at all output voltage levels. Some systems convert the voltage output to units based on the software, which may not represent units that are comparable across laboratories.

Regardless of the detection system used, most commercial test systems sample the detection device at a rate of 0.5 or 1 kHz (sampling usually about every 1 or 2 msec) for up to 300 msec. This allows for a collection of data points that can be used to calculate the main metrics commonly reported. It is important to remember that the peak startle response is an extremely rapid reflex that begins some 10-14 msec after onset of the eliciting stimulus, and peaks within about 20-40 msec after the onset of the startle stimulus (Davis, 1980; Davis and Eaton, 1984).

Sample trace recordings from load cell and accelerator systems are illustrated in Figure 1. The figures illustrate the three major metrics typically reported for most startle studies. These are: response magnitude or amplitude, latency to response onset, and latency to peak of response.



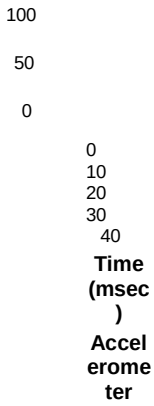

100
50
0
0
10
20
30
40
Time (msec)
Accelerometer

Figure 1. Trace recordings of startle responses from adult male rats exposed to a 115 dB auditory startle stimulus. The left panel shows results from a force transducer (load cell measured in grams of force) while the right panel shows results from an accelerometer. Note that the response onset and peak response times are comparable in both systems. Note that the right panel shows two peaks and in some cases the second peak is the maximum peak. Latency to startle onset is the time between the onset of the startle stimulus and the onset of the startle response.

5. Test Procedures

Many experimental (e.g., environmental conditions, lighting, noise, test apparatus) and organismic factors (e.g., age, test history, sex) are known to alter behavior (Claassen, 1994; Slikker et al., 2005; Tyl et al., 2008). In addition, there are a number of factors that specifically influence startle responses (Davis and Eaton, 1984; Hoffman and Ison, 1980). Procedures to control these factors should be described in the protocol or experimental design and reported in the results. These include:

- Environmental factors such as sound levels, temperature in test area, lighting conditions, time of testing relative to light-dark cycle;
- Testing of treatment groups should be counterbalanced over test device, time of day and each animal should be tested individually;
- Standardized procedures (chemicals and drying time) for cleaning chambers, including procedures used between testing animals and between testing days;
- Identify how animals are loaded into chambers and how individual test sessions are started. If animals are stagger-started, then identify precautions used to minimize disturbance of animals already in the test chambers while other animals are loaded into test chambers. If animals started simultaneously, identify procedures used to load all animals quickly in test chambers and then start activity recording in all animals;
- Testing history of the animals. It should be clearly described whether the animals have been tested for any other behavioral evaluations prior to startle testing;
- Identify if a period for habituation to the chamber is included in test procedures (e.g., 5 min habituation to chambers prior to startle trials).

6. Data Reporting

In order to permit critical evaluation of startle data reported in DNT studies, reports should include sufficient detail to ensure that it is clear how data were collected and computed. These settings affect the raw data generated, as well as the software-computed measures and in turn will affect comparisons across studies between and within laboratories and comparisons across ages (e.g., preweaning, adult) within a study. The information that should be provided includes:

- Details of the recording instrumentation (i.e., whether the system uses a load cell or an accelerometer) including software;
- Details on how the recording system was calibrated and how this calibration was used to calculate response units (e.g., grams, Newtons);
- Definitions of all calculated variables (e.g., peak response magnitude, latency to peak

response);

- Details on the testing chambers, including background noise levels;
- Details on the generation of acoustic signals (frequency, duration, rise times, inter-trial intervals), and how these signals were calibrated;
- Details on procedures used to select response data. The first peak may be considered the startle response even though it is not always the maximum peak. Indeed, it is not entirely clear how multiple peaks should be evaluated (Grimsley et al, 2007) (Figure 1 shows examples of multiple peaks);
- Criteria used to define non-startle responses as general movement (e.g., long latency values for first peak would not be a startle response; small peaks may reflect movement rather than startle). Note that animals do not always startle on every trial and this should be addressed in data analysis. While averaging trials over blocks may help to lessen the impact of trial-to-trial variability, treatment-related increases in the number of low- or no-response trials may be detected by examining individual data.
- Some laboratories include "blank" trials, which record response measurements without an eliciting stimulus. This could reflect baseline activity, or non-stimulus related startles, and should be reported against which to judge the response data.

The main results that are of interest from DNT startle testing are:

- The effects of chemicals on the overall startle amplitude (average of startle response collapsed over all trials or total session mean);
- Effects of chemicals on habituation in startle magnitude over the test session;
- Comparison of effects between the two ages of testing;
- Sex differences in effects of chemicals on startle response amplitude and/or habituation.

In order to evaluate the effects of chemicals on these parameters, it is essential that DNT reports provide clear information on study design, precise definitions of the dependent variables, and the statistical procedures used to evaluate the results.

The terminology used in reporting these measures can vary among laboratories and, therefore, in DNT reports. Table 2 lists the variety of measurement names that were used in a total of 45 DNT studies submitted to the US EPA. This represents actual data submissions and is presented for cross-reference when reviewing startle data. In many cases the amplitude variable is the maximum amplitude of the response. Note that average response amplitude differs from the maximum response magnitude in that it incorporates the average over the entire recording period rather than the peak response magnitude. Response latency may be measured as latency to response onset or to the peak startle response. Care should be exercised when reading the study in order to determine exactly what the measurement is intended to represent.

Table 2. Common reporting terms for startle response data and biological measurement interpretation	
Reported Term	**Measurement**
amplitude (Newtons)	Amplitude
amplitude (volts)	Amplitude
maximum amplitude (no units)	Amplitude
maximum amplitude of the response (mV)	Amplitude
maximum impulse (arbitrary units)	Amplitude
maximum response (Vmax)	Amplitude
maximum response amplitude (Vmax)	Amplitude
maximum startle (voltage)	Amplitude
mean peak startle (no units)	Amplitude
peak amplitude (grams)	Amplitude
peak response amplitude (v)	Amplitude
reflex peak amplitude (V)	Amplitude
response (g)	Amplitude
latency (msec)	Latency
latency (Tmax)	Latency
latency to maximum response (msec)	Latency
latency to peak (msec)	Latency
latency to peak of the response (msec)	Latency
latency to peak response	Latency
mean time to maximum amplitude	Latency
peak latency (msec)	Latency
reflex latency (msec)	Latency
time of maximum startle (msec)	Latency
time to maximum response (Tmax)	Latency
time to maximum startle (msec)	Latency
time to peak amplitude (msec)	Latency
percent inhibition	Sensory inhibition
prepulse inhibition	Sensory inhibition
average response (grams)	Undefined in report
average response amplitude (mV)	Undefined in report
response duration (msec)	Undefined in report

6.1. Dependent Variables

There are a number of dependent measures that can be used depending on equipment setup and software parameters, and it is important that the dependent variables recorded and analyzed are unambiguously described. Based on the biology of the response, the following are the most common startle response metrics:

- Maximum startle response –maximum amplitude of the response, recorded in grams for load cells and as equipment-specific units for accelerometers;

- Habituation - calculated metric for characterizing the change in startle response magnitude over repeated exposure to the startle stimulus.
- Latency to onset or peak – time in msec from the startle stimulus onset to the beginning or to the peak of the startle response

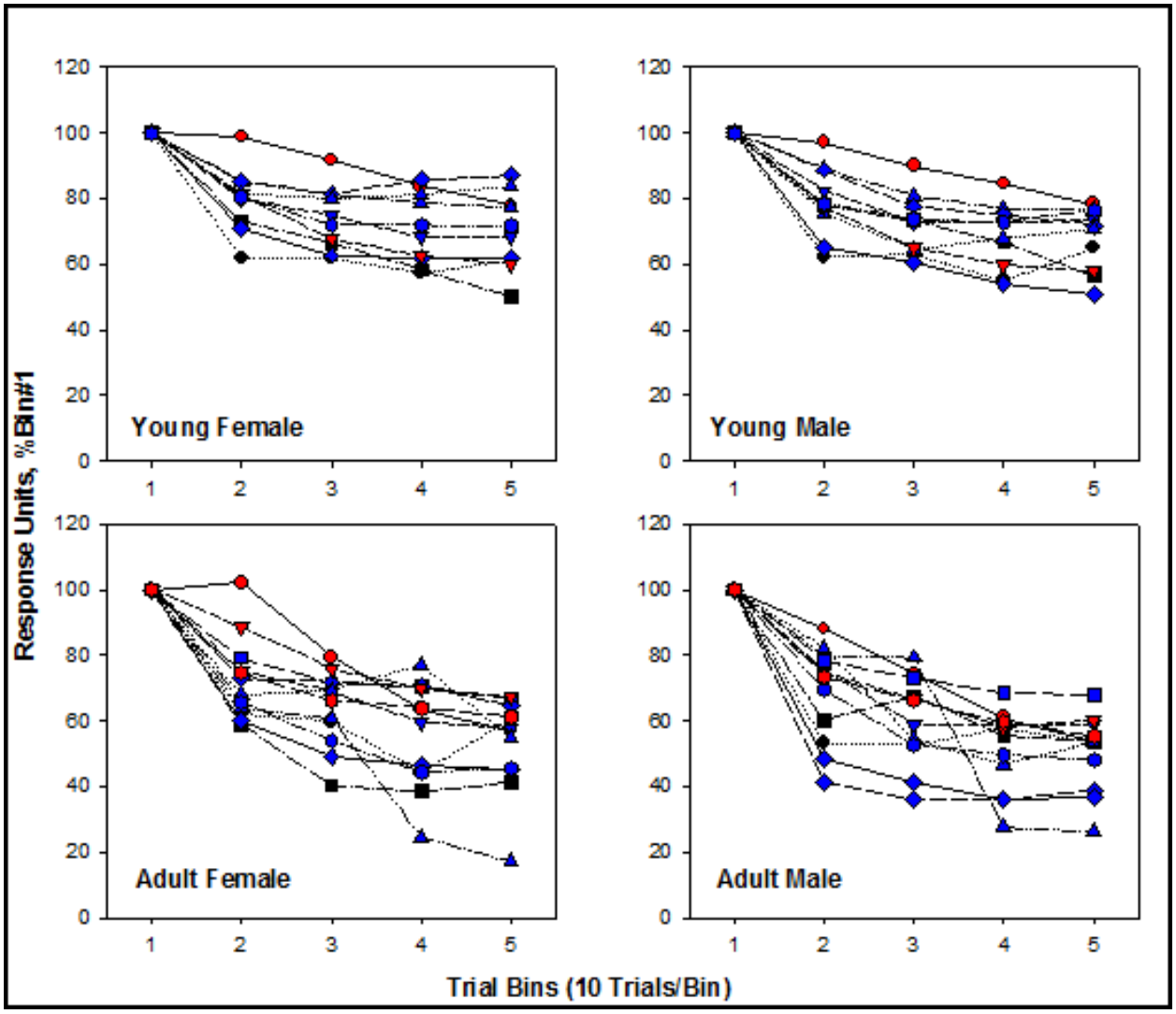


Figure 2: Habituation data for amplitude of young (PND20-24) and adult (PND58-68) male and female rats. These data were submitted from a number of different DNT testing laboratories (not identified individually). Data are group mean responses for controls plotted as a function of 5 10-trial blocks within a test session (data from Raffaele et al., 2008).

Averages of the response magnitude and latency values should be calculated from individual trial data. Both the averages for blocks of 10 trials (average for trial block) as well as the total session average (average for all 50 trials) should be reported. These averages should include treatment and sex factors. A measure of habituation must be reported. While habituation can be assessed in a variety of ways, the main characteristic of habituation is a decrease in a response following repeated exposure to a stimulus that does not involve sensory adaptation/fatigue or motor fatigue (Leussis and Bolivar 2006; Grissom and Bhatnagar 2009; Thompson and Spencer 1966; Thompson 2009). The precise method used to determine habituation should be clearly indicated. Note that the initial trial typically has the highest startle response, and when averaging across blocks this high value contributes to the higher values for the first block. Typically, habituation is inferred from statistical analyses that demonstrate a significant effect of trial block that is due to a decrease in startle response amplitude between the beginning and end of the test session. Figure 2 illustrates control habituation data from multiple laboratories for males at both young and adult ages (Raffaele et al., 2008). Habituation may also be inferred from slope measurements (negative slope for response magnitude from first to last trials) or difference in scores between first and last trial blocks. Regardless, the specific method used to determine habituation must be clearly described.

6.2. Reported Data

In order to evaluate treatment effects, the following level of detail should be reported:

- Startle measures broken down by dose group, sex and trial blocks at each test age. Note that the US EPA DNT guidelines and the OECD 433 require that data are reported for 5 blocks of 10 trials/block for a total of 50 startle trials;
- All dependent variables reported as means ± standard deviations for each treatment cell (trial block, treatment group and sex combination). Additional statistical measures, such as the median or mode, may also be presented where appropriate (e.g., unequal variance, non-normal or skewed distributions);
- Total session amplitude and latency (i.e., collapsed across all trials) reported as mean ± standard deviations;
- Details on calculations or analysis for habituation measures clearly described.

7. Data Analysis

Startle data must be statistically analyzed using methods appropriate for repeated-measures data. Statistical analyses specific for DNT data are described by Holson and colleagues (2008). As with all developmental toxicity data, the unit of analysis must be the litter. Where a male and female are sampled from the same litter and data from both are included in data analysis, the data must be analyzed with sex nested within the litter, using litter as a random factor, or a hierarchical analysis with sex as a matching factor. Data analysis should use a repeated measures analysis of variance (ANOVA) (or comparable procedure) that incorporates treatment and trial blocks in the analysis. Separate ANOVAs on results for each sex separately can normally only be justified after finding a significant sex-by-treatment interaction in an overall ANOVA.

Using ANOVA, demonstrating habituation would usually require the detection of a significant trials (or trial block) effect that is associated with decreased startle response magnitude over testing. Other analytical approaches that demonstrate a significant decrease in response magnitude (e.g., regression approaches) can also be employed. It is not appropriate to perform separate comparisons among treatment groups at different trial blocks in the absence of a statistically significant interaction between time and treatment.

While there are a number of statistical models for analysis of startle data, repeated-measures ANOVA is the most generally accepted approach (Holson et al., 2008). With mixed-effects-modeling, a covariance matrix that best fits the data can be selected from many available types of covariance matrices. For analysis of results within a specific age group, factors include dose and trials as the repeated-measure factor, keeping in mind that litter must be included where there is more than one animal from each litter. Generally, interactions should be examined first, as significant interactions indicate that main effects are modified by a second factor. Potential interactions would include treatment-by-trial blocks, trial block-by-sex, treatment-by-sex, and treatment-by-trial block-by-sex. Depending on the nature of the interactions, main effects may also be examined. More specifically, crossed interactions usually preclude the use of main effects while for parallel interactions main effects can still provide valuable information.

The impact of body weight on startle response magnitude has been clearly established. For example, it has been repeatedly demonstrated in studies of the ontogeny of the response where body weight gain increased the amplitude of the response (Dean et al., 1990). In addition, where male body weight exceeds female body weight in mature animals there is a corresponding increase in the amplitude of the startle response. The use of body weight as a covariate in analyzing startle response data can be useful to reduce error associated with body weight. However, the use of body weight as a covariate for startle response amplitude data is appropriate only when there are no treatment-induced effects on body weight (Csomor, 2008; Holson et al., 2008; Tabachnick and Fidell, 2006). On the other hand, failing to control for body weight could result in erroneous conclusions, especially regarding sex differences. There are many examples from the literature where a simple correlation between a chemical's effect on body weight and the startle response may be independent of treatment effects on startle responses. For example, developmental exposure to methimazole results in a small decrease in body weight in male animals but no corresponding effects on startle response. At the same time female body weight is unchanged but the startle response magnitude in females is decreased (Henck et al., 1996). There are also study results where no changes in startle response magnitude occur at young or old test ages despite decreased developmental weight gain (Buelke-Sam et al., 1998). Similarly, in young animals, decreases as well as increases in startle response have been recorded in animals having concomitant treatment-related decreases in weight gain during development (Goldey et al., 1995; Kobayashi et al., 2005). Some have reported no effects on startle response after weight loss of up to 25% (Crofton and Knight, 1991). Figure 3 illustrates an example of the independence of weight loss and effects on startle response magnitude. The upper figure shows a dose-related decrease in PND22 female body weights in the absence of any effect on startle response amplitude. The lower graph shows the converse situation where a dose-related decrease in startle response magnitude occurs in the absence of any dose-related change in body weight. Taken together, even though both age-related changes in body weight and sex differences in body weight do influence startle response magnitude, the available data indicate that treatment-induced effects on body weight are not necessarily associated with treatment-

induced changes in startle response magnitude. In the end it is best to examine both body weight and startle data and determine the most appropriate or valid approach for analysis.

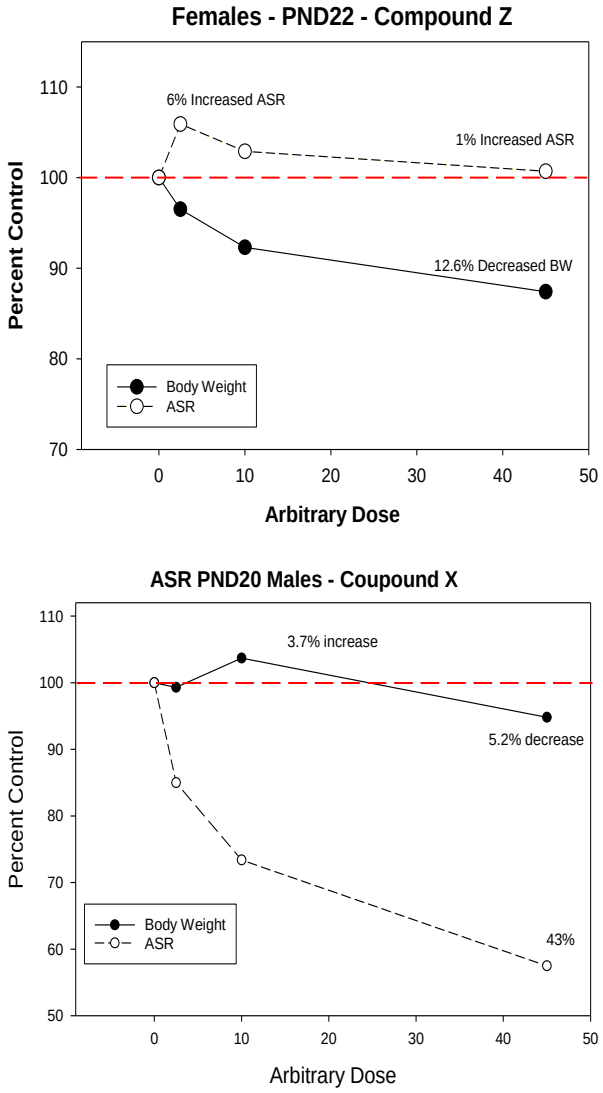


Figure 3. Example of independence of treatment-induced changes in body weight and startle response amplitude (see text for details).

7.1. Statistical Models

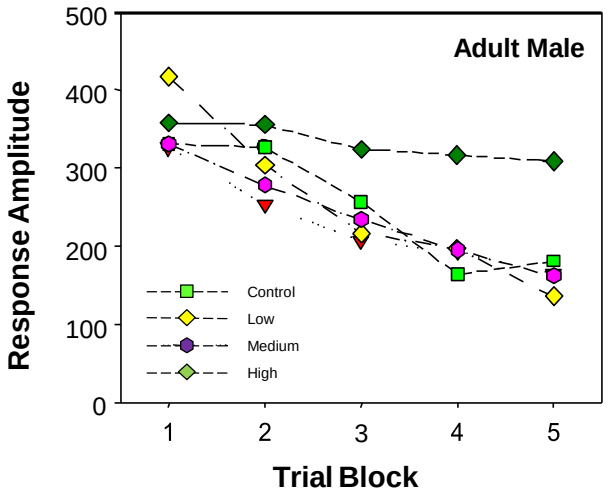
Below is a brief description of the types of statistical effects that may be obtained from analyses of DNT startle data.

- Main effects:
  - Treatment: indicates that chemical exposure affects the overall level of startle amplitude (or latency) collapsed across all other factors (i.e., sex,

trial blocks, litter). This is effectively a direct statistical test of the effect of treatment on the overall total session startle since trial blocks and sex are ignored in this statistical test. A significant treatment effect indicates that the exposure affects total session startle amplitude (or latency);

- Sex: indicates a difference in startle response between males and females while all other factors (treatment, trial block, litter) are held constant;
- Intersession Trial block: indicates that startle response varies across trial blocks in the test session. A significant effect of trial block is required to demonstrate habituation. Note that a significant effect of trial is necessary but not sufficient to demonstrate habituation. A significant effect of trial may indicate an increase in startle responses, or some other non-monotonic pattern, over trial blocks, while habituation requires a decrease in startle responses over trial blocks.

Figure 4. Example of treatment-related alteration in startle habituation. Adult males were tested for 50 trials and data are presented as means of five 10-trial blocks (see text for details).



- Interactions:
  - Treatment-by-Trial blocks: indicates that the effect of trial blocks varies among treatment groups (or alternately, the effect of treatment varies between trial blocks). The precise nature of the interaction must be evaluated to determine how treatment groups differ over trial blocks since it is possible to have a significant trial-by-treatment interaction that does not indicate habituation (e.g., treatment increases startle responses over trials in a test session but the size of the increase varies between treatment groups). This is illustrated in Figure 4 where there is a similar decrease in startle amplitude over the five trial blocks in all groups except the high dose group, which shows much less habituation. This was evidenced statistically as a trial block-by-treatment interaction. Acceptable approaches to assess the interaction include simple effects tests or specific contrasts, at each trial block. Depending on the data, another simpler approach is to treat trial block as a continuous variable (not as a categorical variable) in the

model, and the habituation is described by the slope (coefficient of trial block). A significant treatment-by-trial block interaction indicates the slopes were different among the dose groups;

- Treatment-by-sex: indicates that the effect of treatment differs between sexes. Note that the nature of the interaction must be evaluated, usually with simple effects or specific contrasts. It is possible to have a significant treatment-by-sex interaction that reflects that males and females are both significantly affected by the treatment in the same direction but the magnitude of the change differs between males and females;
- Sex-by-Trial blocks: indicates a difference in habituation between males and females;
- Treatment-by-Trial blocks-by-sex: indicates that the interaction between sex and treatment varies over trial blocks (or, alternately, that the effect of treatment varies between males and females over trial blocks). This interaction is normally required to demonstrate that the effect of treatment on habituation differs between sexes. As with other interactions, the nature of the interaction must be evaluated using mean contrast interaction tests, simple effects, or specific contrast. It is necessary to show that the change in startle response over the test session is influenced by treatment and that this change in startle response is altered differently between sexes.

In addition to ANOVA results described above, the actual magnitude of change should be considered since large effects are sometimes not statistically significant due to high variability but could be of substantial biological significance (also see discussion of biological significance in section 8.4). This is especially useful where critical significance values (e.g., p-values) are inconsistent (e.g., results between age groups) due to variability in data, while the magnitude of effects may be comparable. Lastly, actual significance values of statistical tests should be reported rather than critical p-values like $p < 0.05$ since all modern statistical software programs report actual p-values. Reporting results of statistical testing using only critical p-values such as $p < 0.05$ can result in qualitatively different interpretation of test results where, for instance, $p=0.048$ and $p=0.051$ when there is no solid basis for treating these two probability values as substantially different (Holson et al., 2008).

A number of relatively common issues occur in the data analysis of DNT reports that can limit the interpretation of study results and may even produce misleading results. These include:

- Sex is often not included in the overall data analysis, with results for males and females analyzed separately. This approach not only precludes direct testing of sex effects and treatment by sex interactions, it also reduces statistical power to detect treatment effects;
- Trial block is not included in data analysis with results reported only for total session startle response (i.e., collapsed across all trials). Not only does this approach preclude any possibility of detecting treatment effects on habituation, it may mask treatment effects where the effect of treatments varies over time blocks;
- Using body weight as a covariate in data analysis when there is a treatment-related effect on body weight (see discussion in section 7). Using body weight as a covariate is acceptable when there are no treatment effects on body weight.

8. Interpretation of Startle Response Results

8.1. Properties of startle control data

Regardless of the system used to measure the auditory startle response, there are a number of behavioral properties in the measured data that should be evident and should serve as a reference point for evaluating reported results. Differences in the detection method (load cell vs accelerometer) must be understood (see section 4) as they influence the use of control data and the comparability of results between ages. Properties of startle measurement that should be present in control animals include:

- In adults, there is normally a sex-dependent effect on response amplitude with adult males exhibiting larger startle response amplitude than adult females due to larger body and muscle mass;
- Habituation of startle amplitude should be evident in animals tested at or after weaning;
- Control animals should demonstrate age-appropriate amplitude and latency measurements.

The reviewer is directed to two papers that contain a wealth of exemplary data on control startle data from regulatory studies (Raffaele et al., 2008, Tyl et al., 2008). Selected examples from these two papers and the peer-reviewed literature are presented herein.

Table 3 presents data that illustrate some comparisons that are expected in submitted data with respect to age, sex, and body weight. Note that systems calibrated to an absolute standard (e.g., grams, Newtons) should show increased startle response between young and adult, and between adult male and female animals; this may not be the case when the output is in units that are not absolute, or may not be consistent across ages. Assuming that the testing system is calibrated to an absolute standard (e.g., grams, Newtons), then there should be an apparent increase in the startle response between young, post-weanling animals compared to adults. This can be seen in Table 3 where males and females were tested at two ages (PND23) and (PND61) and are due to the significant weight gain that occurs between these two ages (note that in some strains body weight can increase 4 or 5-fold). Second, there are no significant differences in response amplitude of PND23 males and females. This is because there is less impact of sex on body weight or muscle mass at that age, compared to mature animals. Third, there is smaller response amplitude in adult females compared to adult males. In this study, females at PND61 weighed about 20% less than males. Fourth, habituation occurred in all groups. This is clearly evident by a 30-40% decrease from the first block of trials to the last. The statistics presented in the report support this position with a main effect on trial block.

Table 3. Example of auditory startle response amplitude data, from PND23 and PND61 male and female rats. See text for details. (modified from Tyl et al., 2008)

Testing Age	Trial Block	Dose (as multiples of the lowest dose) 0	1	5	25
		Males			
PND23	1-10	379.6 ± 98.8	439.3 ± 149.2	372.9 ± 99.2	302.5 ± 122.8
	11-20	266.6 ± 64.6	297.7 ± 63.3	297.4 ± 103.2	200.3 ± 64.4*
	21-30	233.3 ± 63.1	261.3 ± 53.3	265.9 ± 91.2	176.7 ± 48.0
	31-40	223.7 ± 50.9	235.4 ± 47.0	251.0 ± 72.6	144.6 ± 54.1**
	41-50	208.0 ± 55.4	220.7 ± 45.7	221.6 ± 62.8	148.0 ± 51.1**
PND61	1-10	1351 ± 370.6	1692 ± 563.1	1394 ± 525.3	1305 ± 393.7
	11-20	898.6 ± 258.7	1053 ± 280.6	952.8 ± 367.3	871.5 ± 228.9
	21-30	861.2 + 346.1	960.0 ± 336.2	846.6 ± 350.7	878.1 ± 324.9
	31-40	769.7 + 279.9	963.1 ± 350.3	832.4 ± 332.1	758.3 ± 229.3
	41-50	701.9 ± 307.9	929.6 ± 368.5	859.5 ± 356.0	752.9 + 270.6
		Females			
PND23	1-10	374.9 ± 121.3	406.3 ± 177.0	353.2 ± 108.2	297.4 ± 62.7
	11-20	332.9 ± 143.1	306.7 + 130.3	281.0 ± 122.5	232.8 ± 45.9
	21-30	267.5 ± 67.0	294.3 ± 174.7	252.9 ± 165.0	222.7 ± 55.0
	31-40	234.7 ± 58.3	230.1 ± 128.6	190.0 ± 59.2	202.4 ± 53.7
	41-50	226.3 ± 44.1	249.0 + 140.1	199.1 ± 72.7	189.9 ± 55.6
PND61	1-10	1064 ± 225.1	974.4 ± 222.8	1161 ± 442.2	1237 ± 321.2
	11-20	874.2 ± 170.2	913.2 ± 277.0	1025 ± 441.5	963.5 ± 156.5
	21-30	772.0 ± 234.1	831.4 ± 305.2	908.7 ± 588.7	926.2 ± 203.1
	31-40	677.8 ± 270.4	813.2 ± 442.1	816.2 ± 505.0	832.1 ± 315.4
	41-50	652.6 ± 185.4	633.5 ± 378.1	716.1 ± 380.7	798.6 ± 220.6

[a] N = 10-14/group. Mean data ± SD. *Statistically significantly different from control value at $p < 0.05$. **Statistically significantly different from control value at $p < 0.01$.

Table 4 illustrates control data that have some inconsistencies with the known biology of the startle response in terms of age and sex. These data were generated using an accelerometer with output listed as millivolts for the amplitude data (Vmax), milliseconds for the latency data (Tmax), and average amplitude in millivolts (Vave). There is consistency in the fact that the amplitude values at young ages have no apparent sex differences, and that amplitude values for older females are lower than values in older males. In addition, the latency data are mostly biologically coherent, in that the peak amplitude occurs at 22 to 34 msec. A number of issues in this dataset, however, are apparent. The response amplitude for PND20 males is slightly higher than the PND60 males and the amplitude of the PND20 females is more than double the response for PND60 females. Another issue is that the lowest amplitude responses in PND60 females have longer latencies. Normally larger responses have longer latencies to the peak response, although there are many reasons this may not occur. Additionally, there is no intra-session information provided to determine if habituation occurred. Since both the EPA and OECD 433 guidelines require habituation data, this report could be considered as not meeting this requirement.

Table 4. Example of auditory startle response data from PND20 and PND60 male and female rats (modified from Tyl et al., 2008).

	Dose Group			
	Control	Low	Medium	High
MALES				
PND20				
Vmax (mv)	214.3 ± 94.9	182.7 ± 84.6	157.3 ± 61.7	123.4 ± 57.3
Tmax (msec)	24.9 ± 3.6	24.1 ± 3.4	25.4 ± 3.5	24.3 ± 3.1
Vave (mv)	45.5 ± 18.6	39.0 ± 16.9	33.8 ± 11.0	26.6 ± 11.6*
PND60				
Vmax (mv)	209.9 ± 129.0	189.2 ± 76.0	126.0 ± 81.5	98.6 ± 66.6*
Tmax (msec)	31.5 ± 4.8	29.3 ± 5.5	32.9 ± 6.3	33.6 ± 5.3
Vave (mv)	46.5 ± 28.0	40.3 ± 16.6	29.0 ± 19.2	21.6 ± 14.4*
FEMALES				
PND20				
Vmax (mv)	181.2 ± 50.8	182.5 ± 54.1	160.6 ± 47.9	129.0 ± 56.6
Tmax (msec)	22.7 ± 2.6	24.0 ± 2.6	24.6 ± 4.6	25.3 ± 2.0
Vave (mv)	39.5 ± 10.0	37.9 ± 10.5	34.4 ± 10.9	27.4 ± 12.5
PND60				
Vmax (mv)	78.2 ± 36.1	78.7 ± 30.7	79.6 ± 40.4	80.1 ± 37.0
Tmax (msec)	34.3 ± 4.4	32.9 ± 4.7	33.9 ± 4.5	30.9 ± 4.7
Vave (mv)	16.7 ± 7.6	17.3 ± 7.8	16.2 ± 7.0	16.5 ± 6.7

[a] N = 9-10/group. Data are mean ± SD

* Statistically significantly different from control group value, $p < 0.05$.

The issues noted in Table 4 should lead to a more thorough review of the report itself. The first step should be a careful examination of the testing equipment, calibration of noise stimuli, and calibration of the testing platform. It is possible that the equipment was properly calibrated and that the gain (circuit amplification) was increased at the young age. Some equipment allows this to be done via alterations to the sensitivity settings or gain switches when testing young (and smaller) animals. If the system was properly calibrated and the sensitivity was increased in the young animals, then this would explain that the lower startle response on PND60 actually reflects the change in instrument sensitivity setting rather than an actual decrease in startle response amplitude. Historical control data from the conducting laboratory would also be useful in determining whether these data are consistent with previous use of this equipment. Without such an explanation, it would be appropriate to question the adequacy and reliability of the control data, thus obviating any need to look at possible treatment-related results.

8.2. Variability in auditory startle data

As with all biological measures, there is variability inherent in the auditory startle response. As mentioned above, separation of experimental and inherent variance can be difficult (Cohen, 1977). There is an excellent review of the variability in auditory startle data from an ILSI workshop (Raffaele et al., 2008). The findings of that report indicated that regulatory testing

laboratories had coefficients of variation (CVs; standard deviation as a percent of the mean) that varied from 20-110% and that some laboratories are much better than others in controlling experimental variability. Where high levels of variability are reported, checking raw data can help determine if treatment effects are masked by odd or unusual individual animal scores that inflate variability and thus decrease power to detect treatment effects. Another reason for high variability is that the data may not be normally distributed, and a transformation may be needed before analyzing the data. Review of the laboratory's historical control data (see section 8.1) will also help in assessing the variability.

8.3. Positive Controls

The EPA DNT Guideline requires testing laboratories to provide positive control data that demonstrate the ability of the test equipment to detect both increases and decreases in response amplitude. Positive control data are crucial for evaluating the relative proficiency of a testing laboratory in detecting chemical-induced changes in the measured endpoint, and serve as reference standards for the measured endpoint. Positive control data help demonstrate the dynamic range of a biological response, verify the proficiency of laboratory personnel in a realistic environment, and can help to characterize intra-laboratory and inter-laboratory reliability of a test method. Positive control data are valuable in a weight-of-evidence approach to help determine the biological significance of results and provide confidence in negative results. A number of reviews of the use of positive controls in DNT studies are available (Crofton et al., 2004, 2008; Maurissen and Marable, 2005).

The methods used for positive controls should be completely described, and must be the same as those used in the study being evaluated. For example, equipment used and strains and numbers of animals tested should be the same. Sessions should contain the same number of trials and use the same stimuli. Stimulus test parameters (stimulus intensity, duration, inter-trial intervals, etc.) should be similar. The testing laboratory should use animals exposed in a manner that mirrors the DNT protocol, although age-matched animals that are exposed just prior to the time of testing are also acceptable (Crofton et al., 2008). The positive control data should have been collected within a reasonable time frame before the data collection for the DNT study being evaluated (e.g., within the last few years). New positive control data should also be collected when personnel or other critical laboratory elements change. This new positive control data, while more important for interpretation of treatment-related effects, can also be used to determine whether a shift has occurred in the control data collected by the testing laboratory. In addition, the reviewer should look to see if the same patterns are evident between sexes and ages.

During assessment, care should be taken to ensure that the positive control treatment is effective. In a review of positive control data submitted to the US EPA, 3 of 16 laboratories never submitted positive controls for startle testing, 4 laboratories failed to demonstrate proficiency (i.e., lack of effects of positive controls), and 3 laboratories were deemed to have only partially established proficiency (e.g., lack of statistical significance, lack of appropriate age animals). Only 5 of 16 testing laboratories were judged to be proficient based on sufficient positive control data (Crofton et al., 2004).

The lack of adequate positive control data can result in low confidence in negative data which in turn increases the possibility that negative results may in fact be a false negative finding. A

simple example illustrates the importance of positive controls for data interpretation. In this particular case, the test laboratory submitted startle data from males and females tested on PND23 and PND62. There were no effects reported for any aspects of startle testing in this study. There was evidence of habituation, but only in the adult age. No historical control or positive control data were available for PND23. The positive control data from this laboratory were available for female adult rats only and the treatment was without effect. Consequently, it was not clear whether the test laboratory's procedures were effective in detecting any treatment-induced changes in startle responses. In a case like this the deficiency should be noted in the positive control data.

8.4. Statistical versus biological significance

Relying solely on statistical differences or a fixed percent change in response (e.g., 20-25% change) is not always appropriate because of the inability to differentiate inherent variability from experimental error (Cohen, 1977). Excessively large variances in data, regardless of the source, make detection and interpretation of treatment-related findings difficult. In case of experimental-error induced variability, the use of statistics increases the rate of false negatives. Large variance due to inherent variability and the use of a fixed change in the response will also increase false negatives. Although increased variance in treated groups may affect statistical power to detect significant differences (Muller and Benignus, 1992), it may also be an important indicator of a treatment effect (Cory-Slechta et al., 2001).

8.5. Dose-response results

The last 30+ years of studies of effects of chemicals on the auditory startle response have shown that chemicals are capable of both increasing and decreasing the amplitude response, although decreases are more commonly found (Crofton, 1992). The classic pharmaceutical principle of increasing dose and increasing effect is a clear indication of an adverse effect of chemical exposures. These kinds of effects are easy to interpret as effects of exposure.

More difficult to interpret are non-monotonic effects where there are small changes at the low or mid dose, with no changes at the high dose. Often these occur at one age or in one sex only, or occur only with some but not all trial blocks in the session. A careful look at all the startle data in the study can help determine if effects may be a false positive.

Figure 5. Startle data from adult males and females. The submitted assessment revealed a significant increase in mean session startle response amplitude at the mid dose only and in males only (upper panel). Re-analysis of the data using a model that contained sex nested within the litter revealed a lack of statistical significance of treatment on any variable (see text for details).

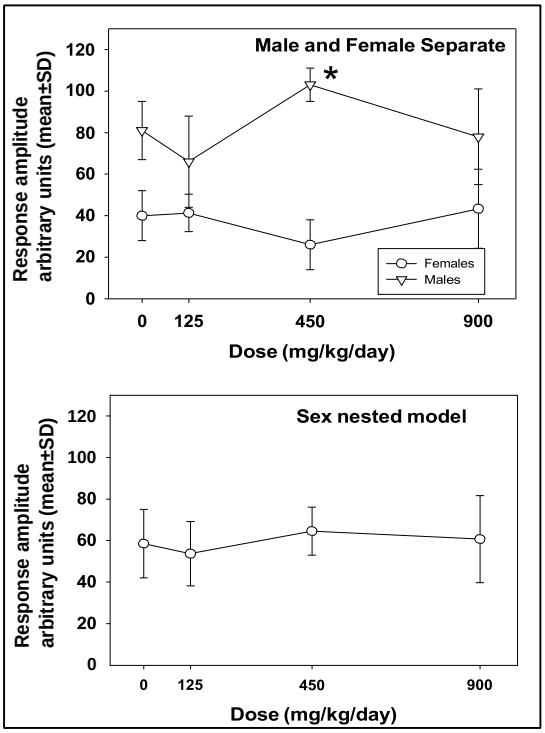


Figure 5 illustrates this example from a DNT study that had the following findings: 1) control data revealed similar amplitudes in PND23 males and females, increased amplitudes in adults compared to PND23 animals, and smaller amplitudes in adult females compared to males; 2) habituation was present in both the young and adult animals, and 3) CVs in the control data ranged from 18 – 33%. The data fit the biological background for the startle response and the variance was well within an acceptable range (see Raffaele et al., 2008). There were no significant effects of treatment on any variables in PND23 males or females. In adults there was no treatment-related effect on habituation. However, there was a main effect of treatment in the males, with a significant increase in the group mean of the middle dose group only. Review of the study methods revealed that one male and one female were tested from each litter for all dose groups, but the sexes were analyzed separately. Re-analysis of the data using a model that accounted for sex within the litter revealed a lack of statistical significance of treatment on any variable. A visual inspection of the data plotted in Figure 5 reveals why. The amplitude in the mid dose male group was approximately 19% higher than the control group and had a very low CV of only 8%. Females at this same dose group had slightly smaller amplitude relative to controls. Including sex in the statistical model to directly test the effect of sex revealed a lack of any sex-by-treatment interaction and presentation of data from both sexes combined clearly reveals no effect of treatment. A review of the positive control data from this laboratory demonstrated that the test device was capable of detecting a decrease in amplitude (using clonidine), and an increase in amplitude (using DDT). These data are consistent with other positive control data for startle tests (Crofton et al., 2008).

8.6. Treatment-related Habituation results

Both the US EPA and OECD 433 guidelines suggest assessing auditory startle response habituation. For auditory startle testing, habituation is demonstrated by a reduction in startle amplitude over the course of repeated presentation of the startle stimulus. Like other measures of habituation, it reflects non-associative learning that is fundamental for interactions with the environment. Under normal circumstances, habituation of startle response amplitude may be demonstrated in several ways (see section 6.1). A difference in the rate of habituation between young and adult animals should be evident in control data. It is also important to note that, unlike response amplitude, startle response latencies do not normally show habituation.

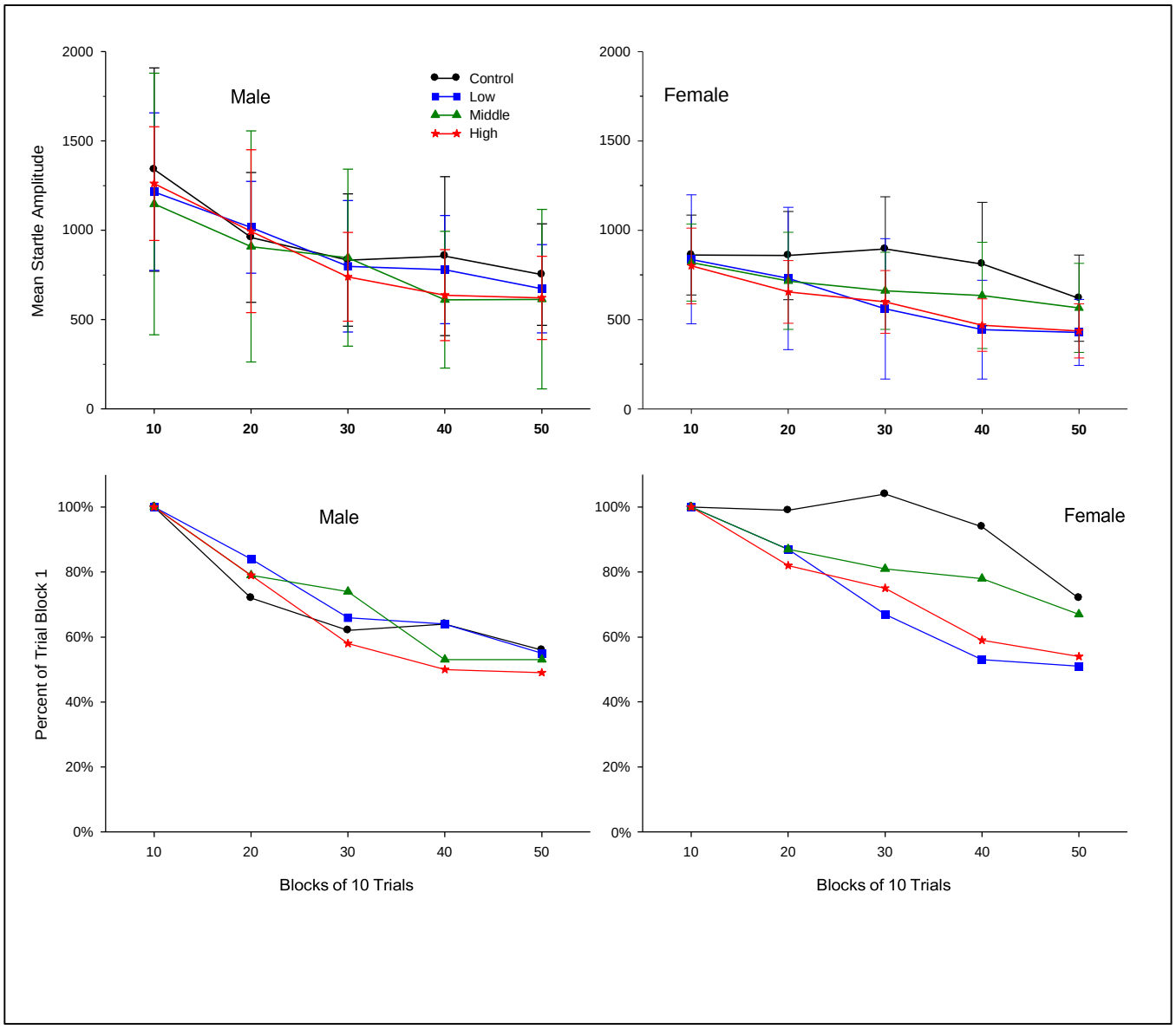
An example of startle habituation data that present challenges for interpretation is illustrated in Table 5 and also plotted in Figure 6 (Tyl et al., 2008). Male and female offspring were tested on both PND22 and PND62. There were no treatment-related effects at PND22 in males or females. Analyses of data from PND62 revealed an interaction of treatment and trial block for females only, with significantly decreased amplitudes (up to 45% in the low and high doses, but not in the mid dose group) and only in the 21-30 and 31-40 trial blocks. While these data appear to indicate a treatment-related *increase* in habituation, there are two major challenges with this interpretation. The first is that the effects are not dose-related since there are no significant effects in the mid dose. Non-monotonic data of this sort should be identified for closer scrutiny. The second problem in this data set is the seemingly abnormal habituation in the control females. Each of the three treatment groups show the expected decrease in amplitude over the 5 trial blocks (up to that equal to ~40-50%), while the female control groups show a decrease only at the last time block. In the fourth block (31-40) there is only a 6% decrease relative to Block 1-10. A careful review of positive control and historical control data revealed this laboratory has repeatedly demonstrated habituation in adult female rats that is in the range of 40-50%. Thus the data in the low and high dose groups demonstrate a more normal amount of habituation, rather than a treatment-related response. These data suggest that the effects found in only a limited number of trial blocks in a non-dose dependent manner, coupled with the lack of habituation in the control group, likely represent a false positive finding.

Table 5: Adult male and female rat auditory startle amplitude data from a DNT study with chemical X. Numbers are group mean ±SD. Numbers in brackets are amplitudes as percent of each dose group Trial Block 1-10.

Trial Block	Control	Low dose	Mid dose	High dose
Males				
1-10	1338.6 ± 569.5 [100]	1214.8 ± 440.3 [100]	1146.0 ± 732.8 [100]	1261.0 ± 318.2 [100]
11-20	959.1 ± 363.1 [72]	1015.6 ± 256.6 [84]	907.6 ± 646.8 [79]	993.5 ± 455.7 [79]
21-30	832.7 ± 370.0 [62]	797.4 ± 368.6 [66]	846.0 ± 495.4 [74]	738.2 ± 248.5 [58]
31-40	854.3 ± 444.9 [64]	778.9 ± 303.6 [64]	610.1 ± 383.6 [53]	636.1 ± 254.5 [50]
41-50	751.0 ± 284.4 [56]	671.6 ± 247.5 [55]	613.3 ± 503.1 [53]	620.9 ± 232.6 [49]
Females				
1-10	860.9 ± 224.1 [100]	837.1 ± 360.9 [100]	819.5 ± 215.4 [100]	800.2 ± 212.1 [100]
11-20	858.2 ± 247.3 [99]	730.8 ± 398.0 [87]	715.7 ± 271.9 [87]	654.7 ± 175.3 [82]
21-30	894.4 ± 292.6 [104]	559.4** ± 392.2 [67]	661.1 ± 215.7 [81]	598.8* ± 175.8 [75]
31-40	810.9 ± 345.7 [94]	443.2** ± 276.7 [53]	635.0 ± 297.0 [78]	469.0** ± 146.2 [59]
41-50	619.8 ± 242.1 [72]	428.2* ± 185.1 [51]	564.9 ± 249.7 [67]	435.8 ± 151.9 [54]

*Statistically significantly different from control value at $p < 0.05$.
**Statistically significantly different from control value at $p < 0.01$

Figure 6. Plotted values from Table 5 Adult male and female rat auditory startle amplitude data from a DNT study with chemical X.



9. General Interpretation

The auditory startle response in rodents is a simple behavior controlled by a simple neural circuit involving sensory, neural and motor systems (described above). Because of this, treatment-induced changes in the startle response can usually be ascribed to alterations in one or more of these systems and this reflects a treatment-induced disruption in nervous system functioning even if the specific location or nature of the alteration is not clearly identified. For instance, alterations in sensory thresholds or alterations in neuromuscular functions may both impact auditory startle responding. Treatment-induced alterations in habituation can usually be interpreted as treatment-related effects on non-associative learning that is mediated by the central nervous system. In the interpretation of auditory startle data some issues that should be considered are described below.

- In evaluating startle test results the reviewer should verify that the testing procedures and the data reported meets the requirements of DNT test guidelines. Animals were tested at appropriate ages, both sexes of animals were tested, testing was counterbalanced among treatment groups and sexes, animals were selected from all litters, etc. If the study design, animal selection, or test conduct does not comply with DNT test guidelines, the startle test data results may be inadequate. If inadequate information is provided to assess this, then additional information should be obtained from the reporting laboratory;
- It is important that the study report provides adequate details on the methodology and test equipment used and on the software parameters used for data collection and analysis. For example, reports should clearly identify if equipment is reporting the first response, or the largest response, which are not necessarily the same. Similarly, reports that use the average response reported by some systems may be the average voltage over the entire recording interval. These results most likely will differ considerably from results from other laboratories that use actual startle response peak values;
- The study report should clearly report the actual measures that were obtained (i.e., peak response, peak first response, average response, latency to response onset, latency to response peak). The report should clearly indicate how habituation metrics were calculated (e.g., compare first block and last block of trials). Reporting only the mean response collapsed over all trials precludes determining habituation and may also mask any trial-by-treatment interactions;
- Appropriate statistical analysis for the reported data is essential. Separate analysis of startle data for males and females is not appropriate and precludes determining if sex differences in treatment effects occur. Litter must be taken into account in all data analysis. Comparison between ages can only be performed if equipment used allows direct comparison between ages;
- Assessing positive control data, historical control and concurrent negative control data is essential to determine if startle response data is within normal and acceptable ranges. Where control animal data are abnormal, startle study data may not be usable. It is important that control data reported are reasonable and consistent with reports from the same laboratory in the past. In addition, both the startle response magnitude and latency values in the study report should be consistent with historical control data from the same laboratory using the same equipment. Such historical control data should also include measures of habituation and demonstrate age- and sex-related startle response. Concurrent control should exhibit

normal habituation response pattern (this should be compared against historical controls as well), normal sex differences (males exhibiting larger response than female in adulthood), and reasonable latency values (> 10 msec for onset, approximately 20-40 msec for peak response). Study reports that have latency values that deviate substantially from these may be problematic or may reflect inappropriate data selection (wrong software settings) or analysis.

Table 6. Potential interpretation of some of the most common problems for startle results found in DNT study reports.

	Data Problem	**Potential interpretations**
1	Lack of age dependent increase in control responses (e.g., PND21 vs PND60)	If not explained by equipment issues (e.g., sensitivity gain changes between age groups), then treatment-related effects, or lack thereof, may not be interpretable
2	Lack of sex dependent differences in control responses of adult rats	If not explained by equipment issues (e.g., sensitivity gain changes), then treatment-related effects, or lack thereof, may not be interpretable
3	Lack of evidence of habituation at young age (around weaning)	Animals may have been tested at too young an age, check historical and positive control data for habituation at that specific age. If age is not the cause see #4 below
4	Lack of evidence of habituation at adult age	Lack of habituation suggests improper experimental conduct or data analysis. Verify that data were analyzed so that habituation was tested directly. If data were analyzed appropriately then the data should not be used for decisions based on habituation. Main effects of treatment may still be useful.
5	Significant effects on latency with no impact on amplitude	Since effects on latency without changes in amplitude are extremely rare and may not be biologically reasonable, efforts should be taken to ensure that testing equipment and data algorithms are correctly identifying response peak times.

6	Significant effects on startle amplitude in only one sex only	Lack of effects in both sexes is not a valid reason for dismissing treatment-related effects. Chemicals that interfere with some endocrine systems during development may have sex-selective effects.
7	Significant effects of treatment only at the low or middle dose in one sex.	If sex is not included in the model, then reanalyze data including sex. If data re-analyses reveal no sex-by-treatment interaction and no main effect of treatment, then the reported effect was likely a false positive.
8	Dose-related effects of increasing magnitude with a lack of statistical significance coupled with excessive variation	Large effects on the amplitude of the startle response may be masked by excessive variability. Expert judgment is required to determine the biological significance in light of the lack of statistical significance. Check historical and positive control data. Check to ensure large variability is not due to extreme values for few test animals.

10. References

Adams J (1986) Methods in Behavioral Teratology. In: Riley EP, Vorhees CV (eds) *Handbook of Behavioral Teratology* Plenum Press, New York, pp 67-97.

Adams J, Buelke-Sam J, Kimmel CA, Nelson CJ, Reiter LW, Sobotka TJ, Tilson HA, and Nelson BK. 1985a. Collaborative Behavioral Teratology Study: protocol design and testing procedures. *Neurobehav. Toxicol. Teratol.* **7**, 579-586.

Adams J, Oglesby DM, Ozemek HS, Rath J, Kimmel CA, and Buelke-Sam J (1985b) Collaborative Behavioral Teratology Study: programmed data entry and automated test systems. *Neurobehav. Toxicol. Teratol.* **7**, 547-554.

Brunjes PC and Alberts JR (1981) Early auditory and visual function in normal and hyperthyroid rats. *Behav. Neural. Biol.* **31,** 393-412.

Buelke-Sam J, Cohen IR, Wierda D, Griffey KI, Fisher LF, and Francis PC (1998) The selective estrogen receptor modulator, raloxifene: a segment II/III delivery study in rats. *Reprod. Toxicol.* **12**, 271-288.

Claassen V (1994) *Neglected factors in pharmacology and neuroscience research: Biopharmaceutics, animal characteristics, maintenance, testing conditions.* Vol 12. Elsever, Amsterdam, Netherlands, 1994.

Cohen J (1977) Statistics for the Behavioral Sciences. Academic Press, New York, NY.

Cory-Slechta DA, Crofton KM, Foran JA, Ross JF, Sheets LP, Weiss B, and Mileson B. (2001) Methods to identify and characterize developmental neurotoxicity for human health risk assessment. I: behavioral effects. *Environ. Health Perspect.* **109 Suppl 1**, 79-91.

Crofton KM (1992) Reflex modification and the assessment of sensory dysfunction. In: Tilson HA, Mitchell CL (eds) *Neurotoxicology* Raven Press, New York.

Crofton KM, Foss JA, Hass U, Jensen K, Levin ED, and Parker SL (2008) Undertaking positive control studies as part of developmental neurotoxicity testing. *Neurobehav. Toxicol.* **30**, 266-287.

Crofton KM and Knight T (1991) Auditory deficits and motor dysfunction following iminodipropionitrile administration in the rat. *Neurotoxicol. Teratol.* **13**, 575-581.

Crofton KM, Lassiter TL, and Rebert CS (1994) Solvent-induced ototoxicity in rats: an atypical selective mid-frequency hearing deficit. *Hear. Res.* **80**, 25-30.

Crofton KM, Makris SL, Sette WF, Mendez E, and Raffaele KC (2004) A qualitative retrospective analysis of positive control data in developmental neurotoxicity studies. *Neurotoxicol. Teratol.* **26**, 345-352.

Crofton KM and Sheets L (1989) Sensory dysfunction: Assessment using reflex modification of the startle response. *J. Am. Coll. Toxicol.* **8**, 199-211.

Csomor, PA, Yee, BK, Vollenweider, FX, Feldon, J, Nicolet, T, and Quednow, BB. (2008) On the influence of baseline startle reactivity on the indexation of prepulse inhibition. Behav. Neurosci.**122**, 885-900

Davis M (1980) Neurochemical modulation of sensory-motor reactivity: acoustic and tactile startle reflexes. *Neurosci. Biobehav. Rev.* **4**, 241-263.

Davis M and Eaton RC (1984) The mammalian startle response. *Neural Mechanism of Startle Behavior* Plenum Press, New York, pp 287-351.

Davis M, File SE, Peek HVS, and Petrinovich L (1984) Intrinsic and extrinsic mechanisms of habituation and sensitization. *Habituation, Sensitization, and Behavior* Academic Press, New York.

Davis M, Gendelman DS, Tischler MD, and Gendelman PM (1982) A primary acoustic startle circuit: lesion and stimulation studies. *J. Neurosci.* **2**, 791-805.

Dean KF, Sheets LP, Crofton KM, and Reiter LW (1990) The effect of age and experience on inhibition of the acoustic startle response by gaps in background noise. *Psychobiol.* **18**, 89-95.

Fay RR (1988) *Hearing in Vertebrates: A Psychophysics Databook* HIll-Fay Associates.

Fechter LD, Young JS, and Annau Z (1986) Reflexive measures. *Neurobehavioral Toxicology* The Johns Hopkins University Press, Baltimore.

Geyer, MA and Swerdlow, NR (1988) Measurement of startle response, prepulse inhibition, and habituation. Current Protocols in Neuroscience, John Wiley & Sons, Supplement 3, 8.7.1-8.7.15

Grimsley CA, Longenecker RJ, Rosen MJ, Young JW, Grimsley JM, Galazyuk AV. (2015) An improved approach to separating startle data from noise. *J. Neurosci. Methods* **253**, 206-217

Grissom N and Bhatnagar S. (2009) Habituation to repeated stress: Get used to it. *Neurobiol. Learn. Mem.* **92,** 215-224.

Goldey ES, Kehn LS, Rehnberg GL, and Crofton KM (1995) Effects of developmental hypothyroidism on auditory and motor function in the rat. *Toxicol. Appl. Pharmacol.* **135**, 67-76.

Goldey ES, O'Callaghan JP, Stanton ME, Barone S, Jr., and Crofton KM (1994) Developmental neurotoxicity: evaluation of testing procedures with methylazoxymethanol and methylmercury. *Fundam. Appl. Toxicol.* **23**, 447-464.

Henck JW, Frahm DT, and Anderson JA (1996) Validation of automated behavioral test systems. *Neurotoxicol. Teratol.* **18**, 189-197.

Hoffman HS and Ison JR (1980) Reflex modification in the domain of startle: I. Some empirical findings and their implications for how the nervous system processes sensory input. *Psychol. Rev.* **87,** 175-189.

Holson RR, Freshwater L, Maurissen JPJ, Moser VC, and Phang W. (2008) Statistical issues and techniques appropriate for developmental neurotoxicity testing: A report from the ILSI Research Foundation/Risk Science Institute expert working group on neurodevelopmental endpoints. *Neurotoxicol. Teratol.* **30,** 326-348.

Ison JR and Hoffman HS (1983) Reflex modification in the domain of startle: II. The anomalous history of a robust and ubiquitous phenomenon. *Psychol. Bull.* **94**, 3-17.

Ison JR. (1984) Reflex modification as an objective test for sensory processing following toxicant exposure. *Neurobehav. Toxicol. Teratol.* **6,** 437-445.

Ison JR and Russo JM. (1990) Enhancement and depression of tactile and acoustic startle reflexes with variation in background noise level. *Psychobiol.* **18**, 96-100.

Kobayashi K, Tsuji R, Yoshioka T, Kushida M, Yabushita S, Sasaki M, Mino T, and Seki T. (2005) Effects of hypothyroidism induced by perinatal exposure to PTU on rat behavior and synaptic gene expression. *Toxicol.* **212**, 135-147.

Koch M (1999) The neurobiology of startle. *Prog. Neurobiol.* **59**, 107-128.

Leussis MP and Bolivar VJ. (2006) Habituation in rodents: A review of behavior, neurobiology, and genetics. *Neurosci. Biobehav. Rev.* **30,** 1045-1064.

Li L, Du Y, Li N, Wu X, Wu Y. (2009) Top-down modulation of prepulse inhibition of the startle reflex in humans and rats. *Neurosci. Biobehav. Rev.* **33**, 1157-1167.

Maurissen JP and Marable BR (2005) Neurotoxicity test validation, positive controls and proficiency: are chemicals necessary? *Neurotoxicol. Teratol.* **27**, 545-551.

Parisi T and Ison JR (1979) Development of the acoustic startle response in the rat: ontogenetic changes in the magnitude of inhibition by prepulse stimulation. *Dev. Psychobiol.* **12**, 219-230.

Raffaele KC, Fisher JE, Jr., Hancock S, Hazelden K, and Sobrian SK (2008) Determining normal variability in a developmental neurotoxicity test: a report from the ILSI Research Foundation/Risk Science Institute expert working group on neurodevelopmental endpoints. *Neurotoxicol. Teratol.* **30**, 288-325.

Muller KE and Benignus VA. (1992) Increasing scientific power with statistical power. *Neurotoxicol. Teratol.* **14**, 211-219.

Sheets LP, Dean KF, and Reiter LW (1988) Ontogeny of the acoustic startle response and sensitization to background noise in the rat. *Behav. Neurosci.* **102**, 706-713.

Shnerson A and Willott JF (1980) Ontogeny of the acoustic startle response in C57BL/6J mouse pups. *J. Comp. Physiol. Psychol.* **94**, 36-40.

Slikker, Jr. W, Acuff K, Boyes W, Chelonis J, Crofton KM, Dearlove G, Li A, Moser VC, Newland C, Rossi J, Schantz S, Sette W, Sheets L, Stanton M, Tyl S, and Sobotka TJ. (2005). Behavioral test methods workshop: A summary. *Neurotoxicol. Teratol.* **27**, 417-427.

Swerdlow NR, Geyer MA (1998) Using an animal model of deficient sensorimotor gating to study the pathophysiology and new treatments of schizophrenia. *Schizophr. Bull.* **24**, 285-301.

Tabachnick BG and Fidell LS. (2006) Using Multivariate Statistics. (5th edition), Pearson, Boston MA.

Thompson RF and Spencer WA. (1966) Habituation: a model phenomenon for the study of neuronal substrates of behavior. *Psychol. Rev.* **73,** 16-43.

Thompson RF. (2009) Habituation: A history. *Neurobiol. Learn. Mem.* **92,** 127-134.

Tyl RW, Crofton K, Moretto A, Moser V, Sheets LP, and Sobotka TJ (2008) Identification and interpretation of developmental neurotoxicity effects: a report from the ILSI Research Foundation/Risk Science Institute expert working group on neurodevelopmental endpoints. *Neurotoxicol. Teratol.* **30**, 349-381.

Williams JM, Hamilton LW, and Carlton PL (1975) Ontogenetic dissociation of two classes of habituation. *J. Comp. Physiol. Psychol.* **89**, 733-737.

Wise LD, Allen HL, Hoe CM, Verbeke DR, and Gerson RJ (1997) Developmental neurotoxicity evaluation of the avermectin pesticide, emamectin benzoate, in Sprague-Dawley rats. *Neurotoxicol. Teratol.* **19**, 315-326.

Young JS and Fechter LD (1983) Reflex inhibition procedures for animal audiometry: a technique for assessing ototoxicity. *J. Acoust. Soc. Am*. **73**, 1686-1693.

MODULE - D

EVALUATION OF LEARNING AND MEMORY DATA

MODULE D - INTERPRETATION OF LEARNING AND MEMORY DATA

Introduction

Learning and memory are theoretical constructs that are sometimes difficult to separate experimentally. Learning is defined as a relatively permanent change in behavior that is the result of experience. Memory is defined as the retention and use of acquired information about locations, events, or temporal order to modify subsequent behavior. Learning and memory can be inferred from changes in behavior. Most complex behaviors require the combinatorial participation of many neural systems that process different attributes of an experience (Gold, 2003). In the process of investigating the biological basis of learning and memory, neuroscientists, psychologists, and behavioral pharmacologists have devised a variety of specialized tests to distinguish different types of learning and memory, multiple definitions have emerged, and distinct neural systems attributed to specific types of learning and memory have been identified (Gold, 2003; Kihlstrom et al., 2007; Peele and Vincent, 1989; Squire, 2004; Vorhees and Williams, 2014a).

The simplest type of learning, non-associative learning, is exemplified by processes such as habituation or sensitization and has been addressed in the module on interpreting motor activity data and so will not be considered further here. Associative learning is the process by which an association between two stimuli (classical conditioning) is learned. Learning and memory are inextricably intertwined as the capacity for learning presupposes an ability to retain the knowledge acquired through experience, while memory stores the background knowledge against which new learning takes place. These processes have been assessed in a variety of ways including classical conditioning paradigms, instrumental learning, spatial learning, and sequential learning and are probed to a greater or lesser degree by different test procedures.

Some learning and memory tasks include two components of memory, i.e., reference or working memory. Reference memory involves remembering the rules of a given task that has been learned. Reference memory is acquired with repeated training and persists from days to months. This memory is an integral component to acquisition and is an example of how the lines blur between assessments of learning and assessments of memory. Reference memory is required for efficient completion of multiple test trials. In a maze task, reference memory might include acquiring the knowledge of the location of arms in the maze that are never rewarded, or learning that an escape platform exists and is in the same location over trials. Distinct from reference memory, working memory is typically defined as a short term memory for an object, stimulus, or location that is used within, but not typically between testing sessions. Working memory uses flexible stimulus-response associations and is sensitive to interference (Green and Stanton, 1989; Olton, 1979). Working memory in animals can be viewed as having a limited lifetime or being task-relevant for only a brief period. A common feature of working memory tasks that can be exploited in variety of simple test systems is that performance decreases as the delay over which a memory is held increases (Carter et al., 1995). Unfortunately this feature has only rarely been exploited in DNT guideline studies.

Learning and memory depend upon the coordinated action of different brain regions and neurotransmitter systems constituting a functionally integrated neural network (D'Hooge and DeDeyn, 2001). The main learning areas and pathways are similar in rodents and primates but specific task requirements preferentially engage some brain regions more than others

(Eichenbaum, 2000; Stanton and Spear, 1990). Very superficially speaking, spatial learning is believed to be primarily based in hippocampal function, fear-based learning involves the amygdala, and attention tasks require prefrontal cortex input (Clark et al., 2000; Eichenbaum, 2000; Squire, 2004).

The variety of tests available in the literature to assess learning and memory and their potential application for guideline testing have been reviewed (Graham et al., 2012; Paul et al., 2009; Peele and Vincent, 1989; OECD Guidance Document 151, 2013; Vorhees and Williams, 2014a, 2014b). A number of considerations must guide the selection of tests for assessing learning and memory in a regulatory context, not the least of which are cost effectiveness, sensitivity and selectivity. Typically methods used in DNT guideline studies tend to be simpler tests; the ones that appear most frequently are the M, E, Y, and T mazes, Morris Water Maze (MWM), passive avoidance (PA), and Cincinnati or Biel Maze (see Table 1). Since these tests have been used in almost all submitted DNT studies, this document focuses only on those. Less often, spatial delayed alternation and delayed matching to sample tasks have been employed. The latter two are potentially more rigorous evaluations of memory, as they can be designed to incorporate time delays. However, they are more complex and labour intensive to run.

It is important to recognize that various tests are assessing different forms of learning or memory and it is reasonable to expect that the effects of any one chemical may vary depending on the specific test. It is also well documented that any given test is unlikely to be sensitive to effects caused by all different types of chemicals (Peele and Vincent, 1989; Tilson and Mitchell, 1984). Similarly, it is not anticipated that a performance impairment detected in a learning and memory test would necessarily be reflected in outcomes from brain pathology or brain morphometry, and vice versa.

In addition, tests that are appropriate for one age of animal may not be appropriate for a different age. Certainly the physical and neurological substrates that are probed by any given task must be fully developed before that task can be applied to evaluate chemical-induced impairment (i.e., animals must be able to swim, locomote, hear and see to use tests requiring these abilities). Certainly all of these abilities are fully developed in animals at the time of weaning, the earliest time point recommended for learning and memory tests in guideline studies. For some tasks, weanling animals are not as proficient as adults, and this relative capacity is due to the immaturity of the brain circuitry that underlies the learning rather than physical ability to perform (Brown et al., 2005; Ehman and Moser, 2006; Green and Stanton, 1989).

Table 1. Summary of typical tests of learning and memory submitted in guideline studies of DNT testing.

Test System	Frequency of Report in North American Regulatory Submissions (as of 2014)	Cognitive Ability Assessed	Comment
Letter Mazes M, Y, E	M : 34 Y : 12 E : 3	Position Discrimination	Simplest form of learning assessed most common but least sensitive test procedure
Morris Water Maze	9	Spatial Learning	Typical acquisition is test of reference memory with single probe trial to test retention at 24 hr
Passive Avoidance	33	Associative Learning	Single trial learning often followed by immediate recall (typically < 1 hr) to assess learning and delayed recall (hrs to days) to test memory. Multiple acquisition trials are often but unnecessarily run as most of learning occurs within the first trial
Biel Water Maze	6	Sequential Learning	Multiple T-mazes. Combined reference and working memory task. Intramaze cues detract from its sensitivity
Cincinnati Maze	2	Sequential/ Egocentric Learning	Multiple T-mazes, improvement over Biel maze. Combined reference and working memory task.
T-Maze	2	Delayed Alternation	Typically this procedure involves position discrimination. Incorporation of delays can enhance sensitivity. Reference and working memory assessed. Although a more rigorous test than simple position discrimination, it is not frequently used. Expert consultation is advised.
T-Maze	1	Delayed Matching to Position	Infrequently used. Expert consultation is advised.
Schedule-Controlled Behavior	0	Delayed Matching to Position	Although suggested by US EPA Guideline 870.6500, it has not been used to date. Expert consultation is advised.

1. Guideline Requirements

As cognitive function is one of the basic processes that can potentially be disrupted by chemical exposure during development, tests of this parameter are included in many studies that are designed to evaluate DNT (e.g., USEPA OPPTS 870.6300 and OECD TG426). The U.S. EPA test guideline for a DNT study (US EPA OPPTS 870.6300) does not specify which test to perform, but rather states more generally that: 1) associative learning must be assessed as either a change across several repeated learning trials or sessions, or in tests involving a single trial, with inclusion of some condition that controls for non-associative effects of the training experience; and 2) the tests should include some measure of memory (retention) in addition to original learning (acquisition) where appropriate. The guidelines also specify that the choice of test to evaluate learning and memory in a given circumstance should consider findings in other toxicity studies for that chemical. Measures of learning and memory are also influenced by sensory, motor and motivational variables (Tilson and Mitchell, 1984). The interpretation of data from a learning/memory task should incorporate consideration of these variables by using the appropriate control procedures or conducting independent assessments (e.g. ability to swim in a swim maze and find a visible platform). While regulatory test requirements for OECD and EPA are comparable for Learning and Memory testing as part of DNT studies, there are some differences that should be considered. These are summarized in Table 2. Both guidelines require a minimum of 10 animals/sex/dose and testing in the early postweaning period and in the young adult; all litters should be represented. Note that while the OECD guidelines state testing at PND25, testing may need to be several days later if pups are not weaned before PND25. Task selection should always consider the developmental state of the animal. Although not specified in the guidelines, US EPA has also recommended that testing should be performed in different groups of animals at these two ages if the same test is used. This is consistent with OECD's recommendation that different tests or different animals be used at the two ages as repeated testing of the same animals in the same test is likely to decrease sensitivity due to carry over effects. If different tests are evaluated such that learning on one does not interfere with or facilitate learning on the other, the same animals can be tested at the weaning and adult ages.

Table 2. Learning and Memory testing requirements in EPA and OECD guidelines. The specific requirements are essentially the same in both EPA and OECD 426 guidelines. The OECD 443 extended 1-generation reproductive toxicity study guideline does not mention cognitive testing except as a potential additional measure where needed.

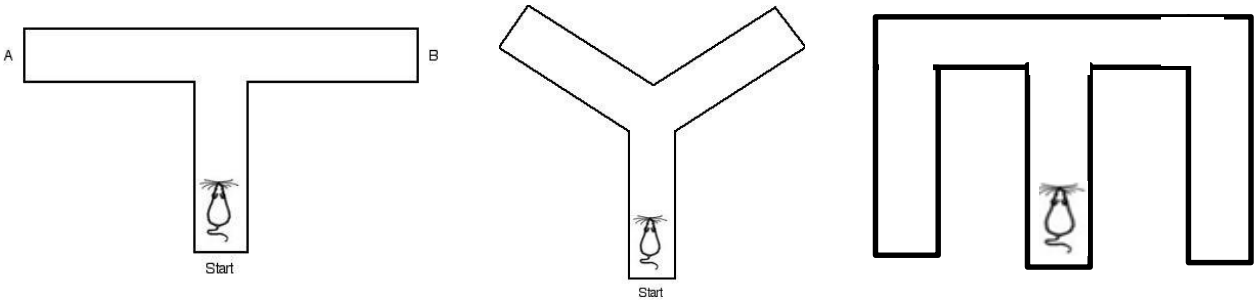
		US EPA OPPTS 870.6300	OECD 426	OECD 443 Extended 1-Gen
Age of testing	OECD: PND25±2 and PND60 and older US EPA: around weaning and around PND60	√	√	Not applicable
Minimal Test Requirements	Test of Task Acquisition- Learning, a change across several learning trials or a single trial.	√	√	

	Test of Retention- Short or long term memory. Cannot be interpreted in absence of documentation of asymptotic acquisition on same task.	√	√	Not applicable
Performance Controls	If the cognitive test reveals an effect of the test compound, additional tests may be conducted to rule out alternative interpretations based on deficiencies in sensory, motivational, and/or motor capacities.	√	√	Not applicable
Choice of Procedure	Not specified. When possible, learning and memory test should be chosen based on demonstrated sensitivity to the class of compound under investigation. Same or different tests may be used at different ages.	√	√	Not applicable
Suggested Learning and Memory Tests	US EPA: Delayed-matching-to-position, olfactory conditioning, acquisition and retention of schedule-controlled behavior. OECD: US EPA recommended tests plus Morris water maze, Biel or Cincinnati maze, radial arm maze, T-maze.	√	√	Not applicable
History of Animal	OECD: Different animals must be used at each test age. US EPA: No specific guidance provided. *Generally accepted that same animal can be tested at both ages if different tests are used.*	√	√	If data indicate need for cognitive testing, it must be integrated without compromising the integrity of the other evaluations in the study.

2. Test Description: Position Discrimination Learning Using Letter Mazes (Y, M, E, T)

Position discrimination is a very simple form of maze learning and, perhaps because of its simplicity, has been the most commonly used test of learning in DNT guideline studies submitted to date. It is not, however, a widely applied test of learning in the academic literature, due to its lower sensitivity in detecting behavioral impairments. In this task the animal has to make a binary choice to turn left or right to escape the water (see Figure 1 for example diagrams). These mazes are in the shape of a Y, M, E, or T, earning them the moniker 'letter mazes' and can be either land- or water-based. In the former, animals are generally food restricted and rewarded for correct turning response by delivery of a food reward. As food restriction procedures may become more complicated (especially in growing pups, or in cases where there is already a treatment effect on body weight), water-based maze procedures are more routinely used in submitted regulatory studies. In the water-based version of this task, the reinforcer, which motivates the animal to learn the task, takes the form of an escape from the water, via a platform positioned at the end of the correct choice arm. The following discussion will be restricted to water-based letter mazes used to test for position discrimination learning.

Figure 1. Various renditions of simple position discrimination mazes, T, Y, and E/M. Animals are placed in the 'Start' arm and must make either a left or a right turn to swim to the escape platform. The correct position should be counterbalanced across animals in each treatment group.



Mazes used in the assessment of position discrimination learning can be configured as a T, Y, E or M but all have the same basic form of straight alley 'start arm' at the end of which is a choice point for the animal to turn left or right; only one of which is rewarded by the presence of an escape ladder or platform. The trials are run as discrete entities, with a relatively brief but consistent intertrial interval. Only one arm position is rewarded for any given animal during training, but the correct arm should be counterbalanced across animals (i.e. equal number of animals per arm position, per dose group). The consequences of an 'error' (e.g., turning left instead of right) are fairly inconsequential to the animal, since it simply turns around and swims to the opposite 'correct arm'. If the maze is small, this only minimally increases the time in the water and time to escape.

Position discrimination is often described as a reference memory task as the same arm is designated the correct response arm throughout the test (e.g., always turn right). Reference memory is distinct from working memory which can also be assessed in this type of apparatus. A working memory task (e.g., spatial delayed alternation) requires the animals to alternate between

the right and left arms for sequential trials (e.g., if you went right last time, go left this time). Working memory tasks are cognitively more demanding than simple position discrimination, and they require more extensive training for successful performance. Working memory tasks are disrupted by treatments that negatively impact function of the hippocampus (Green and Stanton, 1989; Freeman and Stanton, 1991; Watson et al., 2009) yet leave reference memory (position discrimination) relatively intact (Chadman et al., 2006; Watson et al., 2006).

Because position discrimination learning develops early in postnatal life, as young as PND15, (Green and Stanton, 1989; Brown et al., 2005; Pagani et al., 2005; Kenny and Blass, 1977; Smith and Bogomolny, 1983), both weanling and adult animals can be readily assessed for this simple form of learning. However, because of potential carry-over learning, the same animals should not be assessed in the same task at both adult and weanling ages. For younger animals, the size of the maze should be scaled accordingly to accommodate the smaller body mass of the animal and to avoid fatigue that could interfere with ability to perform the task.

2.1. Procedures for Position Discrimination Learning

As almost all spatial learning studies submitted for regulatory evaluation are water-based mazes, the following description of procedures will be limited to tasks of this type. Many of the same principles, however, are also applicable to land-based mazes. Details completely describing the maze set-up, size, and procedures should be included in the report.

Straight Alley Test for Swim Speed, Acclimation, Side Bias.

To ensure that the animals are physically able to perform the task and to familiarize them with the minimal task requirements – swimming and escape from the water -- animals are initially placed in a straight alley. Straight channel swim speed provides information regarding any potential deficit in motor function. An escape ladder or platform is clearly visible at the other end of the alley. The animal's swim speed can be calculated, a parameter that should be assessed if latency measures are used as metrics for learning. If video tracking is used for data collection, swim speed should also be calculated on every trial. Motor dysfunction could potentially confound interpretation of learning deficits if a decrease in latency is used as the main measure of learning for the task, in that an increase in latency could be the result of a decrease in swim speed, rather than an inability to learn.

In some procedures, the first learning trial may also serve as an acclimation trial, in which animals are allowed to explore both arms and an escape opportunity is provided in the opposite arm that is entered. The 'correct' arm for that particular animal for all subsequent trials is the arm opposite of the one first entered. This procedure may guard against a side preference bias that some animals may exhibit. Ideally, designation of the correct side should be counterbalanced within and across treatment groups.

Task Acquisition.

Animals are released in the start box and allowed to swim to the choice point and select a left or a right turn. If the correct choice is made, animals will find an escape route at the end of the alley. If an incorrect choice is made, the animal remains in the maze until it swims to the correct alley or until a set period of time (typically 60-90 seconds) has elapsed. In the case of the latter, animals may be guided to the escape platform by the experimenter. Alternatively, the animal may simply be removed from the incorrect arm and the trial ended (non-correction procedure).

Note that here, and in many other tests, setting a maximum time for testing produces closed-ended values that can impact statistical analyses (see Section 9). Acquisition testing typically occurs within one day, with a set maximum number of trials for each animal and a fixed intertrial interval (ITI). With a short ITI relative to maximal trial duration of 60-90 seconds, it is possible that fatigue may compromise performance in those animals, increasing variability and confounding interpretation of latency data. However, a review of submitted data has not identified this as a common occurrence.

Generally, testing for each animal will continue until learning has been demonstrated according to a pre-defined level of proficiency (i.e., a learning criterion has been met) or until some pre-defined maximum number of trials has been administered. Criterion performance is typically set at 5 consecutive correct trials, at which point testing is ended for that animal. Different criteria may be utilized but must be explicitly defined and represent a valid index of task acquisition. Taking testing out to the maximum number of trials even if criterion has been achieved in fewer trials facilitates the evaluation of learning curves and statistical analyses. However, this practice does run the risk of having 'overlearning' which may decrease sensitivity of retention tests performed at a later date (see below).

Retention.

For position discrimination learning, retention is usually evaluated one week or so after the acquisition phase. The protocol for retention testing is generally identical to that for acquisition, with a specific maximum number of trials for each animal and a learning criterion of a set number of consecutive correct trials. The correct goal is the same as that designated during acquisition. If the animal has retained the learned information, criterion is frequently achieved more quickly than in acquisition. Note, however, that only the first trial during that second week can be considered "memory", since subsequent testing uses the information from the previous trial, not trials one week prior. Performance after the first trial can be considered re-learning.

It is not possible to test retention in animals that did not achieve criterion performance of the task during the acquisition phase. Such animals should be excluded from the retention testing, and the failure to acquire the task should be noted in the presented results. When this is the case, it should be noted that there will be differing sample sizes across groups. Furthermore, such a practice leads to testing only a selected population, which may lead to bias in the outcomes. Alternatively, the acquisition phase could be extended until all animals achieve criterion performance prior to testing retention.

2.2. Standard Test Conditions for Position Discrimination Learning.

All standard test conditions described below should be maintained for appropriate conduct of position discrimination learning:

- Maze size and water temperature should be specified, within acceptable limits, and appropriate to the age and size of the animal being tested;
- The animal should be placed in the start position of the maze facing the back wall of the maze;
- The procedure used if an animal does not reach the platform within the maximum test time must be specified (e.g., whether it is immediately removed or taken to the platform);
- The duration of time animal is left on the escape platform on completion of each trial

must be consistent and specified;
- The intertrial interval (the delay between removing the animal from the platform and the beginning of the next trial) and handling and holding conditions between each trial must be consistent and specified;
- Criterion performance in the acquisition phase must be achieved in all subjects evaluated for retention learning.

2.3. Metrics to Evaluate Learning and Retention in Position Discrimination Learning

There are a few basic dependent measures that are used to evaluate position discrimination learning: escape latency, number of errors, and trials to criterion. These may be the only measures when data are collected by an observer, but more information (e.g., path length, direction) may be obtained if video-based tracking systems are used. The basic endpoints are described in more detail below. Some are redundant and should be highly correlated, but with slight nuanced differences.

Latency to escape

This measure reflects the time from release of the animal in the maze until it reaches the correct goal and exits the maze. Latency will typically decrease over trials, reaching some asymptotic level. Note that latency is to some extent dependent on the size of the maze/length of the alleys, and thus varies from laboratory to laboratory. Standardization of alley length or minimal requirements would facilitate interpretation and comparison across studies. Extremely short latencies may indicate a very small maze, and may complicate interpretation of results, especially when presenting differences as percent change across trials.

Errors

It is imperative that an explicit operational definition of an 'error' is provided (see Section 7). Errors will typically decrease over trials, with criterion performance set as a specific number of consecutive trials with no errors. The number of errors by treatment group for each trial should be noted, as well as the number of animals in each group making any error for each trial.

Trials to criterion

Criterion performance may be measured as the number of trials before criterion is reached (trials to criterion) or as the number of total errors each animal makes over the testing session (errors to criterion). In the latter case, it is important to determine whether criterion was in fact obtained, i.e., whether the total errors to criterion are inclusive of all animals or only those that reached criterion. In some cases, the percentage of animals in each group that achieved criterion performance may be provided. In these cases, the total number of trials or errors is only a single number and as such cannot provide a learning curve that shows change over time. For this reason, evaluation of error data over trials provides a better representation of learning.

The percentage of correct trials (number of correct trials for a given animal divided by the total number of trials completed by that animal) may be useful when there are differences among animals in the number of trials run.

Typically, animals that do not meet criterion are considered non-learners, and are not tested

for retention or re-learning. As mentioned above, this practice may lead to different sample sizes as well as selection bias across groups. Furthermore, the use of criterion data alone may be misleading, leading to the classification of non-learner being applied for different situations. In the example shown in Table 3, the rats show very different patterns of behavior, yet all are considered non-learners by the definition used in this particular laboratory. Non-learners were defined as rats that did not perform five consecutive errorless trials for a maximum of 15 trials. Specifically, rat 1 makes some number of errors on every trial. After the 11th trial, it is not possible for criterion to be met, so that rat is not tested for the last 4 trials and a maximum value of 15 is recorded. Rat 2 shows a decreasing number of errors over trials, and reaches criterion but only on the 15th trial; thus this animal by the laboratory's definition is considered a non-learner. Rat 3 makes relatively few errors, reaching as many as four but not five consecutive trials with no errors. As shown in the last column of Table 3, these rats display different outcomes in terms of other endpoints, namely total and mean errors as well as trials with no errors, which are considerably more informative than the trials to criterion or number of rats reaching criterion. Rat 3 in this example had fewer errors and more errorless trials than the other rats, indicative of a more consistent pattern of learning.

Table 3. Example of individual data (number of errors per trial) for three rats. Total number is 15 trials, and criterion performance is 5 consecutive trials with no errors. NT indicates that the rat was not tested.

Trial	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Trials to criterion	Total and mean errors	Trials with no errors
Rat 1	5	2	4	1	2	3	2	1	5	1	2	NT	NT	NT	NT	15	28 2.5/trial	0
Rat 2	3	2	3	2	0	2	1	2	0	2	0	0	0	0	0	15	17 1.1/trial	7
Rat 3	1	0	0	0	2	0	0	1	0	1	0	0	0	0	1	15	6 0.4/trial	11

2.4. Data Reporting for Position Discrimination Learning.

In order to evaluate the effects of chemicals on these parameters, it is essential that DNT position discrimination reports provide clear information on study design, the measures (dependent variables), and the statistical procedures used to evaluate the data. The main results that are of interest from DNT learning and memory assessments are:

- The effects of chemicals on acquisition, i.e., initial learning rate;
- Effects of chemicals on memory, i.e., retention of the learned response;
- If position discrimination was assessed at both weanling and adult ages, is the deficit seen at both ages;
- Sex differences in effects of chemicals on learning and/or retention.

Dependent Variables

It is important that the dependent variables recorded and analyzed are unambiguously described. For position discrimination, the dependent measures include the following:

- Latency to reach the correct goal, and how this changes over trials;
- Number of errors (incorrect turns within each trial), and how this changes over trials;
- Cumulative as well as average number of errors across all trials (usually reported as a maximum of one per trial, although more errors could potentially be committed);
- Number of trials that were not completed (e.g., animal did not reach goal within allotted time);
- Trials to criterion;
- Numbers (proportions) of each group that did (or did not) achieve criterion;
- Errors to criterion – similar to cumulative number of errors but eliminates animals that failed to reach criterion.

Reported Data

In order to evaluate treatment effects, the following details should be reported for each sex at each age group. In cases where data are non-normally distributed, different measures of central tendency may be more appropriate (e.g., median, mode).

- Mean ± standard deviations of latency per treatment group on each trial (not averaged over all trials);
- Mean ± standard deviations of total number of errors to criterion performance per treatment group;
- Median or mode trials to criterion performance per treatment group (which may be more appropriate than mean);
- Median or mode for number of errors per treatment group on each trial (which may be more appropriate than mean because of the type of data, e.g., count, or the underlying distribution);
- Number of animals in each treatment group meeting criterion on each trial as well as number of animals reaching criterion performance;
- Individual trial data for each animal should be included in an appendix;
- Summary data should be presented in tabular and graphic form.

Data should be reported separately for acquisition and retention assessments. All data should be reported on an individual and group basis. Individual data should include every trial performed for each animal, including the first (acclimation) trial. Note that some laboratories frequently fail to include data for the first trial(s). Also, note that the number of errors is usually reported as a maximum of one per trial, although more errors could potentially be committed. It may be necessary to request the record of all the errors for all trials, including the first.

2.5. Interpretation of Position Discrimination Results

The following points should be taken into consideration in the interpretation of the data:

- Control animals should demonstrate a clear and statistically reliable reduction in both latencies and errors over trials. The absence of this in control animals indicates lack of a ‘learning curve’ and as such the test does not contain the minimal requirements to detect a change in behavior over time and does not measure learning;
- A treatment-induced increase in the number of errors or trials to criterion is evidence of a learning impairment;
- A treatment-induced decline in the number of animals reaching criterion performance is evidence of a learning impairment;

- When slight differences in swim speed are present, it is advisable to consider their relative contribution to the magnitude of the impaired performance. Latency increases in the presence of slower swim speeds in treated animals may indicate performance or motoric deficits which can confound the interpretation of cognitive impairment. On the other hand, faster swim speeds induced by treatments may mask impairments based on latency measures. In cases where swim speed is impacted by treatment, other metrics (i.e., errors to criterion, number of animals reaching criterion performance) may be more reliable to determine presence of a learning impairment;
- Variability must be considered in interpretation of findings (Raffaele et al, 2008). Excessive variability can make it impossible to demonstrate that learning has occurred, and will also make it difficult to detect changes in performance between control and treatment groups. Variability that is minimal or absent may also indicate a lack of sensitivity of the task to detect potential cognitive impairments, and should also be of concern;
- Positive control data as described below in Section 8 must also be included to display competency that the laboratory, using the same procedures and protocols, can detect impairment if one exists.

3. Test Description: Spatial Learning using the Morris Water Maze Test

The Morris water maze (MWM) (D'Hooge and De Deyn, 2001; Morris, 1981, 1984) is widely used in the field of cognitive neuroscience as well as neurotoxicology, and is supported by a rich literature regarding the effects of drugs on learning and memory and involved brain regions. Place training in the Morris maze assesses spatial learning abilities, which are believed to be primarily hippocampal-based (Clark et al., 2000; Squire, 2004; Eichenbaum, 2000), whereas memory probes tap the nature, persistence and organization of memory representations within the cerebral cortex.

In the MWM, animals rely on distal cues in the environment to navigate from various start locations around the perimeter of an open swimming arena to locate a submerged escape platform. As with the water-based position discrimination mazes, escape from water provides the motivation, and no food or drinking water restriction is required. Spatial learning is assessed across repeated trials and reference memory is determined by preference for the platform area when the platform is absent. Although working memory is less frequently assessed using this task, protocols have been developed and a number of papers have been published demonstrating working memory deficits using the MWM (Dudchenko 2004; Dudchenko et al., 2013; Vorhees and Williams, 2006). The working memory version of the task has not, however, been used in guideline toxicological assessments and so will not be discussed further here.

Measures of learning and memory are influenced by sensory, motor and motivational variables (Tilson and Mitchell, 1984). The interpretation of data from a learning/memory task should incorporate consideration of these variables by using the appropriate control procedures or conducting independent assessments (e.g. ability to swim in a swim maze and find a visible platform). Recently, the MWM has become popular for the fulfillment of International Congress for Harmonization (ICH) and Food and Drug Administration (FDA) juvenile studies. Thus, while not as common in standard DNT studies as the letter mazes, there are a number of studies that have and will use this test paradigm. The MWM has been successfully used to detect learning and memory deficits in young and adult rats and mice. As with position discrimination

learning, the tank used in the MWM is scaled to the size of the animal and a number of optimal parameters have been tested and summarized in recent reviews by Vorhees and Williams (2006; 2014a; 2014b). A number of variations of the MWM basic protocol can be applied to probe for different aspects of spatial learning and memory. However, despite extensive use of the MWM, the task has not always been optimally implemented in guideline or literature-based studies. Some of this stems from an under-appreciation for specific aspects of the apparatus and testing procedures that are most salient for obtaining the most reliable and interpretable results.

The maze itself is a simple circular pool that should be at least three-quarters filled with water (see Figure 2). It is critical that the interior is made such that it is as close to being featureless as possible to eliminate as much as possible any intramaze cues to help the animal locate the platform. It is also important that the escape platform lies just below the surface of the water and cannot be visually detected by the animal as it swims. Making the water opaque or making the maze and platform the same color are examples of methods that may be used to camouflage the platform. The platform should be placed far enough from the tank wall so that the rat cannot find it by simply swimming close to the wall around the tank (thigmotaxic behavior). The goal is to have the animal locate the hidden escape platform using spatial cues positioned around (outside of) the maze. The distal cues, e.g., objects or pictures, should be large and not color-based since rodents have relatively poor and monochromatic vision. As there are no proximal cues available, i.e., cues within the maze itself, the use of distal cues provides the most effective strategy to locate the hidden escape platform. The escape platform is typically 10 cm in diameter for rats (smaller for mice), made of clear acrylic, and may have a rubberized surface to increase the animal's traction to climb on once located. Scaling of the size of the pool and the size of the escape platform can highly influence the difficulty of the task – the larger the pool and the smaller the platform, the more difficult the task. The search area to escape target ratio is an important variable to consider and optimal ratios vary between mouse and rat (Vorhees and Williams, 2006; 2014a; 2014b). The maze is divided into quadrants with cardinal positions arbitrarily designated as north, south, east, or west (N, S, E, W). While there is no accepted optimal water temperature, studies have shown that rats learn better with water at 19-25º C but higher or lower than that will impair learning (Salehi et al., 2010). Details completely describing the water maze set-up, size, and procedure should be included in the report.

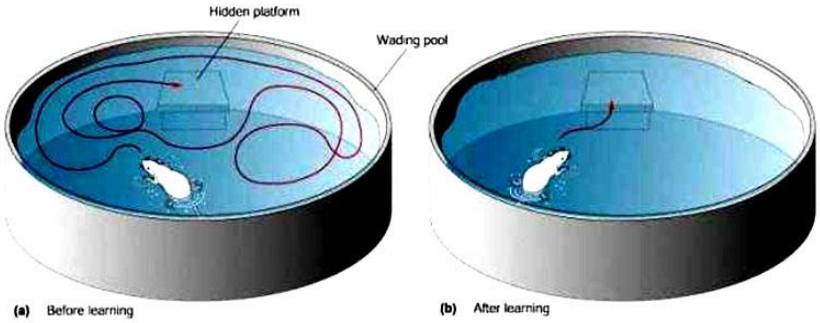


Figure 2. Morris Water Maze. Animals must navigate a large pool to find a hidden escape platform submerged below the water surface. Different starting points are used on each trial but the position of the platform remains fixed over trials. Distal cues are placed on the walls surrounding the maze and remain invariant over the course of the training

The platform is positioned in one of the four quadrants and animals start trials from one of the four cardinal positions as described below. The escape platform is typically positioned in the middle of one of the quadrants and remains there throughout testing. The animal is released into the water facing the wall of the tank and must search for the submerged platform. For video tracking, a camera is suspended from the ceiling and positioned over the middle of the maze and attached to a recording device. The animal's position in the pool is recorded, from which the software computes a number of variables, the most common of which are distance and time measures. Other measures that may be included are positioning data throughout the trial that provide a spatial analysis of swim patterns (Blokland et al., 2004); however, such data are rarely reported in guideline studies.

Two important performance control features should be included to rule out treatment-related visual and motoric impairments that can confound interpretation of cognitive deficits identified using this task. These include swim speed and cued-trials and are discussed in more detail below.

3.1. Procedures for Spatial Acquisition

Acquisition.

Place or spatial learning is the most basic MWM procedure. Typically either 4 trials are administered on each of 5 consecutive days or 2 trials/day are administered for 10 to 12 days. Most protocols use four start locations, N, S, E and W. Each start location is represented each day and locations are randomly chosen within a day, without replacement. As such, all start positions are used each day in 4 trials/day protocols and across 2 days in 2 trials/day protocols; the same start point is never repeated within a day. The order of start points is also varied from day to day. These are designed so that the animal is not able to learn a specific order of right or left 'headings' to locate the platform, but rather must use the distal spatial cues. Animals are typically run in squads of 4 such that the interval between trials is ~ 5 minutes if all 4 animals are run in order on each trial. After locating and mounting the escape platform, the animal is left to rest there for 15-30 seconds. If an animal fails to find the platform within the allotted time (typically 60-90 seconds), it is guided there or placed on the platform for the same brief period. The water should be gently stirred between trials to prevent rats following urine trails from previous animals (Means et al., 1992). Using a heating pad under or around the cage may keep the animal warm while drying.

Below, Figure 3 depicts learning in an individual rat using video tracking during early (left) and late (right) learning trials. Path length drawn in these tracings is automatically calculated by software and is highly correlated with latency. Figure 4 illustrates some examples of learning curves based on latency to find the escape platform in a 2 trials/day protocol (left) and a 4 trials/day protocol (right). Propylthioruacil during early development impaired the acquisition of spatial learning in the adult. A different pattern was seen with developmental treatment with tebuconazole where longer latencies were seen in the high dose group on all days of acquisition.

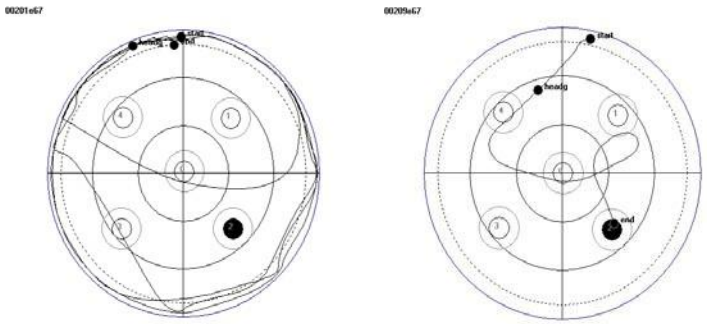


Figure 3. Individual tracings of the same rat on the first and last day of spatial training. The video system imposes landmarks (e.g., platform positions): platform 2, darkened circle in the right bottom quadrant, is the target. On the first day, the rat swims mostly near the tank wall, and does not find the platform within 60 sec. On the last day the rat quickly heads in the correct direction and has a latency of 10.7 sec (Moser, personal communication).

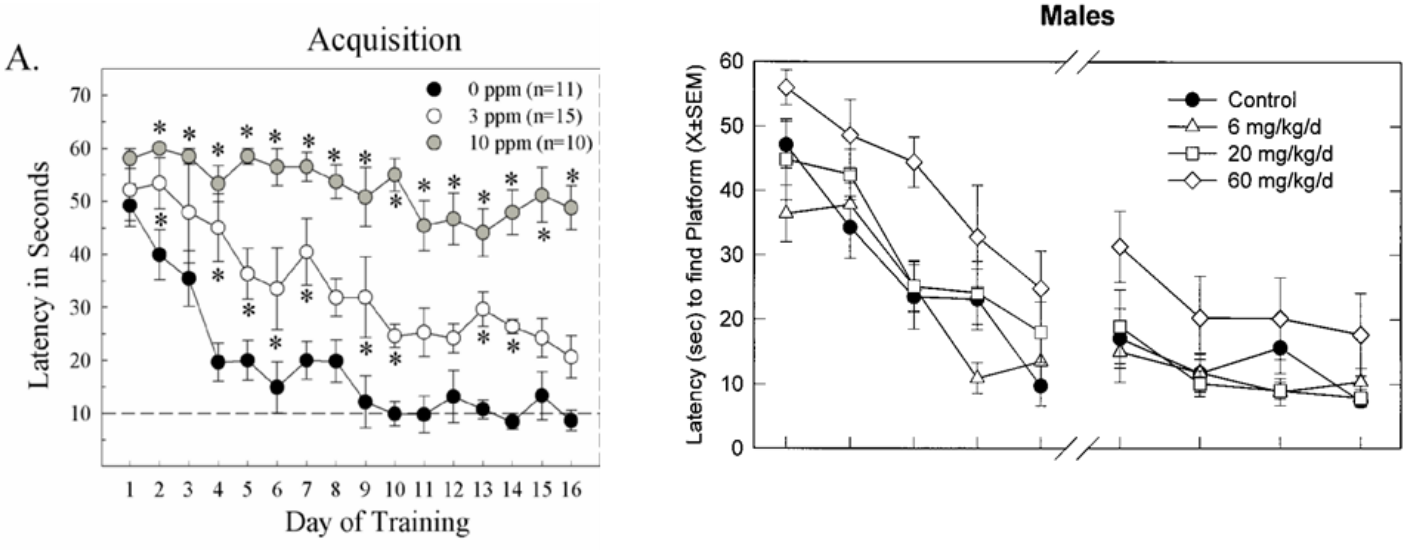


Figure 4. Examples of acquisition data (latency to find platform) across training trials for two different chemicals. On the left are latency data (X±SEM) from animals with thyroid hormone insufficiency following developmental exposure to propylthiouracil. Control animals show a decrease in latency over training trials (2 trials/day), with longer latencies in exposed animals. Very little learning is evident in the high dose group. Significant main effects of dose and day, and a significant dose-by-day interaction were detected in this study (Gilbert and Sui, 2006). On the right are data on rats exposed developmentally to tebuconazole, in which there was a significant effect on learning at the high dose (Moser et al., 2001). There were no significant group differences on the very first trial, but when analyzed across daily blocks (Monday through Friday the first week, Monday through Thursday the second week, 4 trials/day) there was an overall dose effect (no significant dose-by-day interaction). This indicates that the treated rats were able to learn but the rate of learning was slower than controls. In both studies, swim speed was not altered by treatment.

Retention.

To assess reference memory at the end of learning, a probe trial is given in which the platform is removed. During this time, the animal searches the area where the platform had been located. A duration of 30 or 60 sec is adequate to observe bias for the target area, and with longer durations the animals may begin to search more randomly. The most common method is to administer one probe trial 24 h after the last acquisition day. With some procedures, the probe trial is administered immediately following the last learning trial but this is not desirable as such a procedure cannot differentiate between short- and long-term memory. Performance of the probe trial on the same day as acquisition may reflect memory for the most recent training session rather than retention of the spatial location from the previous day. A longer interval between the last training trial and the probe trial is essential if reference memory is to be determined independently of the memory of the last training session.

Performance Controls.

Test guidelines specify that when cognitive deficits are detected, performance control procedures (e.g., swim speed and cue learning) should be conducted to dissociate cognitive deficits from sensorimotor performance impairments. Two often-used performance controls are assessment of swim speed and visual function. It is common practice in published literature reports to include data on both of these parameters as they are often deemed critical for data interpretation.

a) Swim Speed Performance Control.

As in position discrimination learning, it is necessary to ensure that the animals are physically able to perform the task – swimming and escaping from the water. This can be conducted in a separate apparatus such as a straight alley as described for position discrimination, or can be conducted in the MWM if the system is equipped with tracking from which swim speed can be calculated using path length and latency measures. Swim speed provides information regarding whether there is any deficit in motor function, which could potentially confound interpretation of learning deficits if a decrease in latency is used as the main measure of learning for the task – i.e., an increase in latency could be the result of a decrease in swim speed rather than an inability to learn, or, conversely, faster swim speeds may mask learning deficits based on latency measures. Since test systems that use video tracking can calculate swim speed for all test trials, this information could be used to assess motor abilities throughout training.

b) Visual Performance Control.

A cued version of the learning task could be conducted as another sensorimotor performance control in the MWM in order to dissociate a cognitive impairment from a sensory impairment. Cue learning differs from place learning since the cues are internal rather than external to the tank. During this trial the platform is moved to a different quadrant (typically the opposite quadrant), and is raised above the water surface and made clearly visible (colored disk or position marked with flag). The animal simply has to swim to the visible platform. When multiple visual trials are conducted, both the platform position and starting point must be moved randomly on every trial. Cue learning combined with assessments of swim speed provides information on the animal's sensory and motoric competence as well as motivation to escape.

Other Learning and Performance Tasks.

Other procedures using MWM include reversal learning and a working memory version of the task. Both have been reported in the literature and may provide additional, and perhaps more sensitive, assessments of learning and memory (Gilbert and Sui, 2006; Hoh et al., 1999; Morford

et al., 2002; Vorhees and Williams, 2006; Williams et al., 2002; 2003). Although these are not typically implemented in DNT studies, a brief description and the potential utility of these tasks are described below.

a) Reversal Learning.

Once animals have acquired the task to asymptotic and comparable degrees across treatment groups, reversal learning can be evaluated. In the MWM, reversal learning simply entails placing the escape platform in the opposite quadrant used during initial training and determines the number of trials required for the animal to recognize the shift. An identical procedure to that used during acquisition is used and if these trials are to be run, they may precede cue learning as the visual performance control. Typically, control animals adjust to the new location within 3-5 trials. Perseveration to the previously correct position is common in brain-damaged animals and may reflect a failure to recognize or adjust to the new task requirements (e.g., Whishaw and Tomie, 1997; Castañe et al., 2010). Continued responding with the previously correct choice may stem from an inability to recognize the change in the task contingencies (what the animal needs to do in order to escape), or an inability to learn a new response pattern, both examples of cognitive inflexibility. Alternatively, treated animals could exhibit fewer errors than controls on the reversal challenge, due to poorer retention of the initial task and thus less 'interference' as the requirements change under reversal conditions.

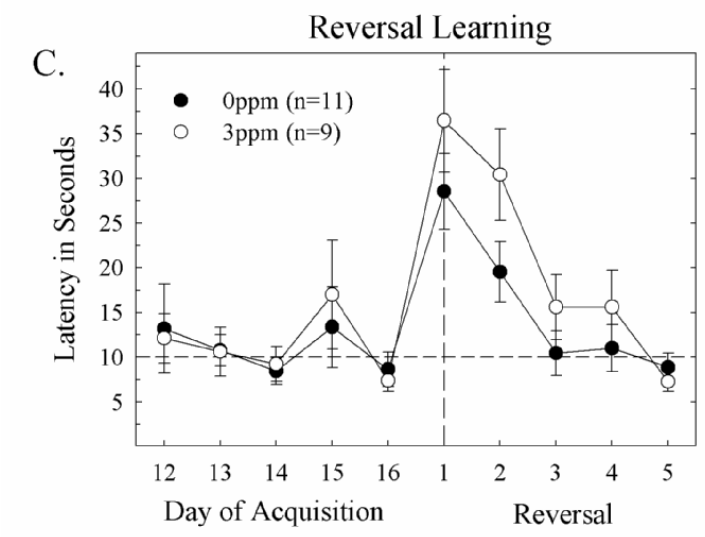


Figure 5. Reversal learning in a subset of animals that were exposed developmentally to propylthiouracil. The rats had acquired the task to criterion performance (final days of acquisition on left), and the platform was then placed in the opposite quadrant to evaluate reversal learning. Both groups quickly acquired the reversal, but the treated animals exhibited slightly longer latencies suggestive of a reversal learning impairment.

b) Working Memory.

Working memory refers to short-term retention of information, and may be measured using a match-to-position version of the MWM (Vorhees and Williams, 2014a). In this procedure, the platform is placed in a new position each day. The first trial allows the animal to identify the new position, and subsequent trials on that same day test the rat's memory for that location. While 2 trial sets/day are adequate to observe improved performance, additional trials may be

used to provide a more robust evaluation of working memory.

3.2. Standard Test Conditions for the Morris Water Maze

As with simple position discrimination learning, all standard test conditions described below must be maintained for appropriate application of the MWM. Additional requirements are specific to the MWM as detailed below. Standardization of maze size is recommended and a minimal diameter is critical. Small mazes are simpler to solve and the dynamic range for detecting suboptimal task resolution is limited. Larger mazes provide a more challenging spatial task, and represent more sensitive means to detect learning and memory impairments (Vorhees and Williams, 2014a).

- Maze size and water depth and temperature should be specified, within acceptable limits, and appropriate to the age and size of the animal being tested;
- The submerged escape platform is not visible;
- Positioning of the animal at the beginning of each trial should be done with the head facing the wall of the tank and the start point chosen randomly over trials;
- The duration of time that the animal is left on the escape platform after each trial must be consistent and specified;
- Intertrial interval (the delay between removing the animal from the platform and the beginning of the next trial) and handling and holding conditions between trials (where the animal is placed between trials, if the animal is towel dried, if the animal is warmed) must be consistent and specified;
- Stable and comparable performance in the acquisition phase is essential for appropriate interpretation of learning;
- Performance control procedures including swim speed and cued learning trials should be incorporated.

3.3. Metrics to Evaluate Learning and Retention in the Morris Water Maze

There are a number of dependent measures that are evaluated for MWM learning, but submitted studies typically include assessment of latency and path length to find the hidden platform, and search time in the correct quadrant on probe trials. Video tracking provides data on numerous variables, whereas fewer measures are possible if data are collected by a technician with a stopwatch. The most common variables are described in more detail below.

Latency to Escape over Trials.
This measure reflects the time from release of the animal in the maze until it reaches the escape platform on each trial. Note that latency is to some extent dependent on the size of the maze and thus varies from laboratory to laboratory. A decrease in latency over the course of training provides evidence of learning.

Path length to Locate Hidden Platform.
Path length is the distance travelled to find the platform on each trial. As with latency, a decrease is evidence of learning. This measure is highly correlated with escape latency, but note that there is no upper cut-off for path length as there is for latency (maximum trial duration).

Search Parameters during Retention Probe Trial.
Time spent searching in the correct quadrant as well as the number of passes over the platform position are typical measures assessed during probe trials.

3.4. Data Reporting for the Morris Water Maze

In order to evaluate the effects of chemicals on these parameters, it is essential that MWM reports provide clear information on study design, the measures (dependent variables), software algorithms that define the variables, and the statistical procedures used to evaluate the main results. The main results that are of interest from DNT learning and memory assessments are:

- The effects of chemicals on acquisition, i.e., initial learning rate;
- Effects of chemicals on memory, i.e., retention of the learned response;
- Visual deficits;
- Swim Speed;
- Sex differences in effects of chemicals on learning and/or retention;
- Preexisting differences across groups, evaluated by data from the first trial or first day of testing.

Dependent Variables
It is important that the dependent variables recorded and analyzed are unambiguously described. Averaging data across trials within a single day into daily blocks may be appropriate, especially in instances where trial and block are included in the statistical analyses and there are no significant interactions therein. For MWM learning, the dependent measures may include the following:

- Latency to reach the escape platform, and how this changes over trials;
- Path length (distance travelled) to reach the escape platform and how this changes over trials;
- Percent of time in each quadrant on probe trial (s);
- Number of platform crossings on probe trial (s);
- Latency to reach escape platform that is clearly visible to the animal (to control for sensory integrity that could influence maze performance independently of spatial learning);
- Swim speed (to control for motor impairments that may impact escape latencies independently of spatial learning).

Reported Data
In order to evaluate treatment effects, the following level of data detail should be reported for each sex at each age group. In cases where data are not normally distributed, different measures of central tendency may be more appropriate (e.g., median, mode). Data should be reported separately for acquisition and retention assessments.

- Mean ± standard deviations of latency per treatment group on each trial;
- Mean ± standard deviations of path length per treatment group on each trial;
- Mean ± standard deviations of percent of time spent in each quadrant on probe trial per treatment group;

- Mean ± standard deviations of number of platform crossings per treatment group on probe trial;
- Mean ± standard deviations of latency and path length per treatment group on cue learning trials;
- Mean ± standard deviations of swim speed per treatment group on probe trial;
- Individual trial data for each animal should be included in an appendix;
- Summary (means and standard deviations) should be presented in tabular and in graphic form.

All data should be reported both on an individual and group basis by trial and by day over treatment groups. Individual data should include every trial performed for each animal, including cue learning trials when available.

3.5. Interpretation of Morris Water Maze Results

The following points should be taken into consideration in the interpretation of the data:

- Control animals should demonstrate a clear and statistically reliable reduction in both latencies and path length over trials. The absence of these properties in control animals indicates lack of a learning curve and as such the test does not satisfy the minimal requirements to detect a change in behavior over time and does not measure learning;
- Control performance on a single probe trial run 24 hours after final acquisition must show an increase in percent time in the correct quadrant that exceeds chance levels of 25%;
- A treatment-related decline in performance on a single probe trial 24 hours after final acquisition can only be interpreted as a retention deficit if comparable levels of learning have been achieved in all treatment groups on the final day of acquisition;
- First trial data should be closely examined to determine the potential presence of a pre-existing motoric difference among treatment groups that may skew learning curves;
- Measures of sensory and motoric function should be assessed; for example, slower swim speed producing longer latencies may be misinterpreted as a learning deficit;
- Variability should be within a reasonable range as described below in "Interpretation of Treatment-Related Effects".

4. Test Description: Associative Learning using Passive Avoidance

Unlike most of the other learning tasks currently used in the guideline DNT study, passive avoidance (PA) is not a water-based maze. PA was originally designed as a task to investigate the impact of various interventions on memory consolidation. It was specifically designed as a type of one-trial learning, so that the relationship between the timing of the intervention (with respect to the timing of the learning and retention trials) and task performance could be evaluated. This led to an improved understanding of the process of memory consolidation, in that interference (e.g., by drug administration) at various times before or after initial learning would lead to different impacts on whether or not the animal remembered the experience (Ehman and Moser, 2006; Peele et al., 1990). It is also a very useful way to evaluate whether or not a drug or chemical has a specific effect on one aspect or another of learning or memory, again based on evaluating response following different types of chemical administration paradigms.

The basic procedure involves a device with two connected chambers, separated by a barrier

(usually an opening covered by a sliding door) (Jarvik and Kopp, 1967). During the training or acquisition phase, the test subject is placed into one compartment, a bright light is turned on and the barrier removed allowing the animal to enter the other chamber which remains darkened (see Figure 6). Since the bright light can be aversive to rodents that prefer the dark, the test subject is motivated to quickly enter the dark chamber to escape from the light. Upon entering the dark chamber, the door is closed and a mild footshock delivered to the animal. This single pairing between the dark chamber and the occurrence of footshock is usually sufficient to ensure that the animal will avoid the dark chamber in the future. On the second trial (accurately defined as the first retention trial), the animal is placed back in the original chamber (lighted side) and the time to enter the dark side (previously associated with footshock) is measured. Typically, footshock is not administered on the retention trial. The latency to enter the shock-associated side (step-through latency) provides a measure of avoidance learning. This task is referred to as "passive" avoidance since avoiding the shock-associated dark chamber does not require any active response on the part of the animal, as it just remains in the original lighted chamber. Depending on the strength of the aversive stimulus (shock magnitude and duration), rats typically learn this task following a single trial, avoiding entry to the dark chamber on all subsequent trials. Subsequent retention or "memory" of the aversiveness of the shock-associated dark chamber is measured by again placing the animal in the lighted non-shock side of the same device following a specific interval (varying from 1 hour up to 1 week for most DNT studies).

Because of its relative ease of use and the large neurotoxicology literature for this task, PA has become one of the most frequently submitted learning and memory paradigms for guideline DNT studies. In multiple laboratories, it has been used at one or both time points (weanling and adult), as the required learning and memory evaluation. However, the specific procedure used in DNT studies varies widely from laboratory to laboratory, and differences in procedure have the potential to greatly affect the sensitivity of the task to detect treatment-related memory impairments.

Figure 6. Passive Avoidance Chamber. Animals are placed in lighted side of the chamber and receive a mild foot shock when they enter the preferred darkened side of the chamber. The ability to suppress the motivation to move to the dark side of the chamber is assessed the following day and taken as a measure of memory.

4.1. Procedures for Passive Avoidance

Initial Acquisition

Typically animals are placed on one side of a two-chamber box that is brightly illuminated, and allowed to acclimate for a period of 20-60 seconds. A door is raised to allow access to the darkened side of the chamber, and animals are allowed to move from the lighted area towards the darkness. Upon entry to the dark chamber, the door between the light and dark sides is closed and a mild brief foot shock is delivered. The animal is then removed from the chamber and returned to the home cage until retention testing. Since this task is designated primarily as a memory test with one-trial learning, there is no metric for 'acquisition'. Despite this, several laboratories have adopted a multiple trial 'acquisition' phase, perhaps due to a misinterpretation of guideline specification of learning defined as "a change in performance over time". Although this guideline specification was not intended to apply to single-trial learning procedures like PA, a number of repeated-trials variants of this task have appeared in which the animal is immediately (within 30-60 seconds) placed back into the lighted side of the chamber following foot shock delivery, with repetition of this procedure until some criterion is met, e.g., two consecutive trials in which the animal refrains from entering the darkened chamber for a specified duration. Since this procedure is not typical in the published literature, its utility and interpretation are unclear.

One variant of a repeated-trials procedure with merit has been described by Peele et al. (1990) in which daily trials were administered until a criterion of two consecutive trials (days) of latencies exceeding the cut-off criterion of $\geq$ 600 sec was achieved. An increased number of trials to criterion was demonstrated in animals exposed to iminodipropionitrile (mean of 5 vs 2.7 days to criterion), in addition to deficits in single trial PA conditioning. In this way, single trial PA learning was used to assess 'retention' of PA learning, while in a second set of animals, the repeated trial PA protocol (with trials separated by 24 hour intervals) was used to address 'acquisition'. Unlike the study of Peele et al. (1990), submitted guideline DNT studies have used relatively brief criterion cut-off times (60-180 seconds versus 600 seconds) for entry into the darkened chamber.

In addition to a wide range of number of trials, inter-trial interval, and avoidance cut-off criteria, various shock durations and shock intensities have also been employed. A 1 mA 1-second scrambled foot shock delivered once all four paws have entered the darkened chamber is typical in the literature. However, submitted studies have incorporated durations of 1-5 sec, and shock intensities $\leq$ 1mA. It should be noted that both intensity and duration of the foot shock will influence memory retention (Ader et al., 1972). In some laboratories, different shock durations and intensities have also been used for young versus adult animals. All of these parameters should be considered when evaluating and interpreting PA data.

Retention testing

As described above, the primary metric in PA testing is the measure of retention (memory). Following delivery of foot shock in the initial training phase, animals are reintroduced to the same chamber after a delay under conditions identical to those used during training, except no shock is applied. The latency to enter the dark side of the chamber (shock-associated side) provides the main measure of memory. Typically retention trial cut-offs are 60 or 180 sec in duration, whereas in the study cited above, a criterion of 600 sec was effective in detecting treatment-induced deficits. Short duration retention trials can reduce the sensitivity of the task to assess treatment-induced memory impairments. Because PA is a very effective procedure to invoke avoidance, normally a single acquisition trial is sufficient to produce significant

avoidance of the shock-associated dark side of the chamber. However, shock intensity and duration can influence the efficacy of this single-trial learning to produce avoidance. Typically, retention is reported as 1 or 2 short-term trials (1- 6 hour delays following initial training) administered on the same day of training. Long-term retention is assessed with longer intervals between training and retention testing, typically 1 or 7 days post-training. Latency to enter the darkened side of the chamber constitutes the measure of retention. It is compared to the latency to initial entrance before administration of the first foot shock (i.e., the training trial).

As noted above, some laboratories have employed a multiple training trial approach where more than one acquisition trial is employed (i.e., shock in dark chamber is applied more than once). This procedure complicates the interpretation of memory results because most animals will avoid the shock side of the chamber after only one trial. Because the second acquisition trial is conducted exactly as the first acquisition trial, the subset of animals that do enter the shock-associated dark side of the chamber is actually providing a measure of retention. On the other hand, animals receiving a second shock-dark side pairing will be the subset of animals that have not learned to avoid the chamber after the first shock acquisition trial. While this second shock pairing will improve avoidance, animals receiving two shock-dark side pairings will have different training experience compared to animals that learned to avoid the shock side after only one acquisition trial. It should also be noted that since any second acquisition trial is actually also the first retention trial, the interval between the first and second acquisition trials is also the first retention interval. Because of this, care should be exercised in interpreting retention data where multiple training trials have been employed.

4.2. Metrics to Evaluate Learning and Retention in the Passive Avoidance Task

Results from this test are usually reported as latency to re-enter the area where shock was received on the training trial (s). Various cut-off criteria have been adopted as described above. Errors have been defined as failing to inhibit the response to enter the darkened chamber. Trials repeated within a given retention session until 2 (or more) consecutive 'criterion' trials are reached. As described above, the specific criteria used vary among laboratories.

Latency

Latency may be reported as step-through latency on the first trial at each retention interval and/or as mean latency across all trials at a given retention interval. The latencies for the retention trials are compared to latency to enter the dark side of the chamber on the first trial of the initial training day.

Trials to Criterion

If a criterion performance is specified (e.g., 2 consecutive trials with >60 sec latency to enter the darkened chamber), the number of trials to attain this are reported for each retention interval assessed.

4.3. Data Reporting for the Passive Avoidance Task

In order to evaluate the effects of chemicals on these parameters, it is essential that study reports provide clear information on study design, shock parameters, the measures taken (dependent variables), and the statistical procedures used to evaluate the main results. The main results that are of interest from PA learning and memory assessments are:

- Effects of chemicals on memory, i.e., retention of the learned response evidenced by increased step-through latency;
- Step-through latency on the first trial, which can show pre-training motoric deficits or activity differences;
- Sex differences in effects of chemicals on learning or retention.

Dependent Variables

The dependent measure in PA learning is latency to enter the chamber after the initial training trial (i.e., on retention trials). This is typically assessed on the same day of training with an interval ranging from 1-6 hours, and again 24 hours later.

Reported Data

In order to evaluate treatment effects, the following level of data detail should be reported for each sex at each age group. Results should be provided separately for each trial, and for the acquisition and retention phases. It is likely that data from PA procedures will not be normally distributed, with a high proportion of censored trials (i.e., maximal values of 60 or 180 sec entered for retention trials). In these cases different measures of central tendency may be more appropriate (e.g., median, mode) and nonparametric statistical procedures should be applied (see Section 9). Data reported must include:

- Mean ± standard deviations of latency per treatment group on each trial;
- Individual trial data for each animal, included in an appendix;
- Summary (means and standard deviations), presented in tabular and in graphic form.

All data should be reported on an individual and group basis. Individual data should include every trial performed for each animal. For this procedure, it is not possible to evaluate the study if data from the first acquisition trial are not reported.

4.4. Interpretation of Passive Avoidance Results

The following points should be taken into consideration in the interpretation of the data:

- All animals must cross into the darkened chamber and receive a foot shock on the first training trial in the acquisition phase in order to be included in the analysis. Animals that fail to cross must be excluded from retention assessments since it is not possible to test retention in the absence of learning. If a disproportionate number of animals among the study groups fail to cross on the first trial, the data will be biased and the test results will not provide a valid assessment of learning. The reviewer needs to examine the reported data carefully to ensure that both of these conditions are satisfied and appropriately reflected in results tables;
- With highly variable data on retention trials, one potentially useful analysis that the reviewer could perform, or request be performed by the reporting laboratory, is application of a binning criterion to evaluate latencies, e.g., number of animals in each group that crossed in 15 seconds or less; 30 seconds or less; 45 seconds or less; >60 seconds. Such an analysis may facilitate the comparison of distributions of latencies across treatments and provide a more meaningful context of the magnitude of the differences;

- If a high proportion of animals from all study groups cross into the darkened chamber on the first trial, but a difference in latency exists between treated animals and controls, this may complicate the interpretation of retention deficits;
- Rats must be able to feel the foot shock, or there will be no aversion. A test of shock sensitivity (for pain and reactivity) could be performed to rule out sensory deficits;
- Control animals should demonstrate a clear and statistically reliable increase in latency to cross on the first retention trial (i.e., the trial that is assessed during the training phase, Trial 2), relative to the training trial (i.e., Trial 1);
- In this very robust learning task, latencies in control animals on the subsequent retention trials typically remain close to that observed in the first retention test, but may diminish with repeated exposures and very long intervals;
- Variability should be within a reasonable range as described below in Section 10 "Interpretation of Treatment Related Effects".

5. Test Description: Sequential Learning using the Biel or Cincinnati Maze

Cincinnati Maze

Figure 7. Cincinnati Maze. The animal must make a series of right or left turns to navigate from the start point to the escape platform at the end of the labyrinth. Both forward and reverse paths may be tested. The reverse path is not the mirror opposite of the forward task.

The Biel and Cincinnati mazes are tests of sequential learning, and by that nature, are complex. Both have been used to evaluate the cognitive effects of chemicals on the developing nervous system, and they have been shown to be sensitive to a variety of developmentally neurotoxic substances (Graham et al., 2012; Paul et al., 2009; Vorhees, 1987; Vorhees and Williams, 2016; Williams et al., 2002). Although the mazes have been in use for many years, the type of learning required for the task and the brain systems involved in achieving good performance have not been as thoroughly investigated as is the case for position discrimination (the letter mazes) and spatial learning (as tested in the MWM). Similar to position discrimination and the MWM discussed previously, both the Biel and the Cincinnati mazes are typically run as water mazes, with motivation being provided by the opportunity to escape from the water once the final goal is reached.

The Biel and Cincinnati mazes can be appropriately described as multiple-T mazes (figure 8), in which the animal is placed into the water at a defined start location and is required to remember the correct sequence of right or left turns as a series of 'choice points' (most recently reviewed in Vorhees and Williams, 2016). The Cincinnati maze was adapted from the Biel maze, expanding and increasing the complexity by incorporating a larger number of choice points (9 vs 6), while at the same time preserving the asymmetry of left vs right turns to reach the goal and escape from the maze. During training, each incorrect choice can potentially result in multiple errors, both

within the blind alleys and when the animal returns to the main path of the maze. Because of this, the number of errors made during acquisition frequently exceeds the number of choice points.

As is the case with other types of mazes, testing procedures for the Biel and Cincinnati mazes often vary among laboratories. Both mazes have been used wherein the starting point and goals are moved creating different "paths". Figure 8 shows path A and B for the Cincinnati maze. In path A, each choice takes place at a blind "T", or dead-end, such that there is a 50:50 chance of turning correctly. However, correct turns in path B are located before each blind "T", such that turns made only at dead-ends are never successful. For this reason, path B has been shown to be more difficult, and potentially more sensitive, than path A (Vorhees and Williams, 2016). In most instances, there will be two acquisition phases for this type of maze. After learning on one path, the goal and starting points of the maze will be switched and the animal will learn the maze in the second path. Thus, in addition to evaluating the animals' ability to learn a path through the maze, the test will also evaluate an animal's ability to recognize that the correct choice has changed and to respond by changing their behavior and learning a new path. This adds an additional level of complexity to the task and thus increases its sensitivity to detecting cognitive deficits, in that animals not impaired on the initial learning phase may be impaired in the 'reversal' task. Long term retention can also be evaluated by retesting the animals on one or both paths after several days with no testing.

5.1. Procedures for the Biel or Cincinnati Maze

Initial Acquisition

Prior to acquisition, the animal must be acclimated and its swimming ability evaluated during several trials in a straight channel. This also allows opportunity for the rat to learn that there is a platform for escape. On the following day, the animal is placed in the maze at the start point, facing toward the wall. Once placed in the maze, the animal navigates the various choice points until it reaches the goal arm that contains a hidden platform allowing escape from the water.

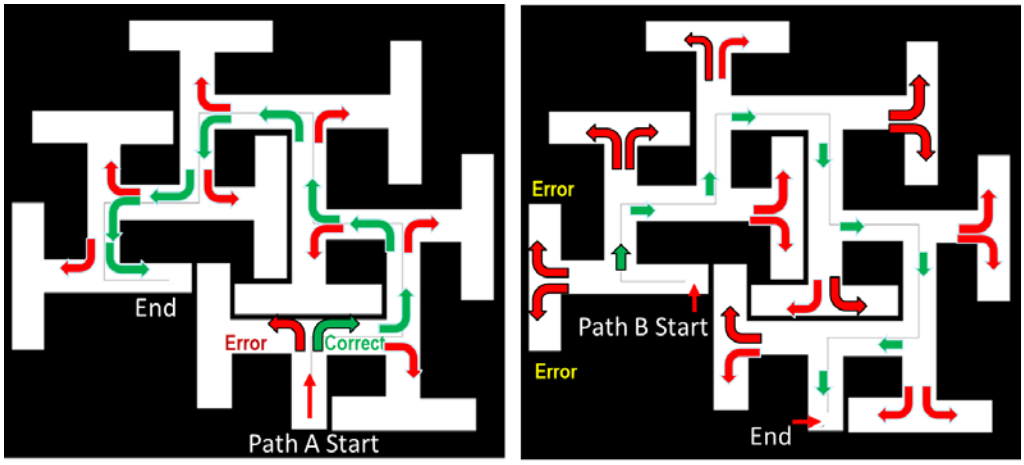


Figure 8. Path B is not the converse of Path A in the Cincinnati or the Biel Maze. In Path A of this Cincinnati maze, the animal is forced to go left or right at the cross point of the 'T'. In Path B, there are always more than a simple left/right decision to be made and more blind alley entries

as described above. Path B is not the direct opposite of Path A and as such, does not constitute true 'reversal learning'. (Vorhees and Williams, 2016).

Both errors and latency are recorded during the trial. Each incorrect entry should be counted as an error, and a clear definition of errors should be provided. For each trial, a maximum latency (trial duration) is defined, typically 3-5 minutes. Correction procedures may vary among laboratories, but generally the animal will be guided through the correct path to the goal of the maze, at which point it is removed from the water. The number of trials per day may also vary (usually 2-4), as may the number of days of training (2-4) and the inter-trial interval (at least 5 minutes, to prevent over-fatigue). Although it is reasonable for these parameters to vary among laboratories, each laboratory should have a defined procedure that is consistently followed for all animals tested in that study.

'Reversal' acquisition:
After the initial path has been learned (i.e., met criterion), or after a set number of trials has elapsed (regardless of performance), the start point and goal of the maze can be reversed, and the ability of the animal to learn the new path evaluated. This is technically not a 'reversal' trial as the animal does not simply have to make the opposite choice at each choice point. Rather, the path for initial learning is distinct from the path for this second learning phase. However, general procedures for this second phase of acquisition should be identical to those used in the first phase, with the same number of trials per day, run under similar conditions. In some cases, the second acquisition phase may include additional days of training, as there is usually a large increase in errors during the first few acquisition trials in the 'new' path as the animal initially applies previously learned responses to the changed situation.

Retention testing:
After the second acquisition phase, the animal may be re-tested on the initial path, in order to determine retention. This re-testing may occur immediately following the second acquisition phase, or there may be a delay of several days between the acquisition and retention testing. In some cases, retention for both maze paths may be evaluated.

5.2. Metrics to Evaluate Learning and Retention in the Biel or Cincinnati Maze

Results from this test are usually reported as latency to reach goal and number of errors per trial.

Latency
Latency should be measured from the time the animal is placed in the start area until the animal enters the designated goal area. Similar concerns regarding latency measures as described above for position discrimination and the MWM also apply to data from the Cincinnati/Biel maze. In multiple choice mazes in particular, however, latency data seldom reflect error performance in a one-to-one fashion. This most likely results from animals deliberating at each choice point before committing to a specific path as learning progresses, resulting in longer latencies in the absence of higher errors. As in other less complex water-based mazes, swim speed can also affect latency measures and potentially confound interpretation.

Errors
Error scores are generally reported as the total number of errors per trial. Unlike the position discrimination maze (for which only one error per trial is generally reported), the number of

potential errors per trial in the Biel/Cincinnati maze is limited only by the time limit imposed on each individual trial (the animal must decide among two directional options at each choice point, and some choices may also lead to back-tracking through the maze). Although the pattern of errors made by each individual animal within the maze could potentially provide useful information, DNT study reports generally include only the total number of errors per trial. There may be an occasional animal that stops searching and swims in a small area for the rest of the trial, and low error rates would be recorded. Averaging group data with these low rates could underestimate a real deficit. The incidence of such cases should be reported and evaluated as a potential treatment effect. It is important that a clear definition be provided of what constitutes an error and how errors cumulate over choice points within a given trial.

Errors may be averaged within days (which will decrease the variability in results); however, errors should also be reported for each individual trial, as it is useful to evaluate changes in the number of errors both within and between days. This analysis can provide information regarding the animal's memory for different retention intervals. Due to the design of the evaluations using the Cincinnati/Biel maze in a typical DNT study, the animal is tested for a defined number of trials. Thus, there is no measure analogous to the trials to criterion or errors to criterion metric used for position discrimination protocols.

5.3. Data Reporting for the Biel or Cincinnati Maze

In order to evaluate the effects of chemicals on learning, it is essential that study reports provide clear information on study design, the measures (i.e., clear operational definition of the dependent variables), and the statistical procedures used to evaluate the results. The main results that are of interest from Cincinnati/Biel maze learning and memory assessments are:

- The effects of chemicals on acquisition, i.e., initial learning rate;
- Effects of chemicals on memory, i.e., retention of the learned response;
- If more than one path is tested, effects of the chemical on the animal's ability to adapt its response and learn a new one (i.e., comparison of learning for the initial path and the second path);
- Effects of the chemical on swim speed;
- Sex differences in effects of chemicals on learning and/or retention.

Dependent Variables

The dependent measures include the following:

- Latency to reach the designated goal, and how this changes over trials;
- Total number of errors for each trial;
- Swim speed (to control for motor impairments that may impact escape latencies independently of spatial learning).

Reported Data

In order to evaluate treatment effects, the following level of data detail should be reported for each sex of each age group. Results should be provided separately for each maze path tested, including the order in which the paths were evaluated (and for any repetition of paths tested again at the end of the study). In cases where data are not normally distributed, different measures of central tendency may be more appropriate (e.g., median, mode).

- Mean ± standard deviations of errors per treatment group on each trial;
- Mean ± standard deviations of latency per treatment group on each trial;
- Mean ± standard deviations of errors per treatment group by day;
- Mean ± standard deviations of latency per treatment group by day;
- Mean ± standard deviations of swim speed per treatment group on straight alley trial;
- Individual trial data for each animal should be included in an appendix;
- Summary data (means and standard deviations) should be presented in tabular and in graphic form.

All data should be reported both on an individual and group basis and by day over treatment groups. Individual data should include every trial performed for each animal. Data should also be labeled by maze path, and the trial on which the path changed should be clearly indicated.

5.4. Interpretation of Biel or Cincinnati Maze Results

The following points should be taken into consideration in the interpretation of the data:

- As described for other test procedures, control animals should demonstrate a clear and statistically reliable reduction in both number of errors and latencies over trials; this pattern should be demonstrated for both the initial maze path and the reversal path, but the shape of the learning curve will differ in the two phases. The absence of these properties in control animals indicates lack of a 'learning curve' and as such the test does not contain the minimal requirements to detect a change in behavior over time and does not measure learning;
- The task requirement for change in behavior during the second phase is a more complex form of reversal learning among common DNT learning and memory procedures. Control performance on the initial trials of the second path should show an increase in the number of errors over both: 1) the number of errors in the last trial of the initial path; and 2) the number of errors in the first trial of the initial path. The increase in errors is typically followed by a rapid (perhaps more rapid than during initial learning) decrease in errors during the later trials (Figure 9);
- Evaluation of performance by treated animals in the second phase is confounded when similar levels of performance are not demonstrated on the first phase (i.e., if treated animals do not learn the first path, they may not be impaired in the early trials of the second path). If this occurs in a small portion of animals, performing an additional analysis with their exclusion may be informative;
- Variability should be within a reasonable range as described below in Section 10 "Interpretation of Treatment Related Effects".

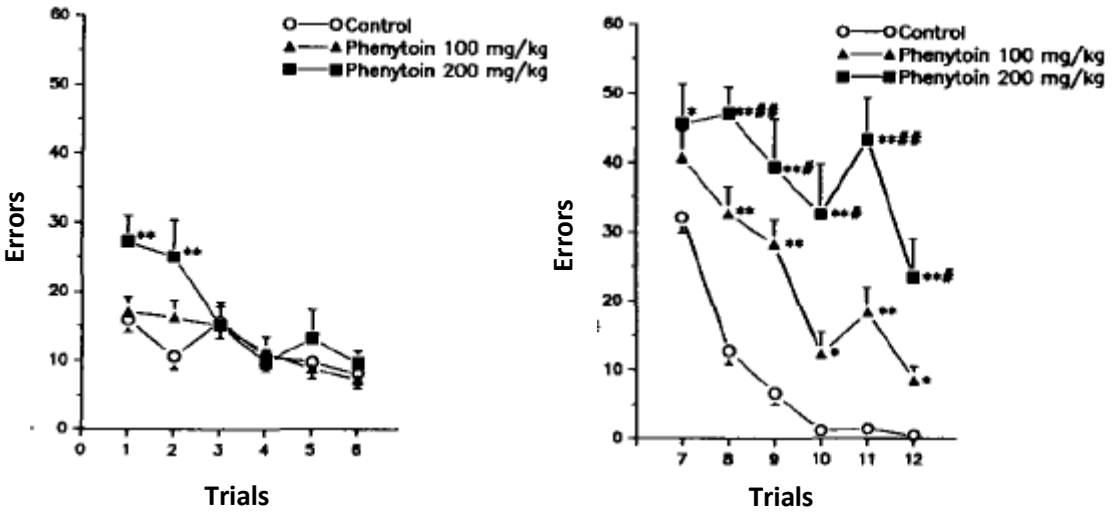


Figure 9. Mean (± SEM) number errors on Path A (left) in the first 6 trials and Path B (right) for final 6 trials in animals developmentally exposed to phenytoin. Errors decreased over trials in both versions of the task, and fewer numbers of errors were committed on Path A (left panel) than the more difficult Path B (right panel). Developmental exposure to phenytoin impaired learning in both versions of the task, but effects were greatly exacerbated under conditions of Path B (Vorhees et al., 1991).

As with the MWM reversal learning described above, treatment-related effects could be manifested in two different ways, depending on how the animal's behavior is changed by the chemical: a) the treated animal could exhibit fewer errors on the reversal than the control animal, due to poorer retention of the initial task (and thus decreased 'interference' as the animal tries to implement a previously learned response in the newly changed conditions) or b) the treated animal could exhibit more errors on the reversal than the control animal, due to an inability to adjust its response in the new situation (i.e., it may continue to respond using the old choice patterns, for a much longer time, due to its inability to recognize that there has been a change in the test conditions and/or its inability to alter its response patterns or learn new response patterns in response to that change in conditions).

The reviewer will need to evaluate data from treated animals in light of these potentially complex response patterns, realizing that impairment can be demonstrated in multiple types of performance, and that the demonstrated impairment pattern could potentially vary among treatment groups.

6. Standard Conditions for All Tests of Learning and Memory

In addition to the multitude of environmental factors (e.g., lighting, noise) that may alter behavior in general (described in Slikker et al., 2005), a number of additional factors can specifically influence assessments of learning and memory. Many of these are common across all behavioral assessments, others are specific to learning and memory, and still others are unique to specific learning and memory tests. The procedures used by the testing laboratory to control these factors should be adequately described in the protocol or experimental design or reported in the results. These include:

- All testing should be performed in a quiet environment with minimal exposure to outside

personnel traffic;

- Temperature, humidity, and lighting conditions in the test area, as well as time of testing relative to light-dark cycle, should be kept constant;
- Overhead lighting, positioning of experimenter relative to the maze, olfactory stimuli (i.e., presence of other animals in the test room) should be controlled, and olfactory and visual cues in/on the maze itself (intramaze cues) should be eliminated. This promotes reliance on extramaze cues and diminishes animal's reliance on a position or cued learning strategy for spatially-based (i.e., MWM, position discrimination) or sequential learning tasks (i.e., Biel and Cincinnati mazes);
- If several mazes are used to increase throughput of testing, treatment groups should be counterbalanced over specific test devices and time of day. Care should be taken to ensure there are no differences in visual perspective among mazes;
- With repeated testing, the animal should always be tested in the same maze and at the same time of day;
- Standardized procedures (chemicals and drying time) for cleaning mazes should be employed, including procedures used between trials, between test animals, and between test days;
- In water mazes, water temperature should be constant across days and neither too warm so as to reduce motivation to escape, or too cold so as to induce hypothermic stress;
- Maze size (alley length, width, wall height, and water depth) should be specified, within acceptable limits, and should be appropriate to the age and size of the animal being tested;
- Positioning of animal in the start position, confinement of animal in a start box or simple release in the start alley should be consistent and specified;
- Correction procedure in event of an error must be described (e.g., if the animal is guided to the correct goal, removed and placed on the goal, or if some other procedure is used);
- Operational definition of correct response vs error should be clearly articulated (e.g., how far does the animal need to go down the incorrect alley in order for an error to be recorded [e.g., head only, nose only, head and shoulders, entire body]);
- Duration of time animal is left in the correct goal box/escape platform on completion of each trial must be consistent and specified;
- Intertrial interval (the delay between removing the animal from the goal and returning the animal to the start box for the next trial) and holding conditions between trials must be consistent and specified (e.g., where the animal is placed, presence of a heat source);
- Training of all subjects to the same level of 'criterion' performance must be specified;
- Wherever data are collected by personnel (as opposed to video tracking, for example), assurances should be made that the observer is blind as to the animal's treatment. In such cases of multiple observers, some measure of inter-observer reliability should be provided.

7. Control Data for Tests of Learning and Memory.

A number of behavioral properties in the measured data must be evident and can serve as a reference point for evaluating reported results.

- Control animals should demonstrate a clear and statistically reliable reduction in both latencies and errors over learning trials in escape-based maze tasks. For PA learning,

latency to enter the darkened chamber should increase and be maintained well above those recorded in the initial trial;

- Animals should attain criterion performance at a faster rate for retention trials compared to the original acquisition learning;
- Variability should be within a reasonable range, based on data from the test laboratory as well as the literature;
- Historical control data may also be useful for interpreting potential treatment effects; however, such data must represent studies using all the exact same experimental procedures. While historical control data can be supportive of conclusions, the most appropriate comparison for interpretation and statistical analyses is the concurrent control group.

To demonstrate that the task is sensitive to detect learning impairments related to chemical exposure, positive control data using a chemical (preferable) or other insult (e.g., physical brain lesion) for which learning deficits have been established should be provided, using exactly the same apparatus, test procedures, and testing conditions as the reported study. The positive control study should have been conducted within a reasonable time frame relative to the reported study, to minimize experimental drift over time. The positive control data presentation should be sufficiently complete to permit evaluation of the sensitivity of the method, including individual data and measures of variability. The positive control study should preferably include a dose-response curve, and exposures should include sufficient internal controls to document that impaired performance is due to changes in learning and memory, not to other effects of the positive control treatment (e.g., changes in motor function or nonspecific physiological reactions to the chemical, which may result from high doses of the same substance).

8. Statistical Analyses

There are many considerations necessary for the appropriate analysis of learning and memory data. Guidance for analysis of DNT data can be found in Holson et al. (2008). For all developmental toxicity studies, the unit of analysis is the litter. That is, littermates cannot be analyzed as if they were independent observations. If the experimental design is such that multiple pups from a litter are tested, the sample size for statistical analysis is the number of litters (not number of total pups). Where a male and female from the same litter are tested, sex must be nested within the litter or used as a random factor. Only in step-down analyses following a significant main effect or interaction with sex can males and females be analyzed separately. Typically, interactions should be tested first, e.g., dose-by-time interactions. This situation applies for all dependent variables. While general linear model ANOVA is historically used for most analyses, homogeneity of variance across treatment groups and/or test times should be noted. In many cases, mixed-effects models allows application of the appropriate covariance matrix that best fits the data.

8.1. Statistical Analysis of Learning Data

Acquisition (i.e., learning) is typically measured over a number of test sessions, with improved performance over repeated testing (or training) being an indicator that the animal has learned the task. In cases where measures are made repeatedly (e.g., latency to find a platform on each trial over days), the within-subject data should be analyzed using repeated-measures analyses. For such data, the overall ANOVA would include the effects of treatment, sex (within litter if

appropriate), and test session (e.g., days or trials), and potential interactions among these. A simple approach to establish that learning has occurred would be a significant change over repeated testing (i.e., significant effect of test session), indicating that the performance has changed (in the appropriate direction). Examples would include decreased latencies to escape water in the MWM (see Figure 4) or decreased number of choice errors in the Biel maze (see Figure 9). Applying such a test to data from the control animals should reveal that the control group has learned the task. This is especially important because it is not possible to unambiguously determine the meaning of any effects associated with chemical exposure if control animals did not learn the task. Assuming that controls do indeed learn the task, an overall treatment effect or interactions with sex and/or dose then can be interpreted as an effect of the treatment. Typically, a significant effect of treatment on learning is seen as a significant treatment-by-time (or day) interaction. Note that simply comparing data from control group animals with treated animals at each test session separately, in the absence of a treatment-by-test session interaction, is not appropriate.

Learning may also be evaluated as a specific endpoint instead of repeated measurements of performance. For example, learning may be measured using trials to criterion or days to criterion instead of latency values from each test session. Since this approach results in only one measure of learning (e.g., days to criterion), a repeated measures ANOVA cannot be used for data analysis and a simple factorial ANOVA would be appropriate using sex and treatment as factors in data analysis.

8.2. Statistical Analysis of Memory Data

Memory may be evaluated in a single trial after learning has been established, and in those cases there are no repeated measures. For example, probe trials typically used in the MWM are normally one trial, measuring time spent in the platform area. Since this measure is typically collected only once for memory testing, repeated measures ANOVA cannot be used. Where there are repeated memory trials (e.g., PA testing on multiple days after shock), the use of repeated measures is necessary.

8.3. Data Distributions

It is important to recognize that the properties of the measurement data for learning (or memory) influence the type of statistical analysis procedure that is appropriate. Parametric statistics, such as ANOVA, require continuous data as the measure (e.g., latency values, distance, etc.) and assumes that the data is normally distributed. There are a number of statistical tests available that evaluate the normality of data sets. Count data (e.g., number of errors) are not continuous and may not be normally distributed; however, while care should be exercised in how such data are analyzed, ANOVA may be appropriate nonetheless. It should be noted that ANOVA is relatively robust for modest departures from normality as well as homogeneity of variance, and may still be used in many cases. However, inappropriately analyzing distributions with marked deviations from normality using ANOVA can underestimate treatment effects. In some cases, transformation (e.g., square root) of the data may improve underlying distributions and variance.

Non-parametric analyses, e.g., Kruskal-Wallis test, may be used in cases of non-continuous data or undefined distributions. For example, learning measurements may be non-normally distributed in situations which trials-to-criterion is used as the dependent variable, and

nonparametric statistics should be used. Visual inspection of the data should be used as well as these tests.

Deviations from normal distribution can be created in some assessments of learning and memory due to a cut-off criterion which generates closed-ended distributions – these are considered censored data. This is especially of concern when the cut-off is actually the value used as the endpoint of learning or memory. For example, maze procedures usually set a maximum trial length (e.g. 60 sec) and a value of 60 sec is assigned to animals that fail to reach the escape platform within 60 sec. While convenient, this 60 sec value actually underestimates the time required for the animal to reach the platform (the actual value is unknown but is greater than 60 sec) and reduces the estimates of variability of the response. This results in a closed-ended distribution where data values beyond the maximum trial length are not measured. One simple but intuitive method for approaching the latency data from a study using PA is provided by the Kaplan–Meier (KM) method of estimating a survival curve as depicted in Figure 10. With letter mazes such as the M-maze, there are often a high number of trials with zero errors, which leads to closed-end data that is left-censored and zero-centric. Low numbers in datasets such as these are not appropriately analyzed by ANOVA, and require more specific analyses (see EFSA 2010 for example with Biel maze data).

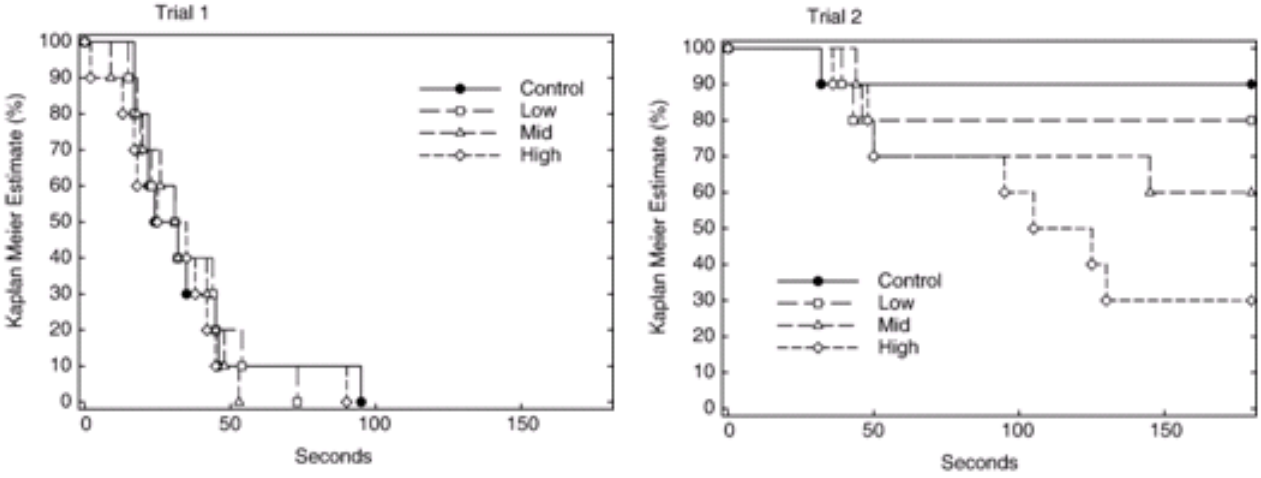


Figure 10. The KM method estimates the probability that the time-to-crossover is greater than a specific time t. For those values that are not censored, the KM estimate is simply the proportion of animals that have not crossed over by time t. For censored values KM is undefined. The KM estimates can then be compared among groups using either the log-rank test or generalized version of the Wilcoxon test (for example, SAS PROC LIFETEST). In these graphs, trial 1 is the cross-over latency before shock. Trial 2, 24 hr later, shows the latency across treatment groups. Most of the controls did not enter the shock compartment, whereas a number of treated rats did. In this example, a conventional analysis suggested that there is no treatment effect, while the Kaplan–Meier approach using log-rank survival tests showed that in fact there is a significant main effect of treatment (Holson et al., 2008).

8.4. Correlated Measures

For many learning and memory procedures, multiple measures of performance are collected. For example, latency, speed, and distance measures are often all reported in water mazes. It is important to recognize that these measures are usually highly correlated or interdependent. For example, because swimming speed is typically computed as the distance divided by latency, latency and speed will be correlated. Even where more sophisticated algorithms in modern equipment software are employed, these two measures would still be highly correlated. Another example would be for the Biel maze where latency to reach the escape platform will typically be greater where animals make more choice point errors simply because errors increase time for task completion. In such a case, the correlations are inherent in the nature of the animal behavior and not in the nature of the computation of the metrics. These intercorrelations between multiple measures of performance have implications for data analysis since performing separate ANOVAs, for instance, on speed and latency data assumes that these measures are independent when they are not. Though multiple measures of performance are frequently analyzed separately, it is important to recognize that they are not independent and in most cases statistical results should confirm comparable results for separate measures. Note that, like multiple uncorrected pairwise comparisons (e.g., multiple t-tests), multiple ANOVAs on correlated learning measures also increase type 1 errors (i.e., finding a significant difference when it does not actually exist). However, it should be noted that these are not as much of concern in the interest of being cautious and health-protective. Multivariate statistical procedures are available that adjust for the intercorrelations, but these are not yet widely used and some are extremely conservative. Another approach to address assessment of multiple correlated measures from a given dataset is to decrease or eliminate these redundancies (e.g., measures only path length or latency, but not both). Unfortunately, this also precludes evaluating some of the information in the data set.

Another option would be to assess the effects of treatment on the expected correlations between multiple measures of learning. This would be especially valuable where treatment may affect one measure but not another. For example, treatments that alter motor function specifically would be expected to influence speed more than choice errors and hence the correlations between these measures would differ between control and treated animals. This may aid in interpretation in cases where speed is not directly measured on every trial.

8.5. Outcomes of Interest

In most cases, the types of effects that can be derived from learning and memory tests include:
Main effects:

- Treatment: the dependent variable differs across groups, irrespective of other factors such as sex and test-session. This indicates the simple case of a treatment effect that is the same in both sexes across training;
- Sex: difference between males and females without consideration of other factors. Since sex differences are well-known for some cognitive tests, this can be used as an assessment of the overall success of the task. For example, females typically do not perform as well as males in spatial tasks (depending on the test parameters), and an overall effect of sex should be expected (e.g., Blokland et al., 2006; Jonasson, 2005; Perrot-Sinal et al., 1996; Roof and Stein, 1999; Vorhees et al., 2008);
- Test session: since learning is defined as a change in behavior as a result of training, there should be an overall effect of test session indicating significant changes from the first to the last test session. This is especially true for the control group when analyzed

separately; indeed, evidence of learning in controls is necessary for interpretation of treatment data as well as satisfying guideline requirements.

Interactions:
Significant multi-way interactions need to be followed using appropriate comparison procedures to assess the precise nature of the interaction. It is also possible to have significant interactions but not have a significant effect on follow-up tests.

- Treatment-by-test session: indicates that the rate of change of the behavior over test sessions differs across treatment groups. This may occur when behavior on the first trial is similar across groups (e.g., random searching for a platform), and differences in this behavior become evident as testing continues. Even if asymptotic performance is achieved, the time (e.g., number of training sessions, trials to criterion) taken to reach this could indicate an adverse effect;
- Treatment-by-sex: indicates that the treatment effect differs between male and females. This may occur when the behavior is affected by treatments in both sexes but the magnitude of effects differs between sexes, or the direction of change is different between sexes;
- Treatment-by-test session-by-sex: this 3-way interaction can indicate multiple outcomes, such as the interaction between treatment and test session differs across sex, or the interaction between test session and sex is altered by treatment. The nature of the interaction must be followed with contrast interaction tests or step-down analyses of one or two variables. A close examination of the data, or the specific hypothesis tested, can influence the choice of subsequent analyses.

In all analyses, the data should be closely evaluated for consistency and expected outcomes, especially in establishing that control animals learned the task. Group differences should be considered in light of the magnitude of effect, since biologically significant changes could be masked by high variability or under-powered studies. Likewise, individual data within the groups should be examined where outliers may suggest animals that are more or less sensitive, or else indicative of high variability.

9. Interpretation of Treatment-Related Effects in Tests of Learning and Memory

As noted above, in evaluating the results, care should be taken to account for possible non-parametric distribution of the data (especially given the frequent use of cut-off criteria for trial number and trial duration). In such cases, the appropriate measure of central tendency should be reported and non-parametric statistical approaches applied.

For positional discrimination data, in addition to evaluating changes in the mean or median values, the reviewer should look at changes in the pattern of errors. Changes in error patterns can sometimes be manifested as an apparent mismatch between the number of errors and the number of trials to criterion. This consideration is illustrated in Table 3 above. Careful examination of individual data may be useful in evaluating differences among means/medians and variability among treatment groups.

As discussed in the Biel/Cincinnati maze section, it is also possible that longer latencies may occur without an increase in errors or a decrease in swim speed. Such a pattern can result if an animal hesitates at the choice point and may still reflect a learning impairment.

For tests in which acquisition occurs over several days and there are multiple trials per day, it should be possible to evaluate within-day acquisition, and frequently to see some decrement in learning on the first trial of each day, when compared to the last trial of the previous day. Data should be presented in such a way that these daily learning curves can be assessed.

It is inappropriate to evaluate retention in the absence of comparable levels of performance between treatment groups in the acquisition phase. Evaluation is straightforward in studies using trials to criteria, whereas in studies with a fixed number of training trials, comparable performance may not actually be reached by the last trial. This highlights the importance of experimental design in interpretation of results.

Variability should be within a reasonable range. A complete lack of variability in control animals may be indicative of an insensitive task, which can result from different aspects of task design. For example, if all animals exhibit perfect performance during retention, then the task may be too easy, the aversive stimulus too strong, or the initial learning criterion set too high. Excessive variability in control animals decreases study power, making it very difficult to demonstrate that learning has occurred, and will also make it difficult to detect changes in performance between control and treatment groups (see Raffaele et al., 2008).

If impairment (e.g., measured changes in latency in position discrimination or the MWM learning) is observed in the presence of a reduction in swim speed, it is not possible to distinguish learning from performance deficits. Changes in swim speed may be indicative of motor dysfunction and should be evaluated in the context of other behavioral assessments of motor function (e.g., motor activity, cage side observations).

If impairment is detected in only one sex, it can still be considered evidence of a learning deficit.

If impairment is seen at only one age, it can still be considered evidence of a learning deficit.

Presence of an atypical dose-response relationship warrants closer evaluation of individual data and consideration of the potential contribution of nonspecific effects that may be treatment-related. For example, increased activity levels may lead to faster swim speeds but more errors, or hypothermia may change the aversiveness of the stimulus, altering motivation. Consideration of other information collected in the context of a guideline study may be informative in evaluation of observed effects in learning tasks.

With these considerations in mind, it is important to emphasize that in the absence of sufficient control procedures or other scientifically sound explanations, any change of performance in this task must be interpreted as an adverse effect on cognitive function.

10. References

Ader R, Weijnen JAWM, Moleman P. (1972) Retention of a passive avoidance response as a function of the intensity and duration of electric shock. *Psychon. Sci.* 26, 125-128.

Blokland A, Geraerts E, Been M. (2004) A detailed analysis of rats' spatial memory in a probe trial of a Morris task. *Behav. Brain Res.* 154, 71-75.

Blokland A, Rutten K, Prickaerts J. (2006) Analysis of spatial orientation strategies of male and female Wistar rats in a Morris water escape task. *Behav. Brain Res.* 171, 216-224.

Brown KL, Pagani JH, Stanton ME. (2005) Spatial conditional discrimination learning in developing rats. *Dev. Psychobiol.* 46, 97-110.

Carter CS, Freeman JH Jr, Stanton ME. (1995) Neonatal medial prefrontal lesions and recovery of spatial delayed alternation in the rat: effects of delay interval. *Dev. Psychobiol.* 28, 269-279.

Castañe A, Theobald DE, Robbins TW. (2010) Selective lesions of the dorsomedial striatum impair serial spatial reversal learning in rats. *Behav. Brain Res.* 210, 74-83.

Chadman KK, Watson DJ, Stanton ME. (2006) NMDA receptor antagonism impairs reversal learning in developing rats. *Behav. Neurosci.* 120, 1071-1083.

Clark RE, Zola SM, Squire LR. (2000) Impaired recognition memory in rats after damage to the hippocampus. *J. Neurosci.* 20, 8853-8860.

D'Hooge R, De Deyn PP. (2001) Applications of the Morris water maze in the study of learning and memory. *Brain Res. Brain Res. Rev.* 36, 60-90.

Dudchenko PA, Talpos J, Young J, Baxter MG. (2013) Animal models of working memory: a review of tasks that might be used in screening drug treatments for the memory impairments found in schizophrenia. *Neurosci. Biobehav. Rev.* 37, 2111-2124.

Dudchenko PA. (2004) An overview of the tasks used to test working memory in rodents. *Neurosci. Biobehav. Rev.* 28, 699-709.

Ehman KD, Moser VC. (2006) Evaluation of cognitive function in weanling rats: a review of methods suitable for chemical screening. *Neurotoxicol. Teratol.* 28, 144-161

Eichenbaum H. (2000) A cortical-hippocampal system for declarative memory. *Nat. Rev. Neurosci.* 1, 41-50.

European Food Safety Authority (EFSA). (2010) Statistical re-analysis of the Biel maze data of the Stump et al (2010) study: "Developmental neurotoxicity study of dietary bisphenol A in Sprague-Dawley rats" EFSA Journal 8:1836. Doi: 10.2903/j.efsa.2010.1836

Freeman JH Jr, Stanton ME.(1991) Fimbria-fornix transections disrupt the ontogeny of delayed alternation but not position discrimination in the rat. *Behav. Neurosci.* 105, 386-395.

Gilbert ME, Sui L. (2006) Dose-dependent reductions in spatial learning and synaptic function in the dentate gyrus of adult rats following developmental thyroid hormone insufficiency. *Brain Res.* 1069, 10-22.

Gold PE. (2003) Acetylcholine modulation of neural systems involved in learning and memory. *Neurobiol. Learn. Mem.* 80, 194-210.

Graham DL, Schaefer TL, Vorhees CV. (2012) Neurobehavioral testing for developmental toxicity. In: Developmental and Reproductive Toxicology: A Practical Approach, Third Edition. RD Hood (editor), Informa Healthcare, London, UK.

Green RJ, Stanton ME. (1989) Differential ontogeny of working memory and reference memory in the rat. *Behav. Neurosci.* 103, 98-105.

Hoh T, Beiko J, Boon F, Weiss S, Cain DP. (1999) Complex behavioral strategy and reversal learning in the water maze without NMDA receptor-dependent long-term potentiation. *J. Neurosci.* 19, RC2 1-5.

Holson RR, Freshwater L, Maurissen JP, Moser VC, Phang W. (2008) Statistical issues and techniques appropriate for developmental neurotoxicity testing: a report from the ILSI Research Foundation/Risk Science Institute expert working group on neurodevelopmental endpoints. *Neurotoxicol. Teratol.* 30, 326-348.

Kenny JT, Blass EM. (1977) Suckling as incentive to instrumental learning in preweanling rats. *Science* 196, 898-899.

Jarvik ME, Kopp R. (1967) An improved one-trial passive avoidance learning situation. *Psychol. Reports* 21, 221-224.

Jonasson Z. (2005) Meta-analysis of sex differences in rodent models of learning and memory: a review of behavioral and biological data. *Neurosci. Biobehav. Rev.* 28, 811-825.

Kihlstrom, JF, Dorfman J., Park, L. (2007) Implicit and explicit memory and learning in M. Velmans & S. Schneider (Eds.), The Blackwell Companion to Consciousness. Blackwell, Oxford, U.K.

Means LW, Alexander SR, O'Neal MR. (1992). Those cheating rats: male and female rats use odor trails in a water-escape "working memory" task. *Behav. Neurol. Biol*. 58, 144-151.

Morford LL, Inman-Wood SL, Gudelsky GA, Williams MT, Vorhees CV. (2002) Impaired spatial and sequential learning in rats treated neonatally with D-fenfluramine. *Eur. J. Neurosci.* 16, 491-500.

Morris RGM. (1981) Spatial localization does not require the presence of local cues. *Learn. Motiv.* 12, 239-260.

Morris R. (1984). Developments of a water-maze procedure for studying spatial learning in the rat. *J. Neurosci. Methods* 11, 47-60.

Moser VC, Barone S, Jr., Smialowicz RJ, Harris MW, Davis BJ, Overstreet D, Mauney M, Chapin RE. (2001) The effects of perinatal tebuconazole exposure on adult neurological, immunological, and reproductive function in rats. *Tox. Sci.* 62, 339-352

OECD/OCDE 426. (2007) OECD Guideline for the Testing of Chemicals, Developmental Neurotoxicity Study.

OECD. Guidance Document Supporting OECD Test Guideline 443 on the Extended One-Generation Reproductive Toxicity Test. Series on Testing and Assessment, No. 151. 2004.

Olton DS. (1979) Mazes, maps, and memory. *Am. Psychol.* 34, 583-596.

Pagani JH, Brown KL, Stanton ME. (2005) Contextual modulation of spatial discrimination reversal in developing rats. *Dev. Psychobiol.* 46, 36-46.

Paul CM, Magda G, Abel S. (2009) Spatial memory: Theoretical basis and comparative review on experimental methods in rodents. *Behav. Brain Res.* 203, 151-164.

Peele DB, Vincent A. (1989) Strategies for assessing learning and memory, 1978-1987: a comparison of behavioral toxicology, psychopharmacology, and neurobiology. *Neurosci. Biobehav. Rev.* 13, 317-322.

Peele DB, Allison SD, Crofton KM. (1990) Learning and memory deficits in rats following exposure to 3,3'-iminodipropionitrile. *Toxicol. Appl. Pharmacol.* 105, 321-332.

Perrot-Sinal TS, Kostenuik MA, Ossenkopp K-P, Kavaliers M. (1996) Sex differences in performance in the Morris water maze and the effects of initial nonstationary hidden platform training. *Behav. Neurosci.* 110, 1309-1320.

Raffaele KC, Fisher JE Jr, Hancock S, Hazelden K, Sobrian SK. (2008) Determining normal variability in a developmental neurotoxicity test: a report from the ILSI Research Foundation/Risk Science Institute expert working group on neurodevelopmental endpoints. *Neurotoxicol. Teratol.* 30, 288-325.

Roof RL, Stein DG. (1999) Gender differences in Morris water maze performance depend on task parameters. *Physiol. Behav.* 68, 81-86.

Salehi B, Cordero MI, Sandi C. (2010). Learning under stress: the inverted U-shape function revisited. Learn. Mem. 17, 522-530.

Slikker W Jr, Acuff K, Boyes WK, Chelonis J, Crofton KM, Dearlove GE, Li A, Moser VC, Newland C, Rossi J, Schantz S, Sette W, Sheets L, Stanton M, Tyl S, Sobotka TJ. (2005) Behavioral test methods workshop. *Neurotoxicol. Teratol.* 27, 417-427.

Smith GJ, Bogomolny A. (1983) Appetitive instrumental training in preweanling rats: I. Motivational determinants. *Dev. Psychobiol.* 16, 119-128.

Squire LR (2004) Memory systems of the brain: A brief history and current perspective. *Neurobiol. Learn. Mem.* 82, 171-177.

Stanton ME, Spear LP. (1990) Workshop on the qualitative and quantitative comparability of human and animal developmental neurotoxicity, Work Group I report: comparability of measures of developmental neurotoxicity in humans and laboratory animals. *Neurotoxicol. Teratol.* 12, 261-267.

Tilson HA, Mitchell CL. (1984) Neurobehavioral techniques to assess the effects of chemicals on the nervous system. *Ann. Rev. Pharmacol. Toxicol.* 24, 425-450.

US EPA, Health Effects Guidelines OPPTS 870.6300, Developmental Neurotoxicity Study, 1998.

Vorhees CV. (1987) Reliability, sensitivity and validity of behavioral indices of neurotoxicity. *Neurotoxicol. Teratol.* 9, 445-464.

Vorhees CV, Herring NR, Schaefer TL, Grace CE, Skelton MR, Johnson HL, Williams MT. (2008) Effects of neonatal (+)-methamphetamine on path integration and spatial learning in rats: effects of dose and rearing conditions. *Int. J. Dev. Neurosci.* 26, 599-610.

Vorhees CV, Weisenburger WP, Acuff-Smith KD, Minck DR. (1991) An analysis of factors influencing complex water maze learning in rats: effects of task complexity, path order and escape assistance on performance following prenatal exposure to phenytoin. *Neurotoxicol. Teratol.* 13, 213-222.

Vorhees CV, Williams MT. (2006) Morris water maze: procedures for assessing spatial and related forms of learning and memory. *Nat. Protoc.* 1, 848-858.

Vorhees CV, Williams MT. (2014a) Assessing spatial learning and memory in rodents. *ILAR J.* 55, 310-332.

Vorhees CV, Williams MT. (2014b) Value of water mazes for assessing spatial and egocentric learning and memory in rodent basic research and regulatory studies. *Neurotoxicol. Teratol.* 10, 75-90.

Vorhees CV, Williams MT. (2016) Cincinnati water maze: A review of the development, methods, and evidence as a test of egocentric learning and memory. *Neurotoxicol Teratol.* 57, 1-19.

Watson DJ, Herbert MR, Stanton ME. (2009) NMDA receptor involvement in spatial delayed alternation in developing rats. *Behav. Neurosci.* 123, 44-53.

Watson DJ, Sullivan JR, Frank JG, Stanton ME. (2006) Serial reversal learning of position discrimination in developing rats. *Dev. Psychobiol.* 48, 79-94.

Whishaw IQ, Tomie J. (1997) Perseveration on place reversals in spatial swimming pool tasks: further evidence for place learning in hippocampal rats. *Hippocampus* 7, 361-370.

Williams MT, Morford LL, Wood SL, Wallace TL, Fukumura M, Broening HW, Vorhees CV. (2003) Developmental D-methamphetamine treatment selectively induces spatial navigation impairments in reference memory in the Morris water maze while sparing working memory. *Synapse* 48, 138-148.

Williams MT, Vorhees CV, Boon F, Saber AJ, Cain DP. (2002) Methamphetamine exposure from postnatal day 11 to 20 causes impairments in both behavioral strategies and spatial learning in adult rats. *Brain Res.* 958, 312-321.

MODULE - E

CONSIDERATIONS IN THE ANALYSIS AND INTEGRATION OF DNT DATA

MODULE E – CONSIDERATIONS IN THE ANALYSIS AND INTEGRATION OF DNT DATA

1. Introduction

The purpose of this module is to provide further points for consideration in performing the analysis and overall integration of results from DNT studies. This would include aspects related to statistics, dose-response and control group considerations. In addition, integration of individual parameters within the DNT study, including maternal effects, along with consideration of findings in other toxicity studies is discussed. Additional considerations such as exposure and human relevance are also addressed.

Several principles for evaluation of neurotoxicological effects have been put forth in guidelines of the US EPA (1998) and OECD (2007). In general:

- An agent that produces detectable adverse neurotoxic effects in experimental animal studies is assumed to pose a potential hazard to humans;
- Developmental neurotoxic effects in animal studies indicate the potential for altered neurobehavioral development in humans, although there may be differences in the specific manifestation of those effects;
- An alteration in behaviors of the offspring, or in the ontogeny of behaviors, is considered to indicate a developmentally neurotoxic change. These may occur with or without neuropathological findings;
- Changes in neuronal organization, structure, or neurochemistry also indicate an effect, with or without accompanying functional effects;
- Data from all potentially relevant studies and effective doses or exposures should be considered in a weight-of-evidence approach to characterizing the potential for developmental neurotoxicity;
- While findings at all dose levels are important, those occurring at doses below levels associated with maternal or general toxicity are generally considered of increased concern;
- While an understanding of toxic mechanisms or pathways can inform an overall evaluation, the lack of such information does not preclude a determination of risk.

With the current DNT test guidelines, there are a number of behavioral tests and other assessments conducted in the offspring. To evaluate such a wide range of outcomes, the data need to be integrated, taking into account consideration of statistical and biological significance. In addition, residual uncertainties resulting from limitations in the conduct of some parameters assessed within the DNT study must also be considered, along with findings in the toxicology database as a whole. The complexities in the nature and the assessment of the endpoints generated in DNT testing make data evaluation a challenge. Reasonable interpretations of the data are required to ensure that false positives (type I error) and false negatives (type II error) are avoided. Approaches to integrating and synthesizing the data are outlined in this module, and are also described in Tyl et al. (2008).

2. Step-Wise Approach To Integrated Analysis Of DNT Study Data

The following outline describes the steps in evaluating DNT data as a whole, and integrating DNT results with the available toxicity database. Data from the individual test measures should be assessed before an overall synthesis can be initiated. Historical and positive control data can be an additional resource for evaluating data but do not substitute for data/findings in the concurrent controls.

Step 1: Evaluate the patterns and/or consistency of effect and the plausibility of each individual endpoint. Separate guidance is available here for motor activity, acoustic startle response, learning and memory, and observations (see modules A-D).

Step 2: Compare the effective dose levels, patterns of change, and the severity of effects across endpoints within the DNT study. Consider changes in some endpoints in the context of others where possible (e.g., influence of motor activity dysfunction on some cognitive tests) (section 3.1).

Step 3: Evaluate general health and growth parameters (e.g., litter size and viability, body weight, systemic toxicity) within the DNT study. Consider the possible impact of such changes on the sensitivity or reliability of the behavioral test results (e.g., decreased ability to detect statistically significant changes due to lower sample size due to pup loss) (section 3.1).

Step 4: Evaluate the effective doses for, and extent of, maternal toxicity (e.g., decreased body weight gain, systemic toxicity), as well as the effective doses and outcomes in the offspring. It is important to take into account the test guideline requirement that the highest dose level "should induce some overt maternal toxicity" (section 3.2).

Step 5: Integrate these determinations with the available toxicity database to evaluate relative sensitivity of DNT effects and consistency of DNT effects with other measures (section 3.3). Additional factors can be considered, including route of exposure and kinetic (absorption, distribution, metabolism, excretion) influences (section 4).

3. Overall Data Synthesis

The major steps outlined above should be followed for the purpose of properly assessing the evidence or data on patterns of related effects within and across toxicity studies. General guidance (OECD, 2004; US EPA, 1998) states that changes in any DNT endpoint can reflect an action on the development of the nervous system, unless data can be provided to support a different conclusion. The level of confidence in the overall conclusion that a given chemical adversely affects the developing nervous system can vary depending upon the weight of evidence. Higher levels of concern are produced with datasets for which multiple, statistically and biologically significant changes are evident. A lack of consistency or dose-response in the data may lower, but not necessarily negate, the concern for potential DNT. The steps and guidance below should be considered in evaluating data from the DNT study itself within the context of the entire toxicity database.

3.1 Integration of offspring data

To describe DNT outcomes, lines of evidence should be assembled and assessed. There is a multitude of endpoints collected in the offspring during the course of a DNT study. Evaluation across these endpoints should consider factors such as experimental procedures and controls, dose response, treatment effects and statistical and/or biological significance and any uncertainties. Graphical presentation of data is helpful for examining possible patterns of behavioral changes (e.g., habituation, learning curves).

3.1.1 Experimental procedures and controls

Experimental and procedural details are necessary to assure that appropriate procedures were put in place, including counterbalancing the testing across time of day, test days, test chamber, dose groups, and assuring that most/all procedures, but especially subjective evaluations, are conducted with testing personnel unaware of the specific group to which an animal has been assigned. An examination of the control data can provide evidence of a well-conducted study. Control data should be assessed for the following:

- Do control animals show the expected response, based on the literature and historical controls? Observed differences between concurrent and historical controls may be due to a number of factors, e.g., a change in background sensitivity of the subjects (genetic drift) over time, or a concurrent control group that is atypical. While such a situation may complicate interpretation of the data, in general, the concurrent control is the most relevant group against which to compare the treated groups;
- Are control data internally consistent for the following factors?
    - Evidence of learning over test sessions, e.g., a significant effect of trials.
    - Habituation of motor activity and startle response over time, e.g., a significant effect of trials.
    - Maturation of certain behaviors with age, e.g., emergence of clear habituation patterns in motor activity over PND17 to 21.
    - Existence of sex differences in certain behaviors, e.g., male rats show better performance than females in spatially-based cognitive tasks.

3.1.2 Dose-response

Dose-response is a key indicator of a treatment-related effect; however, there is currently discussion on the relevance of non-monotonic dose-response curves (e.g., only the lowest dose has an effect, or the dose-response curve does not show a consistent trend in one direction), especially for certain classes of chemicals (Vandenberg, 2014). Although an orderly trend of increasing effect with increasing dose is the clearest indication of a treatment-related finding, there are many issues to consider. In the face of a weak or absent dose-response (i.e., no gradation of effect), the pattern of individual animal data should be examined to identify changes in incidence or severity of an effect that may have been present but not reflected in the group data. Even if atypical, similar dose-response curves across multiple endpoints or dose groups are particularly supportive of an effect. Considerations for assessing inconsistent dose-response data include:

- Inappropriate dose selection may result in unexpected shapes that would be better defined given more doses. The current guideline requirement for the minimum of three dose levels makes it difficult to determine or define precise dose-response curve shapes;
- Influence of kinetic factors (absorption, distribution, metabolism, excretion). Non-linear responses could be reflective of saturation of metabolic systems or adaptation of response;
- Neural systems reflect interplay of both inhibitory and excitatory actions, and the relative influence of these factors may impact a dose response. These may be observed as U-shaped or inverted U-shaped curves: the classic example is the excitation followed by sedation produced by ethanol. If possible, this interplay should be evaluated in conjunction with any related dose-response measures that are available (i.e., similar findings in tests measuring the same/similar underlying neurological construct).

3.1.3 Determining Biological Significance

There is strong evidence of a neurotoxic effect when several related measures in a battery of tests are affected, there is biological consistency in the changes across endpoints, and/or the effects appear to be dose dependent, especially in the absence of systemic toxicity. There are many reasons, however, why these criteria may not be met in current DNT studies. Thus, while the following considerations are important, the lack of such findings does not necessarily negate the concern for developmental neurotoxicity.

Changes in related measures or endpoints, both within the study and across the database, strengthen the evidence of a biologically significant effect. Where related measures or endpoints exist within a database, one could expect similar direction of change as well as dose response. Patterns to look for could include but are not limited to:

- Altered motor activity levels are important to consider. Although motor activity is directly measured in the automated assessment of activity, additional information on motor activity can be gleaned from cognitive tests that include latency as an endpoint. For example, an animal with low activity may not cross into the shock side in a passive avoidance chamber, and thus confound the assessment of learning. Measuring the latency to cross on the first trial will provide baseline data that can differentiate a learning deficit from a difference in activity levels;
- Motor deficits may be evidenced as neuromuscular weakness or inability to meet the motor demands of a test. Such deficits may be suggested by decreased activity levels, decreased startle amplitude, observations of sluggishness, or lethargy. In a water maze cognitive task, poor swimming ability can impact latency; however, it may not actually impair learning. In this case, swim speed information is useful for looking at motor function separate from task performance, but the data from other measures may also be informative;
- Sensory dysfunction would most likely be detected in the auditory startle test; however, other sensory deficits may not. Visual function, while necessary for some tests of learning (especially spatial tasks), is not often evaluated. Likewise, other senses (e.g., olfaction) are ignored. Where tests of other sensory function are included, similarities of effect across the senses would not necessarily be expected;
- Alterations in habituation pattern may be evident in motor activity and/or auditory startle

data. However, since there are subtle differences in the habituation patterns for these measures, the data may not necessarily be concordant. Thus, the presence or absence of concordance across these measures may not alter the interpretation of an effect;

- In adult animals, brain weight is highly conserved in the presence of changes in body weight, and the convention of using organ to body weight ratios is less useful when studying brain weight changes (Sellers et al., 2007). Studies of malnutrition during development suggest that severe reductions in body weight (e.g., <50% of controls) will result in lower brain weight (Fernandez et al., 1985), but this has been less studied with regards to moderate or slight (e.g., <10% of control) body weight differences. In general, effects on brain weight cannot be dismissed even in the presence of body weight differences, and should be considered treatment-related and adverse;
- Body weight differences (reduced body weight or lower weight gain) or delayed maturation in the offspring may affect some of the DNT measures. Data on growth and body weight are necessary for evaluating neurobehavioral measures. Considerations include:

    - Physical development correlates with body weight, such that developmental landmarks and neuronal functions may be delayed in smaller pups. Many of the behavioral tests are influenced by motoric capacity and body size, e.g., auditory startle amplitude may be less in smaller pups. For this reason, if changes are evident in the presence of delayed growth, it may not be possible to distinguish whether changes in performance are related to delayed growth or to delayed neurobehavioral development per se; for example, delayed development (hair growth) results in a hairless pup whose performance may be compromised in tasks where swimming is required;
    - Most often the source of body weight effects is not known, and could be either maternally mediated (e.g., decreased milk quantity or quality, decreased care) or the results of a direct adverse effect on the pup (e.g., systemic toxicity, neurotoxicity that affects feeding ability).

3.1.4 Statistical Significance

For each test, appropriate statistical analyses must be applied to determine significant differences compared to controls. Necessary factors in statistical analysis include:

- Litter as the unit of analysis when multiple littermates are tested (including males and females);
- Sex as a within-litter or random factor for statistical analyses where males and females are sampled from the same litters;
- Treatment, sex, and within-subject repeated measures in an overall design, followed by step-down analyses and post-hoc group comparisons that are guided by the overall main effects or interactions.

In some cases the statistical data may be misleading. Possible explanations for this include:

- Inappropriate methods of analyses (e.g., using parametric analyses for binary data);
- Underpowered statistical measures. Low statistical power in the study can result from

poor experimental control over testing procedures or small sample size. While coefficients of variation (CV, standard deviation as a percent of average values) can be useful quantitative measures of variability, it should be noted that: 1) high CVs can be influenced by numerically low values, and 2) CVs can only be calculated for parametric data;

- Individual outliers that affect group means (and increase group variability). This could reflect overly sensitive or insensitive subsets of animals, but may also be a result of technician or instrument error. Individual animal data are necessary to assess this;
- Censored data based on pre-determined levels of significance (e.g., p-values > 0.05 but < 0.1). The reporting of actual significance probabilities is useful in this regard, especially where treatment effects approach significance criterion values (e.g., p=0.052);
- The magnitude of change may be too low to exceed the statistical threshold. However, a small change may still be indicative of developmental neurotoxicity. Indeed, the magnitude of effect may merely be a function of dose selection.

Tyl et al. (2008) provide case studies that include a discussion of how to balance biological and statistical significance. Statistical significance of treatment effects does not necessarily mean that the effect is biologically relevant. Because of the large number of potential endpoints in DNT studies, care must be exercised in ensuring that experiment-wise error rates are appropriately controlled to minimize false positives (Holson et al. 2008). However, because DNT studies are conducted to screen chemicals for possible adverse effects on the developing nervous system and these studies are often the only study with these biologically important neurological endpoints, the willingness to accept a higher false positive rate instead of a lower false negative rate may be a more conservative and protective approach.

Most importantly, biologically significant, treatment-related findings may occur in the absence of statistical significance. If reported variances in a given behavioral test compared with other reports from the same laboratory (historical controls) or with other laboratories suggest excessive variance within a study, the study itself may be flawed, at least for the parameter under consideration. The utility of these test results should be considered in the context of the other measures in the study, at which point the particular parameter may not be used, or this could warrant a repeat of the study with better experimental control.

Thus, the conclusion of whether or not a chemical exposure has resulted in adverse outcomes on neurodevelopment should not rely solely on evidence of a statistically significant treatment-related effect. Consideration should be given to factors such as multiplicity of statistical findings as well as the level of statistical significance in the weight of evidence determination; however, statistical significance should not override biological significance.

3.2 Integration of maternal data

The DNT test guidelines require that the highest dose level "should induce some overt maternal toxicity", although the criteria for this determination are not described. The DNT test guidelines do not specifically require evaluation of maternal behavior, but body weight (and sometimes food/water consumption) will be available for review. The comparison of toxicity in the dam with that seen in the offspring is important in determining age-related sensitivity for certain regulatory jurisdictions; issues to consider include:

- A large decrease in food/water intake could result in decreased body weight that exceeds an acceptable level, resulting in malnutrition of the dam. While this situation may result in neurobehavioral and physiological impacts on the offspring, the level of maternal malnutrition that consistently leads to impairment of neurobehavioral development has not been determined;
- Maternal care (e.g., nursing, nesting, grooming) is critical for neuronal development. This can be monitored by observations of how well the litters are kept together (the dam usually retrieves pups that go astray), temperature of the pups (coolness to the touch), and the presence of milk bands. While alterations in maternal behaviors should be considered in assessing potential neurotoxic effects in the offspring, these simple assessments are not typically included in regulatory DNT studies;
- Treatment-related differences in litter size, fetal birth weight, and pup growth can result from maternal influences, as well as from direct toxic effects of the chemical on the fetus and/or newborn. Data available from DNT studies are not sufficient to differentiate between maternal and fetal origin of the effect; such findings should be considered treatment-related and adverse.

There are several approaches for designing studies to address whether the effects are attributable to maternal toxicity or direct neurotoxicity on the offspring. The most common of these is a cross-fostering procedure. These additional studies are not typically conducted, and the reviewer is advised to consult a neurobehavioral expert if such studies are submitted.

3.3 Integrating other lines of evidence

After conducting a critical analysis of the data from the DNT study, the reviewer should consider findings from other studies.

- Reproductive toxicity studies, either single or multigenerational. These exposures cover the pre-mating, gestational and lactational periods as well as the post-natal periods in the offspring (F1, F2 generations). This continuous exposure is in contrast to the discrete perinatal exposure of the DNT study. Information from such studies can be used to aid in interpretation of DNT studies; however, differences in the dosing paradigms across studies could lead to different manifestations of toxicity. For that reason, a lack of similar findings across different types of studies should not be used to dismiss discordant effects. Information from these studies that may inform DNT findings include:

    - Developmental delay(s) in physical and reproductive parameters. There may be an expectation of similar effects in reproductive and DNT studies, especially if the studies include comparable doses. As described in section 3.1.3, developmental delays may alter behavioral endpoints, especially when testing at an early age. Pup body weight may be used to infer developmental delay. Many developmental landmarks (e.g. pinna detachment, incisor eruption, eye opening) are highly correlated with body weight, and a delay in physical development often correlates with later appearance of these landmarks. Delayed onset of puberty (delayed preputial separation or vaginal opening) is also often a reflection of delayed physical growth;

- Marked or obvious changes in the behavior of the offspring noted as part of routine clinical observations. Since chemical exposure continues after weaning in reproductive toxicity studies, signs of neurotoxicity may be due to a direct action of the chemical on neuronal function, which may differ from a permanent change in the development of the nervous system itself. The reproductive toxicity study design does not allow differentiation between these types of causality, but either type of effect would be of concern.

- Other toxicity studies. For many chemicals, data from other studies, including those assessing subchronic neurotoxicity and chronic toxicity, may be available. These studies are conducted in adult animals and thus would reflect direct chemical effects on the adult animal rather than effects on the developing nervous system. That said, these data can be compared to maternal data derived from the DNT study, especially where there is similarity in dose levels and route of administration. Effective dose ranges across studies can inform sensitivity of different target organs and toxicity manifestations. Care should be taken to consider the potential impact of other dosing regimens in any comparisons;
- DNT studies in the literature. The exposures and experimental procedures used in DNT studies reported in the peer-reviewed literature differ from those used in the DNT test guidelines in almost all cases. There may be some consistency in DNT effects of some chemicals across studies, but for many others there may be considerable differences; these may result from differences in exposure, test selection and conduct, age at assessment and many other factors. The evaluator should be cognizant of these differences when making comparisons between the literature data and the study under review.

4. Additional considerations for interpretation

Additional issues may be considered in the overall interpretation of DNT data. These include exposure considerations and kinetic factors, both of which impact levels of the test chemical in the fetus/offspring. In addition, kinetic information may inform extrapolation to humans.

4.1 Exposure

The reviewer should consider details of the exposure scenario, which can influence chemical levels in the fetus and offspring:

- In DNT studies, the dam receives the test chemical and fetal exposure is assumed to occur through placental transfer. For postnatal exposure, the DNT guidelines specify that there should be assurance of continued (i.e., lactational) exposure, which requires demonstration of transfer of the test chemical to the pup through the milk. In ideal situations, data demonstrating the presence of the chemical in milk or in the pup would be provided. If data indicate a lack of lactational transfer, this would suggest the need for direct dosing of the pups after birth. While more challenging, direct dosing of pups has been used in regulatory DNT studies and can be accomplished without adversely impacting the study (Moser et al., 2005);
- Dietary exposure (food, water) is considered to be a relevant dose route for potential human exposure, and it is often used in DNT studies. However, food and water intake of

the dam changes over the study, and greatly increases during lactation. Thus, exposure at fixed concentrations (i.e., ppm in diet or water), will result in much higher chemical intake by the dam on a mg/kg bw basis during lactation (Yoon and Barton, 2008). This higher intake could result in additional maternal or pup toxicity during lactation. For this reason, changes in chemical intake over time should be noted. Since water and/or food intake is often measured during the study, actual dose should be calculated and reported. On the other hand, gavage dosing (to the dam or pup) generally provides a consistent amount of compound on the basis of body weight;

- It should be noted that by the second to third postnatal week (towards the end of the lactation period), pups start consuming water and feed in addition to their nursing activity. If the test chemical is administered in the diet or in drinking water, it is likely that chemical intake increases for some of this time due to these multiple routes of exposure (lactational plus direct oral);
- Other relevant exposure routes may include inhalation or dermal, and review of such studies should consider the additional complexities of these exposures. For example, daily inhalation and dermal exposures may be stressful to the pregnant rat, and stress alone can produce adverse outcomes. For inhalation exposure, there are two options for continued exposure after birth, both of which have shortcomings: 1) exposure of the dam only, but this requires separation from the pups for many hours a day, limiting their maternal care and nourishment, or 2) exposure of the entire litter, but this may lead to inconsistent uptake due to the pups nesting and burrowing, preventing access to the vapor. For dermal exposure, direct dosing of the pups should take into account their immature skin and lack of hair after birth.

Kinetic parameters are important in determining whether there are differences in systemic exposure in the young animal compared to adults. In some cases physiological changes underlying absorption, distribution, metabolism, and elimination (ADME) may take days to weeks to reach adult capacities. However, there is often a lack of age-specific kinetic information and therefore one must rely on some generalizations:

- Metabolic capabilities are often low in the very young, and the ontogeny of different metabolic pathways occurs at different rates across species (Saghir et al., 2012). The resulting impact may vary depending on the development of the specific systems through which the test chemical is metabolized. This can result in higher tissue levels in the pup, but it may also result in lower levels of the toxic moiety if it is the metabolite, not the parent compound, that is active;
- Absorption, distribution, and elimination of the test chemical may differ in the young compared to the adult, leading to greater or lower blood levels. Furthermore, the blood-brain barrier and transporter proteins change during development, which may influence uptake into the brain. The impact of these differences in terms of higher or lower tissue levels may depend on specific characteristics of the test chemical (e.g., lipophilicity, molecular size).

5. Human Relevance

There are specific processes and/or critical developmental periods in nervous system development, during which chemical-induced disruptions may result in long-lasting adverse

outcomes. Awareness of differences in brain development and function across species can help inform extrapolation of rodent DNT data to human infants and children.

Nervous system development across species:

- Nervous system development begins in early gestation and continues through adolescence in both rodents and humans (reviewed in Bayer et al., 1993; Rice and Barone, 2000). As the underlying biological processes are similar across species, chemical effects targeting those processes may be comparable. Proliferation and migration of neural and glial cells occur in waves across brain regions. Neural connections are established through differentiation and synaptogenesis, and are influenced by a number of molecular signals. Myelination processes are most protracted, and a period of rapid glial proliferation (known as the brain growth spurt) follows. This occurs postnatally in rodents (Dobbing and Sands, 1979; Dobbing and Sands, 1973) as opposed to other mammals (including humans) where this occurs in utero. These basic developmental processes occur in a specific, tightly controlled sequence that differs across brain regions and cell types. While the general processes and sequence in brain development is conserved across species, the rate and timing of specific processes relative to parturition differ between species. The timing of neurogenesis can be matched across species (Clancy et al., 2007); a web-based model providing temporal comparisons across a number of mammals is available at http://translatingtime.net/;
- Consideration of temporal differences in timing of brain development may provide insight into identifying potential targets for neurodevelopmental effects. However, it should be noted that in guideline DNT studies, exposure occurs throughout the pre-weaning developmental period and effects cannot be attributed to any specific exposure windows (i.e., gestational versus lactational exposure). Since late gestational development in humans corresponds to early postnatal life in rodents, relative exposures of the human fetus and the postnatal pup can become an important issue. For example, adequate postnatal exposure in the newborn rat is needed to evaluate effects that could occur from late-gestational exposure in the human fetus.

Nervous system function across species:

- Nervous system reflexes and responses are highly conserved across species, allowing direct extrapolation of specific behaviors (e.g., startle reflex); however, higher order functions (e.g., learning, memory and attention) are less well understood and defined, and relevant testing procedures may not be entirely homologous between rodents and humans. Behavioral tests of cognition for rodents and humans may not be specifically analogous due to species differences in behaviors, motor abilities, and motivation;
- The issue of comparability of human and rodent cognitive measures in studies of pharmacology and neurotoxicology has generated a large body of literature, much of which is relevant to the evaluation of DNT data. A few reviews of possible interest are found in D'Mello and Steckler (1996), Sarter (2004), and Sharbaugh et al. (2003).

6. Conclusions

The DNT study was designed as a screening tool to investigate a wide range of possible developmental neurotoxic effects, and it should not be considered a diagnostic tool to address specific human neurological conditions. Data from the rat DNT study are expected to be predictive of adverse outcomes in human infants and children, given similarities in nervous system development. Interpretation of DNT data, however, can be challenging and requires close evaluation and consolidation of the results from testing in offspring and maternal animals as well as other available studies. With attention to details and understanding of the test measures, informed decisions can be made regarding the potential for a test chemical to produce developmental neurotoxicity.

In closing, key points for consideration include:

- Assessments should take into account biological plausibility, consistency or pattern of effects, dose response, and systemic toxicity in the offspring. A higher level of confidence in the conclusions is attained in cases of consistent patterns of effect, appropriate experimental methodologies and study conduct, and accurate statistical analyses;
- Sufficient data are not usually available during the DNT evaluation process to allow full causal inferences to be made with respect to human disease conditions, which require availability of a number of experimental and possibly epidemiological studies;
- It cannot be assumed that toxicity in the offspring is due to maternal toxicity. While clear or severe maternal toxicity may result in toxicity to the pups, or may influence neurological outcomes, this is not always the case. Furthermore, the designs of most standard DNT studies do not provide data to address whether the effects are attributable to maternal toxicity or direct neurotoxicity on the offspring;
- Evaluations should take into account available information on kinetic parameters of the test chemical as well as other available toxicological information, to the extent that this other information is available.

7. References

Bayer SA, Altman J, Russo RJ, Zhang X. (1993). Timetables of neurogenesis in the human brain based on experimentally determined patterns in the rat. *Neurotoxicology* 14:83-144.

Clancy B, Finlay BL, Darlington RB, Anand KJS: Extrapolation brain development from experimental species to humans. *Neurotoxicology* 28:931-937, 2007.

D'Mello GD and Steckler T. (1996) Animal models in cognitive behavioural pharmacology: an overview. *Cog. Brain Res.* 3:345-352.

Dobbing J, Sands J. (1973). Quantitative growth and development of human brain. *Archives of Disease in Childhood*, 48, 757-767.

Dobbing J, Sands J. (1979) Comparative aspects of the brain growth spurt. *Early Human Develop.* 3:79-83.

Fernandez SF, Menendez MF, Fernandez BM, Patterson AM. (1985). Malnutrition in utero and during lactation in the rat: Relationship of dams weight gain and development of suckling. *Nutrition Res.* 5:413-421.

Holson RR, Freshwater L, Maurissen JP, Moser VC, Phang W. (2008) Statistical issues and techniques appropriate for developmental neurotoxicity testing: a report from the ILSI Research Foundation/Risk Science Institute expert working group on neurodevelopmental endpoints. *Neurotoxicol. Teratol.* 30, 326-348.

Moser VC, Walls I, Zoetis T. (2005). Principles for direct dosing of pre-weaning laboratory animals in toxicity testing and research. *Int. J. Toxicol.* **24**:87-94.

Organisation for Economic Co-operation and Development. (2004). Draft Guidance Document on Reproductive Toxicity testing and Assessment. OECD Series on Testing and Assessment No. 43.

OECD/OCDE 426. (2007) OECD Guideline for the Testing of Chemicals, Developmental Neurotoxicity Study.

Rice D, Barone S Jr. (2000) Critical periods of vulnerability for the developing nervous system: Evidence from humans and animal models. *Environ. Health Perspect.* 108 (suppl 3), 511-533.

Saghir, SA, Khan, SA, McCoy, AT. (2012) Ontogeny of mammalian metabolizing enzymes in humans and animals used in toxicological studies. *Crit Rev. Toxicol.* 42:323-357.

Sarter M. (2004). Animal cognition: defining the issues. *Neurosci. Biobehav. Rev.* 28:645-650.
Sellers RS, Morton D, Michael B, Roome N, Johnson JK, Yano BL, Perry R, Schafer K. (2007).

Society of Toxicologic Pathology position paper: Organ weight recommendations for toxicology studies. *Toxicol. Path.* 35:751-755.

Sharbaugh C, Viet SM, Fraser A, McMaster SB. (2003). Comparable measures of cognitive function in human infants and laboratory animals to identify environmental health risks to children. *Environ Health Perspect.* 111:1630-1639.

Tyl RW, Crofton K, Moretto A, Moser VC, Sheets LP, Sobotka TJ. (2008) Identification and interpretation of developmental neurotoxicity effects: A report from the ILSI Research Foundation/Risk Science Institute expert working group on neurodevelopmental endpoints. *Neurotoxicol. Teratol.* **30**, 349-381.

US Environmental Protection Agency. (1998) Guidelines for Neurotoxicity Risk Assessment. EPA/630/R-95/001F.

Vandenberg LN. (2014) Low-dose effects of hormones and endocrine disruptors. *Vitam. Horm.* 94:129-165.

Yoon M, Barton HA. (2008) Predicting maternal rat and pup exposures: how different are they? *Toxicol. Sci* 102:15-32.