EPA Public Access

Author manuscript

Neurotoxicol Teratol. Author manuscript; available in PMC 2019 July 16.

About author manuscripts

Submit a manuscript

Published in final edited form as:

Neurotoxicol Teratol. 2017 September ; 63: 24–45. doi:10.1016/j.ntt.2017.07.001.

Recommendations for harmonization of data collection and analysis of developmental neurotoxicity endpoints in regulatory guideline studies: Proceedings of workshops presented at Society of Toxicology and joint Teratology Society and Neurobehavioral Teratology Society meetings

Abby A. Li^{a,*}, Larry P. Sheets^b, Kathleen Raffaele^c, Virginia Moser^d, Angela Hofstra^e, Alan Hoberman^f, Susan L. Makris^g, Robert Garman^h, Brad Bolonⁱ, Wolfgang Kaufmann^j, Roland Auer^k, Edmund Lau^l, Thomas Vidmar^m, and Wayne J. Bowersⁿ

^aExponent Health Science, Inc., USA

^bBayer CropScience, USA

^cUS EPA Office of Land and Emergency Management, USA

^dUS EPA National Health and Environmental Effects Research Laboratory, Office of Research and Development (NHEERL, ORD), USA

^eSyngenta Canada, Canada

^fCharles River Laboratories, Global Developmental, Reproductive and Juvenile Toxicology, USA

^gUS EPA National Center for Environmental Assessment, Office of Research and Development (NCEA ORD), USA

^hConsultants in Veterinary Pathology, Inc., USA

ⁱGEMpath Inc., USA

^jMerck KGaA, Global Pathology & Reproductive Toxicology, Germany

^kUniversity of Saskatchewan, Department of Pathology, Canada

*Corresponding author at: Exponent Health Science, Inc., c/o Rosalie Bambao, 475 14th St #400, Oakland, CA 94612, USA.

Authors' note

The analysis, conclusions, and opinions expressed in this article are solely those of the authors and do not represent the views and policies of the US EPA, Health Canada, or other organizations that may employ the authors or provide them with funding. Abby Li and Wayne Bowers contributed substantially as first and senior author of the papers. All the authors contributed substantially toward development of presentations made at the workshop, and/or significantly as co-authors of the paper. The authors also wish to thank staff members of the US EPA and Health Canada PMRA, as well as other participants in earlier workshops held at PMRA, for helpful discussions that encouraged this venture.

Conflict of interest

Drs. Bolon, Garman, Hofstra, Hoberman, Kaufmann, Lau, Li, Sheets and Vidmar were employed or funded by companies that conduct studies that may have DNT-like endpoints for regulatory submission.

Disclaimer

The views expressed in this article are those of the author (s) and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency.

Transparency document

The <http://dx.doi.org/10.1016/j.ntt.2017.07.001> associated with this article can be found, in the online version.

^lExponent Health Science, Inc., USA

^mBioSTAT Consultants, USA

ⁿDepartment of Neuroscience, Carleton University, Ontario, Canada

Abstract

The potential for developmental neurotoxicity (DNT) of environmental chemicals may be evaluated using specific test guidelines from the US Environmental Protection Agency or the Organisation for Economic Cooperation and Development (OECD). These guidelines generate neurobehavioral, neuropathological, and morphometric data that are evaluated by regulatory agencies globally. Data from these DNT guideline studies, or the more recent OECD extended one-generation reproductive toxicity guideline, play a pivotal role in children's health risk assessment in different world areas. Data from the same study may be interpreted differently by regulatory authorities in different countries resulting in inconsistent evaluations that may lead to inconsistencies in risk assessment decisions internationally, resulting in regional differences in public health protection or in commercial trade barriers. These issues of data interpretation and reporting are also relevant to juvenile and pre-postnatal studies conducted more routinely for pharmaceuticals and veterinary medicines. There is a need for development of recommendations geared toward the operational needs of the regulatory scientific reviewers who apply these studies in risk assessments, as well as the scientists who generate DNT data sets. The workshops summarized here draw upon the experience of the authors representing government, industry, contract research organizations, and academia to discuss the scientific issues that have emerged from diverse regulatory evaluations. Although various regulatory bodies have different risk management decisions and labeling requirements that are difficult to harmonize, the workshops provided an opportunity to work toward more harmonized scientific approaches for evaluating DNT data within the context of different regulatory frameworks. Five speakers and their coauthors with neurotoxicology, neuropathology, and regulatory toxicology expertise discussed issues of variability, data reporting and analysis, and expectations in DNT data that are encountered by regulatory authorities. In addition, principles for harmonized evaluation of data were suggested using guideline DNT data as case studies.

Keywords

Developmental neurotoxicity testing; Extended one-generation reproductive toxicity; Behavior; Learning and memory; Morphometry; Neuropathology

1. Introduction

There is concern globally about the potential for food-use pesticides, pharmaceuticals, and other chemicals to cause neurodevelopmental effects following perinatal exposures (Bjørning-Poulsen et al., 2008; Grandjean and Landrigan, 2014). Neurotoxicity evaluations from studies conducted according to test guidelines of the US Environmental Protection Agency (US EPA, 1998a) and the Organisation for Economic Cooperation and Development (OECD, 2007), as well as the recently finalized extended one-generation reproductive toxicity study (EOGRS; OECD, 2012a) guideline, may play a pivotal role in children's

health risk assessment. Neurological endpoints such as auditory startle habituation, motor activity, and brain morphometry are now being used in regulatory hazard identification and risk assessments; however, data from the same study may be interpreted differently by regulatory authorities in different countries, leading to differences in public health protection, which may, in turn, lead to commercial trade barriers. The issues of which specific tests to use, and data interpretation and reporting for DNT studies are also relevant to juvenile and pre-postnatal studies (ICH S5-R2, 2005) conducted more routinely for pharmaceuticals.

Previous efforts have provided excellent reviews of some of the important issues that are encountered in DNT studies (Crofton et al., 2008; Holson et al., 2008; Tyl et al., 2008; Raffaele et al., 2008). These reports provide general recommendations for improving data quality, statistical analysis, and interpretation. In addition, there have been useful publications on recommended practices for implementing and evaluating the OECD and US EPA DNT guidelines for neuropathology and morphometry (Garman et al., 2001, 2016; Bolon et al., 2006; Kaufmann and Gröters, 2006; Bolon et al., 2011; Kaufmann, 2011). Despite these resources, there is a need for further development of recommendations geared toward the operational needs of the regulatory scientific reviewers who use these studies in risk assessments, as well as the scientists who generate DNT data sets.

The workshops summarized here used the experience of the authors to discuss several of the functional evaluations and postmortem procedures included in DNT studies as well as a number of issues that have emerged from diverse regulatory evaluations of these studies. Important issues addressed are the need to: (a) articulate the primary biological hypotheses that are being tested by the DNT guidelines, (b) understand which dependent variables and potential interactions are of primary interest in addressing the biological questions, (c) understand inherent variability of measures from studies performed in conformance with regulatory guidelines, (d) identify practical considerations (e.g., test equipment and conduct) that influence the ability to test different biological hypotheses, and (e) illustrate approaches to evaluating DNT data that take into consideration inherent variability, historical control data, maternal toxicity, and appropriate statistical analyses.

These workshops (Regulatory Neurodevelopmental Testing: New Guiding Principles for Harmonization of Data Collection and Analysis) were held at the annual 2015 Society of Toxicology (SOT) and the 2015 joint Teratology Society and Neurobehavioral Teratology Society (NBTS - now named the Developmental Neurotoxicology Society [DNTS]) meetings held in San Diego, California, US, and Montreal, Quebec, Canada, respectively. Speakers and co-authors were selected to promote exchange of different perspectives and expertise among toxicologists, pathologists, and statisticians from government, industry and academia (Table 1). This report summarizes issues identified with the performance and interpretation of DNT studies that were addressed at these workshops, with the intent that these discussions will assist in generating improved DNT study data, data analysis, and interpretation. The presentations and views expressed by the authors of each presentation are summarized in each section below; these views may not be those of other authors.

The workshops were introduced by brief remarks about the current regulatory need to improve and harmonize the analysis and evaluation of DNT studies. Dr. Angela Hofstra noted that the disparity in the regulatory interpretation of DNT study results is due, in part, to different expectations and interpretations by regulatory agencies about variability and outcomes of DNT data. In addition, insufficient reporting of methods and data or inadequate analysis of data can contribute to differences in regulatory interpretation between jurisdictions. Dr. Larry Sheets provided an overview of DNT study design requirements, the types of testing included in DNT studies, and addressed how experimental design features of regulatory guideline studies may also contribute to data variability. This presentation also provided historical control data for both motor activity and auditory startle from his laboratory and approaches for qualitative evaluation of patterns of motor activity based on historical control data. Dr. Wayne Bowers extended this discussion of behavioral data variability to factors related to testing parameters (e.g., equipment and software settings), age-related differences in motor capacity, and quantitative approaches for analyzing motor activity data. Dr. Kathleen Raffaele collaborated with Drs. Thomas Vidmar, Edmund Lau, and Abby Li to present approaches for hypothesis-driven statistical analysis of continuous behavioral data using auditory startle data as an example. Using learning and memory as an example, Drs. Virginia Moser and Angela Hofstra highlighted shortcomings in data reporting and analysis encountered by regulatory authorities that influence scientific evaluation. Dr. Moser also discussed approaches for conducting learning and memory tests, an area where there is little specific guidance. Dr. Li discussed three DNT topics that have had an impact on risk assessment decisions regarding relative sensitivity: (1) quality of maternal toxicity assessment, (2) selection of an adverse-effect level, and (3) quality of linear morphometric methods and neuropathology evaluation (presented in collaboration with pathologists Drs. Robert Garman, Brad Bolon, Wolfgang Kaufmann, and Roland Auer). Portions of Dr. Raffaele's and Dr. Li's presentations are captured in separate manuscripts that provide detailed discussion and illustration of optimal approaches for quantitative analysis of behavioral data and for histopathology processing for morphometry measurements (Garman et al., 2016). Consequently, only brief highlights of the latter two presentations are included in this manuscript. The discussion section captures the main points discussed at the meeting.

Given the significant role of the DNT study for hazard identification and risk assessment to protect the developing fetus, infants, and children, Health Canada Pest Management Regulatory Agency (PMRA) has been working in partnership with Health Canada and the US EPA DNT experts to develop internal guidance for how DNT studies conducted according to test guidelines should be analyzed and evaluated based on practical experience of those who have reviewed them (Table 1). The 2015 workshops at the SOT and Teratology Society/NBTS [DNTS] meetings held in the US and Canada provided an opportunity to extend this discussion to the broader scientific community.

2. Global regulatory perspective on harmonization of scientific approaches (Angela Hofstra)

The DNT study provides information on the potential effects of chemicals on prenatal and postnatal toxicity to the developing nervous system. Given the role of the DNT study in assessing risk to the developing fetus and the young, it is imperative that it be conducted and interpreted in a scientifically-robust and defensible manner. However, both the conduct and interpretation are challenging, due to the complexity and lack of familiarity with this study. DNT studies conducted according to current guidelines (Table 2) employ multiple assessments at various life stages and their conduct requires knowledge of developmental toxicity, neurotoxicity, and sophisticated statistics. In addition, DNT studies are conditionally required and, therefore, are less routinely conducted than other standard toxicity studies; consequently, regulatory toxicologists in industry and government may be less familiar with this type of study.

The variability of behavioral endpoints is higher than regulators may be used to seeing, compared to adult standard toxicity or adult neurotoxicity studies that are more frequently conducted (OECD, 2007 Appendix 1; OECD, 1997). In some cases, the variability range from adult neurotoxicity studies is used as a reference point for DNT studies, but these expectations are not necessarily appropriate. This has led to different interpretations of the biological relevance of changes, including non-statistically significant differences from control groups, contributing further to divergent regulatory assessments of DNT endpoints from guideline studies around the globe.

Data from DNT studies can be used as the basis for the regulatory point of departure for risk assessment purposes and to determine the magnitude of uncertainty factors. Of particular importance are vulnerable populations (e.g., infants and children), which are addressed on the regulatory/policy side, with distinct uncertainty factors such as the Food Quality Protection Act (FQPA) safety factor in the United States, the Pest Control Products Act (PCPA) factor in Canada, or hazard classification in the European Union.

Regulatory requirements dictate that Health Canada PMRA must apply the PCPA factor in addition to the traditional uncertainty factors (10× for interspecies and 10× for intraspecies), unless “the PMRA concludes, based on reliable data, that a different factor is appropriate for the protection of infants and children” (Health Canada, 2008). The final determination of the magnitude of the PCPA factor (reduce or increase) involves “evaluating the completeness of the data with respect to exposure of and toxicity to infants and children as well as potential for prenatal or postnatal toxicity” (Health Canada, 2008). US EPA Office of Pesticide Programs (OPP) also must “consider the special susceptibility of children to pesticides by using an additional tenfold (10×) safety factor when setting and reassessing tolerances, unless adequate data are available to support a different factor” (US EPA, 2015).

A regulator in one region may conclude that an endpoint or point of departure for regulatory assessment is the highest dose with no increased sensitivity of the young, while a regulator in a different region may assess the same study and determine that it indicates findings in the young at doses that are not maternally toxic. This could lead to assessments based on points

of departure that are quite different for various regulatory bodies. In addition, composite uncertainty factors that include the FQPA factor (US) or PCPA factor (Canada) can vary by a factor of 3 or 10. Taken together, this can lead to very different risk assessments impacting commercial interests, such as uses, import/exports, classification, labeling, and packaging. These differences could lead to different health protective values and restricted or cancelled uses of a chemical in one country but not another. In the experience of the industry representatives, this situation has had significant impact on international trade of more than just a few chemicals.

Harmonization of key elements will improve the conduct, reporting, interpretation, and assessment of DNT studies and should help reduce inconsistencies in regulatory decision-making. These key elements include selection of tests or specific behavioral measures, best practices in study conduct, and data assessment, including understanding the hypothesis being tested, addressing and understanding the variability, applying appropriate statistical analysis, and evaluating the weight of evidence.

3. Evaluating data variability for neurobehavioral measures: how much is too much? (Larry Sheets)

This presentation discussed how to evaluate data variability for neurobehavioral measures and also sets the stage for later presentations by using motor activity and auditory startle as examples of these principles. First, various sources of variability were discussed, including which sources are inherent to the behavior that is being evaluated and which are ones that the investigator can control. Second, logistical and practical circumstances were outlined that permit the investigator to produce data of consistently high quality. Finally, different ways to determine whether or not the data are excessively variable were discussed.

3.1. Sources of variability

In general, there are intrinsic and extrinsic sources of variability (Raffaele et al., 2008). For the purposes of this discussion, intrinsic, or biological, variability is defined as variability that is inherent in each measured endpoint (e.g. photobeam breaks for ambulatory activity, peak startle amplitude for auditory startle response), which the investigator has a limited ability to control. Extrinsic variability is defined as variability that is caused by experimental or environmental conditions that can be controlled within the context of US EPA and OECD DNT guidelines. Some procedures that may serve to reduce intrinsic variability (e.g., by over-training the animal or selecting a test that is easy for the animal to solve) can also reduce the sensitivity of the test (Raffaele et al., 2008). This is the reason the title to this presentation refers to “How much is too much?”—because too little intrinsic variability is also not necessarily desirable.

Extrinsic variability is affected by factors, such as environmental conditions, that the investigator must manage in order to achieve consistent results. Some extrinsic factors that impact data variability include the time of day when animals are tested, how the equipment is cleaned between animals, and the manner in which the animals are handled by personnel

on a daily basis and in preparation for testing on a given day (Li, 2005; Raffaele et al., 2008; Crofton et al., 2008; Slikker et al., 2005).

Individual animal factors may have a considerable impact on variability as well, and this is considered “intrinsic” variability. For example, animals are typically assigned the same date of birth within a 24-h pre defined period (e.g., all animals born after 4 pm can be considered born on the next calendar day). At early ages of testing specified by DNT guidelines (e.g., PND 13), this difference in the assigned age can have a major impact on behavior, thereby increasing variability. Because it is impractical to determine the hour of birth for each of the 80–100 litters in a study and test the offspring accordingly, the variability due to this age difference is considered to be “intrinsic” variability.

In studies conducted in my laboratory, preweaning assessment of some neurobehavioral measures such as motor activity can have variability of ~30% to > 100%, based on the standard deviation (SD) compared to the group mean (Table 3). This is consistent with the variability in guideline DNT studies by other laboratories (Maurissen et al., 2000; Tyl et al., 2008; Stump et al., 2010; Sheets et al., 2016). A key question is whether this variability is too high for measures that are inherently more variable than other toxicity endpoints, such as body weight (Maurissen et al., 2000; Stump et al., 2010). Our data show that motor activity in PND 60 adult rats is less variable (~17–30%; Table 4), which agrees well with motor activity data derived from studies that were not part of a DNT assessment (e.g., Crofton et al., 1991). While there are important extrinsic sources of variability that should be controlled, there are also intrinsic sources of variability (differences between animals that are independent of any treatment) that are inherent to the behavioral endpoints or impractical to control within the context of DNT test guidelines.

3.2. Logistical and practical considerations when conducting DNT studies

Speaking from personal experience, a considerable and constant challenge associated with DNT studies is to manage the logistics associated with the test guideline. This includes the number and variety of tests in the study design and the large number of animals that must be tested on specific days of age, preferably within a narrow time window during the day to reduce diurnal variability. In the OECD guideline for a DNT study (OECD, 2007), which is also accepted by the US EPA, the treatment is administered (e.g., by gavage or the diet) from gestation day (GD) 6 to lactation day (LD) 21 (Fig. 1). Testing of the dams is limited to measures of food consumption, body weight, and detailed clinical observations during gestation and lactation, and the dams are euthanized without tissues being collected on LD 21. For pesticides, a DNT study is required when there is an indication for neurotoxicity or increased sensitivity of the young, compared to adults, and this study is performed after the results from prenatal developmental toxicity and multi-generation reproduction studies are available. The findings in these studies are used for dose selection and to indicate whether the chemical is a developmental or reproductive toxicant.

Review of the study design clearly illustrates the focus of the DNT study (US EPA, 1998a; OECD, 2007) is testing of the F1 animals, using neurobehavioral and neuropathology endpoints (Fig. 1, Table 2). In general, the test strategy for these studies (Table 2) is to investigate whether treatment during gestation and lactation produces effects on the nervous

system in F1 animals associated with ongoing treatment and to investigate persistent or latent effects after treatment has been discontinued. Finally, the offspring are euthanized on PND 21 or ~PND 70, with perfusion fixation, gross and microscopic brain morphometry, and a thorough examination of the brain (PND 21 pups and PND 60 adults) and other neural tissues (spinal cord and peripheral nerves in adults only) for histopathology (Table 2). Given the emphasis of the study on the offspring and neurodevelopmental endpoints, it is important to recognize that not every finding is evidence of DNT; rather, the finding should be interpreted in the context of all the available information for the test substance, including evidence of toxicity identified in other studies.

Given the circumstances, it is a constant challenge for the staff to manage the workload without making errors or handling the animals in an inconsistent or hurried manner, which would increase variability (Crofton et al., 2008). The logistics involve taking the correct first generation offspring (F1) for each test from each litter (in the DNT case, for 80–100 litters), transporting the appropriate animals to the test room, placing each one into the proper test apparatus, performing the test appropriately, retrieving the animals and returning each to the correct cage, then collecting the next set of animals for the next test session. In a repeated-testing study design, failure to test an animal on one of the required occasions may remove all of that animal's data from the test data set for some statistical models. In between, the technician may help other staff members by retrieving other animals for a different test being performed at the same time in another room. Further, the selection of tests and animals varies each day. Although careful training of technicians can reduce “extrinsic” sources of variability, the logistical requirements of the DNT guidelines contribute to data variability.

An important question that laboratories and regulators grapple with is whether a data set is excessively variable, after considering intrinsic variability. There is no single best reference or “gold standard” for use to gauge variability—it depends on the nature of the data and the specific measures. In some cases, the lower variability of published academic studies has been compared with higher variability in guideline DNT studies, raising the question of whether there is inadequate control of experimental and environmental variables (Raffaele et al., 2008). These comparisons may not be appropriate if the published studies are not conducted according to OECD or US EPA DNT guidelines, which require testing three different sets of 160 pups on multiple behavioral endpoints (OECD, 2007; US EPA, 1998a). For example, some of the academic studies cited in OECD (2007) and Raffaele et al. (2008) include studies in adult males (Crofton et al., 1991), acute age-relevant positive control studies that are not DNT design (Crofton et al., 2008; Marable and Maurissen, 2004), studies that report standard errors across multiple studies (Buelke-Sam et al., 1985), and studies that used substantially-fewer pups and/or a within-litter design that utilized a total of 10 litters per test chemical (Ruppert et al., 1983, 1985; Crofton et al., 1993). Thus, direct comparisons of variability in these types of studies and DNT studies must be done with considerable caution.

To gauge data quality, it is useful to view data from the control animals from a variety of statistical perspectives, including the standard deviation (SD), standard error of the mean (SEM), coefficient of variation (CV), range, and distribution. In addition, the data set (e.g., group mean and SD) should be evaluated to determine whether it varies appropriately with

repeated testing or with age and sex. Next, a data set can be compared with data from other laboratories, but it is important to consider whether the test circumstances are appropriate for comparison. For example, data are generally more variable when the circumstances dictate an extended test day, due to diurnal variation in activity (Raffaele et al., 2008), or testing of animals within the cohort over multiple days, compared to testing fewer animals within a narrow time span all on the same day. It is also useful to evaluate consistency for control animals across a number of studies performed under similar conditions (i.e., historical control data). If such comparisons indicate that the data are more variable than expected, the investigator should review the test conditions and procedures in detail and spend time watching how the animals are being handled and the test is being performed. This should include detailed examination and testing of each test apparatus, because the environmental conditions may vary or there may be differences in equipment function or defects in manufacture.

In general, each litter is represented by one male and one female, resulting in a sample size of 20/sex/dose group. In the US EPA (1998a) and OECD (2007) DNT guidelines, a different cohort of animals (one pup/sex/litter) may be assigned to different behavioral tests. In addition, some F1 animals may be selected for blood collection, while others are taken to necropsy for perfusion fixation, necropsy, and tissue collection. In the EOGRTS (OECD, 2012a), the majority of DNT evaluations (behavior and neuropathology) are performed in the same cohort of one male or female/litter (10 males and 10 females/group). The dams also must be attended to and the surviving F1 animals disseminated to new housing.

For motor activity, the workday over the course of the study can range from one test session (e.g., 30 or 60 min) to six or more sessions, depending on the duration of the test and the number of devices used (Fig. 1). Consequently, the test may be completed between 8 and 9 a.m. on some days and may continue until 4 p.m. on other days, which would likely significantly increase variability due to diurnal patterns of activity (Raffaele et al., 2008) (Fig. 1). One should examine the results to determine whether this source of variability is a problem that must be managed. For example, additional devices with the same calibration could decrease the number of test sessions needed over the course of several hours; however, there could be increased variability associated with the additional time required to load more animals.

3.3. Approaches for evaluating motor activity and auditory startle data

The remainder of this presentation was specific to motor activity and the auditory startle response, including a description of the test procedures and discussion of data for control animals from guideline DNT studies performed in my former laboratory at Bayer. In this context, motor activity refers to spontaneous movement of the animal in a controlled environment. The commercial equipment that is commonly used may provide a variety of measures, but this generally consists of some measure of locomotor or ambulatory activity and potentially a measure of non-locomotor movements, which may include anything from rearing to tremor. US EPA (1998a) and OECD (2007) DNT guidelines state that this test should demonstrate habituation in controls over the course of the test session, as an indication of the animal's ability to adapt to its environment—a simple form of learning.

Motor activity is an apical test, because it can be altered through effects on a number of systems. Interpretation of results may take into account the nature of the effect, including persistence after treatment is discontinued, whether activity is increased or decreased, and whether there is an effect on habituation. The device used for this study design must be automated and suited to testing rats from PND 13 to 60. It must also be reliable and suited to provide consistent results, including habituation over an appropriate duration of time (approximately 30 min to 1 h). US EPA (1998a) specifies that the test session should be long enough for motor activity to reach asymptotic levels by the last 20% of the session. The most common choice in testing laboratories is some type of open-field device, while the figure-8 chambers (Ruppert et al., 1985) have been used in the Bayer laboratory.

The conditions and variables that should be controlled for a test of motor activity include background and intermittent sound, light and visual cues, vibration, time of day, and scent in the device. It is appropriate to attend to some of these issues when the room and test equipment are set up, while others are checked before testing begins on each test day or between test sessions. There is generally a practical limit to how much intermittent sound can be reduced in the test environment, so it is common practice to use a white noise generator to provide consistent background noise that is evenly distributed throughout the area. Scent is a very potent and important factor for rodents, so it is important to manage this variable carefully through the choice of procedures used to clean the equipment between animals, balancing the allocation to devices across treatment groups, and testing sexually mature males and females in different equipment or on different days.

The level and type of activity vary considerably, depending on the size and shape of the device, as well as the distribution of photocells in the device. The figure-8 maze has eight pairs of emitters and detectors placed at specific locations, so activity is measured as beam interruptions when some part of the animal passes through that location. Motor activity is measured as any movement that interrupts a beam, while locomotor activity (ambulation) is measured when the animal breaks two different beams by moving from one location in the maze to another. Our results for control male Wistar rats tested in figure-8 chambers on PND 13, 17, 21, and 60 illustrate a consistent pattern of activity, both within a study and across a complement of several studies performed under the same conditions (Table 3). This pattern includes a progressive, age-related increase in motor activity. On PND 13, the level of motor activity is quite low and the SD approximates or exceeds the group mean, which results in a coefficient of variation from 76% to 138%. Looking at the data in this way indicates that the results are highly variable; however, evaluation of results within and across nine studies shows the results are very consistent, with group means ranging from 46 to 90 counts, compared to group means of 127 to 263 counts on PND 17 (Table 3). Looking at such data, it is easy to determine which results are from PND 13 animals, compared to all other age groups, and it is clear which results are for adults. In my lab, the results within any given study also showed clear age-related differences in within-session habituation, with habituation present on PNDs 17, 21, and 60 but not on PND 13 (Fig. 2). While other laboratories using different equipment and test conditions do not always observe habituation at PND 17, they do consistently report the absence of habituation at PND 13, and consistently report habituation at PND 21 and later (Raffaele et al., 2003; Ruppert et al.,

1985). Furthermore, the results for adults show a consistent pattern of habituation in all nine studies (Table 4).

The auditory startle response (ASR) is measured as the magnitude of the downward force a rat exerts in response to an abrupt acoustical stimulus of sufficient intensity (Davis, 1980; Davis and Eaton, 1984). The ASR is a well-characterized, reflex-based response with a short latency to onset following the presentation of a startle-eliciting stimulus, which is measured using either a load-cell force transducer or an accelerometer (Davis and Eaton, 1984). A load-cell force transducer measures both static (gravitational) force (e.g., the weight of the animal) and movement (typically at 1-msec intervals). This device is calibrated using static weights, and the response is measured in grams. The sensitivity of a load cell is set by the operator to optimize resolution of the output within the device's dynamic range (e.g., to a maximum of 500 g) for young rats or reduced (e.g., to a maximum of 2000 g) to avoid exceeding the capacity of the device to measure the peak response for adults. By comparison, an accelerometer only measures the dynamic movement of the platform, and the ASR is generally reported in millivolts (mV) or arbitrary units. With this device, the sensitivity setting and the enclosure platform are different for young rats versus adults. As a result of these adjustments, the results with an accelerometer are not directly comparable across age groups; in fact, the magnitude of the response (reported in mV or arbitrary units) may appear to be similar or greater in young animals than in adults.

Important factors to manage include calibration of the equipment used to generate the startle-eliciting stimulus and to measure the ASR. As with motor activity, the test guidelines specify a measure of habituation. To achieve this, the test conditions (e.g., inter-trial interval) are optimized to achieve habituation over the course of the 50-trial session (5 blocks of 10 trials each). The latency of the ASR is also assessed to verify that the response being measured is the primary startle response, which is generally a factor only when the ASR has decreased with habituation and the true peak startle response may be less than other downward movements that occur during the sample period.

Using data from the Bayer laboratory, Tables 5 and 6 show ASR results using load cell detectors, with a 120-dB burst of broad-spectrum noise to elicit the response and a 9-s inter-trial interval to produce habituation. Our results from nine DNT studies show modest habituation in PND 22 control Wistar rats over the course of the 50-trial session, with an average 6% to 35% decrease in group mean response amplitude from the first to the last block of trials (Table 5). The results from these studies show a consistent pattern, based on group means, standard deviations, and habituation. For the same complement of studies, the ASR in PND 60 control males is much greater (39%–68%) than on PND 22, and the magnitude of habituation is also much greater (Table 6). The absolute difference in total and within-session group means among these studies is greater on PND 60 than PND 22, due to the much greater dynamic range of the response for individual animals and individual trials in adults, relative to weanlings (Tables 5 and 6). For females, the response is comparable to males on PND 22 but much less than males on PND 60 (data not shown). This difference between adult males and females cannot be accounted for by differences in body weight, because the output from the load cell during the first 8 ms from the onset of the stimulus (before ASR onset) is subtracted from the measure.

In summary, these examples emphasize that it is important to examine DNT data quality using a number of factors and to compare results from a single study with other data within and across labs performed under comparable conditions to determine whether or not variability is excessive. If appropriate, there are a number of steps that can be taken to reduce variability, in order to achieve high-quality and consistent results. However, the logistical considerations addressing DNT guideline specifications (e.g., sample size, age of testing, habituation design) can contribute to the “intrinsic” variability in appropriately designed studies, which may not be observed in other study designs.

Therefore, when reporting and reviewing DNT data, it is important to describe and understand the test conditions and procedures used and what the measures represent. It is also useful to provide historical control data for reference, with sufficient details provided to support the suitability of those data for comparison.

4. New insights into analysis of highly variable data: motor activity as case study (Wayne J Bowers)

This presentation provides different opinions and additional insights from those presented by LS by focusing on discussion of measurement issues and important test equipment and software factors that may affect motor activity measurement. In addition, the critical biological questions addressed by motor activity test data were described, as well as statistical considerations in the analysis of motor activity data. Considering the excess variability in PND 13 activity data, as indicated by the higher CV at PND 13 compared to other ages (see LS section above), specific emphasis was placed on the nature of pre-weaning activity data, especially PND 13 data, from multiple laboratories to illustrate some of the challenges in analyzing motor activity data from DNT studies.

4.1. Measurement of motor activity

As previously described, motor activity is an apical behavioral test that reflects multiple neural inputs, including motor, sensory and cognitive/reactivity. Motor activity data do not allow one to determine specifically which underlying functional capacity is affected by chemical exposure; only that some function is affected. This non-specificity is a disadvantage for characterizing the nature of treatment-related effects; however, this non-specificity is also an advantage, because multiple functional systems are assessed in motor activity testing.

4.2. Activity equipment and software issues

Both US EPA and OECD DNT test guidelines state that motor activity testing should be conducted with automated test equipment. There is a variety of test equipment available for motor activity testing. The most commonly used recording system is photocell-based equipment, but video-based systems are becoming more common. Test chambers are typically square or rectangular open-field-type chambers; other chambers of other shapes are employed less frequently (e.g., figure-8 chamber; Ruppert et al., 1985; Sheets et al., 2016, Crofton et al., 1991). For photocell systems, a bank of transmitters and receivers is mounted outside the activity chamber. Photocell systems use beam breaks in XY and sometimes Z

coordinates to generate raw data, which are then aggregated by system-specific software (usually vendor proprietary software) to generate measures of motor activity, such as ambulatory counts, ambulatory time, ambulatory distance, vertical movement counts, and vertical time. Some systems also provide measures of small movements such as grooming. For photocell-based systems, the number of and spacing between photocells will affect sensitivity to movement; this design consideration can be an important factor in determining photocell measures. This will be important for small animals (such as pre-weaning rodents), where large spaces between photocells may result in small (i.e., young) animals detected as missing by system software. Video-based systems generate data by monitoring the change in location of an image of the animal within the test chamber. The raw data are then aggregated by system-specific software to generate motor activity measures similar to those produced by photocell-based systems. For video-based systems, it is important to optimize ambient light settings to ensure acceptable contrast between animals and the background. It is important to recognize that software settings for both photocell and video systems play a critical role in determining the activity measures generated. Indeed, the actual measures of motor activity reported in DNT studies (e.g., ambulatory counts, vertical counts) are actually generated by equipment software using algorithms that convert photocell beam break sequences into activity measures, or image movement into activity measures. Because of this, most systems permit customization of software settings to take into account animals of different sizes.

There are a number of potential impacts of the equipment and software settings. Because activity test systems can generate absolute activity counts using different algorithms, absolute activity measures might not be comparable among laboratories using different systems or even within laboratories where software settings vary among studies. Similarly, direct comparison of activity measures between ages within the same study may not be comparable, because the same software parameters used to generate activity data at different ages (e.g., where animals differ greatly in size) may exhibit differential sensitivity to movement. Thus, the generated absolute activity measures (e.g., ambulatory counts) may reflect actual differences in activity level, combined with variations in equipment sensitivity to movement. Fortunately, most systems permit the adjustment of software settings and re-analysis of raw data collected by test instruments. Equipment and system settings may also set constraints on the ability to compare absolute values from concurrent controls with historical control values within a laboratory.

Motor activity testing is conducted to assess three primary biological questions: behavioral ontogeny, habituation, and motor function. Behavioral ontogeny for motor activity is an important assessment of neurobehavioral development over the pre-weaning period, which usually is reflected by increases in motor activity from PND 13 to PNDs 17 and 21. Treatment-related alterations in the pattern of pre-weaning motor activity may be related to generalized, non-specific, developmental delays, and this would be associated with other indicators of developmental delay (e.g., altered growth, delay in developmental landmarks). In the absence of other indicators of developmental delay, altered activity in the pre-weaning period generally is interpreted as a disturbance in nervous-system development.

Second, motor activity data provide a measure of habituation, an assessment of non-associative learning that is critical for adaptive behavior. US EPA (1998a) and OECD (2007, 2012a) test guidelines state that habituation in motor activity should be evaluated.

Habituation is characterized as a decrease in response to a stimulus following repeated exposure to that stimulus, and this short-term adaptation plays an important role in filtering the environmental stimuli that influence behavior. Habituation is not measured directly in motor activity testing, but rather is inferred by the pattern of activity over a test session. Under normal circumstances, habituation is reflected by a decrease in activity over the course of a test session (Fig. 3). Thus, the slope or rate of decline in activity measures provides an index of habituation. Treatment-related effects on habituation can either decrease habituation (flatter slope, compared to controls) or increase habituation (steeper slope), with either of these outcomes indicating a potential treatment-related effect on non-associative learning.

Third, motor activity testing provides a general measure of motor capacity and can assess whether treatment impairs overall motor function. Hypoactivity or hyperactivity in animals can indicate disturbances in motor function or disturbances in reaction to environmental stimuli. Typically, hyper- or hypoactivity, without an effect on habituation, could be reflected by parallel shifts higher or lower in the intrasession activity pattern. In contrast, altered habituation can be characterized by a change in the intrasession activity pattern, which can result in increases, decreases, or no change in cumulative total activity levels for a test session (Fig. 3).

4.3. Primary measures and statistical considerations

There are a number of important issues to consider when performing motor activity testing. Some key measures that can be evaluated are locomotion (e.g., ambulatory counts, ambulatory time, ambulatory distance) and rearing (e.g., vertical counts, vertical time). In some systems where photocell counts are reported, these are usually treated as measures of ambulation (locomotion). Most activity measures are highly inter-correlated (e.g., ambulatory counts and ambulatory distance), and this should be considered in data analysis. Some systems report total movement, which incorporates all forms of movement, including ambulation (horizontal), rearing (vertical), when measured, and small (localized) movements (e.g., grooming). This total movement measure should not be used by itself, because it could mask biologically-meaningful alterations in specific motor measures. For instance, increases in small movements that are characteristic of stereo-typy (i.e., increased grooming and small repetitive movements) are incompatible with locomotion, so summing all movements may indicate no motor change, when both small movements and locomotion have changed in opposite directions. Note that some laboratories may use similar terminology in reporting total movement as total cumulative ambulation for the entire duration of the test session (e.g., total ambulation for test session). This total or cumulative activity for one specific activity measure differs from the total movement mentioned above, which indicates the total of all movement measures (e.g., ambulatory and fine movements).

In most cases, the generally-accepted statistical approach for analyzing motor activity data is repeated measures analysis of variance (RM-ANOVA), with treatment, sex, and test time

blocks included as factors in the analysis. Alternative statistical approaches, such as trend analysis, multivariate analysis, and regression approaches may also be used, as long as they include an assessment of the time blocks factor, which is critical to assess habituation. Including data for male and female offspring with sex as a factor in the data analysis is preferred, because separate analysis of male and female data precludes any direct assessment of sex differences in response to treatment effects (i.e. the dose $\times$ sex interaction). Note that when more than one pup/litter is tested, the litter factor should also be included in any data analysis. In addition, if one pup/litter is selected for testing (i.e., one male or one female per litter), combining both sexes in a single analysis may increase power to detect treatment effects. It is also important that the variability and nature of the data distribution are checked to ensure that the ANOVA is appropriately applied.

A number of statistical errors have been found commonly in reports of DNT motor activity data. First, reporting only cumulative session data (sometimes referred to as the total session data) precludes an evaluation of treatment effects on habituation. In addition, cumulative session results can be misleading where there are within-session treatment interactions with time or sex. Second, as already mentioned, reporting statistical results where males and females are analyzed separately is problematic, because it precludes direct assessment of any interaction with sex. Third, some reports follow a significant sex by treatment interaction in the initial global ANOVA (i.e., both sexes combined) with completely independent and separate ANOVAs on male and female data. This is problematic, because the error terms used in these separate ANOVA analyses are not appropriate and can produce inaccurate statistical p-values. Statistical tests to follow-up on sex by treatment interactions should be constructed within each sex using the original global model, so that subsequent analysis and contrast procedures are based on error terms from the initial global ANOVA. The power of this analysis for each sex is higher than when completely separate independent ANOVAs on each sex are performed. Fourth, it is important to check data for extreme values that can distort estimates of the group mean and result in unnecessarily variable data and reduces the statistical power to detect treatment effects. If extreme values are encountered, comparing results of data analysis with and without the extreme values can help determine the impact of the extreme values and how robust are conclusions in the presence and absence of extreme values. In addition, non-parametric statistics using ranks can help reduce the impact of extreme values.

4.4. High variability in PND 13 motor activity data

Many laboratories find that motor activity at PND 13 is highly variable. Assessments of the CV (a method to normalize variability between data sets by expressing variability relative to the group mean) indicate that PND 13 activity data typically have higher relative variability than activity data from older animals (Table 3). This higher variability at PND 13 decreases statistical power to detect treatment effects, relative to statistical power at older ages. However, it is not clear that the CV as a measure of relative variability at PND 13 is appropriate. This is because the CV is based on the group standard deviation divided by the group mean. Because the group mean is a reliable indicator of the group central tendency only when data are reasonably normally distributed, the CV is a reliable approach to comparing relative variability among data sets only when group data are normally

distributed for all data sets. As illustrated in Figs. 4–6, using control data from three independent laboratories, PND 13 activity data are unlikely to be normally distributed; rather these data are usually markedly skewed or bi-modal. This is less likely to be the case for activity data for older animals. As a result, the CV is not a reliable means to assess relative variability between PND 13 and older animals. Indeed, Table 3 shows that the magnitudes of the raw standard deviations at PND 13 are actually numerically lower than those at other ages. While the CV appears higher at PND 13, this is likely due to the highly skewed PND 13 data, rather than an actual increase in relative variability. Thus, there remains a question of whether PND 13 data are actually more variable or simply has different properties from activity data from older animals. In addition, these data distribution issues have implications for statistical procedures used for analyzing PND 13 data, as well as any direct comparisons between PND 13 activity data and activity data for older animals.

For PND 13 activity data, there are a number of biological reasons to expect that the data distribution will not be normally distributed. First, PND 13 animals do not have a fully developed neuromotor system and normally exhibit a different type of ambulation, compared to older animals (Altman and Sudarshan, 1975; Bolles and Woods, 1987; Vinay et al., 2002; Vinay et al., 2000; Clarac et al., 2004). Because the neuromotor system is developing rapidly at this age, some PND 13 animals may have more advanced neuromotor functioning than others, resulting in subsets of animals with different motor function capacities. Based on neuromotor development, one would expect that a high proportion of PND 13 animals would have little locomotor (horizontal) and rearing (vertical) activity. Indeed, motor activity at PND 13 may actually reflect the proportion of PND 13 animals who have achieved a stage where ambulatory capacity has developed (Altman and Sudarshan, 1975). Such differences in biological development will likely affect the data distribution at PND 13. Second, PND 13 rats normally have not reached the developmental stage where eye opening is completed (a development that usually occurs around PND 14–16, depending on the strain). Because of this, the visual sensory information available to PND 13 rats during motor activity testing will differ from the visual information available to older animals. Although the specific role of visual sensory information in motor activity behavior in pre-weaning rats is not clearly identified, and albino rats generally have poor visual acuity, it is nonetheless clear that, due to developmental stage, the overall sensory environment differs between PND 13 rats and older rats. As with neuromotor capacity, subsets of animals at PND 13 may achieve eye opening by the PND 13 test, and thus, activity data distribution at PND 13 may reflect subsets of animals at different stages with respect to visual system development.

The impact of developmental age on DNT motor activity data is readily seen in pre-weaning rats. Fig. 4 shows motor activity data for 10-min time blocks for PND 13 (top), PND 17 (middle), and PND 21 (bottom) rats collected in a figure-8 chamber with photocells. While data for PND 17 animals are reasonably normally distributed at these time blocks, data for PND 13 rats are not normally distributed and are negatively skewed, with a high proportion of zero values. This reflects a large proportion of PND 13 animals exhibiting little or no movement. This pattern for PND 13 animals is also evident in data from another laboratory using different activity equipment (Fig. 5 top), with many animals exhibiting no movement. However, PND 17 (Fig. 5 middle) data also reveal a large proportion of animals with low activity levels. This non-normal data pattern for PND 17 is probably related to equipment

sensitivity, because similar results are evident from another laboratory using similar equipment (Fig. 6 top). While a full analysis of the role of test equipment is beyond the scope here, differences in software and photocell density that determine activity data may account for the results. Finally, data from PND 35 animals (Fig. 6 bottom) show that activity data at this age exhibit the expected normal distribution. Overall, these data indicate that pre-weaning activity data, especially PND 13 data, have a pattern different from post-weaning animals, consistent with the idea that developmental stage constrains the nature of pre-weaning motor activity. Further, the nature of the data properties indicates that pre-weaning data analysis (especially PND 13 data) must take this into account.

There is another consequence of the differences in biological stage of development between PND 13 rats and older rats that was not considered in the previous section. Because of the differences in neuromotor capacity, it is also possible that activity measures collected are not biologically comparable between PND 13 and older animals. As a result, comparisons between activity measures at PND 13 and those gathered at older ages may actually be comparing different motor-based functions, despite employing the same measurement, instruments, and procedures. By way of analogy, this may be comparable to assessing crawling in infants and walking in children by measuring distance moved. While both provide a measure of distance or time in movement, they are fundamentally different behaviors; therefore, it raises the question of whether it is appropriate to directly compare PND 13 activity data with activity data for older animals.

While it seems that direct comparison between activity data from PND 13 and older animals may be problematic, because of differences in developmental stage and capacity as well as differences in data distribution properties, PND 13 data may nevertheless provide useful information on motor system ontogeny. Indeed, the biological capacity differences that characterize PND 13 versus older ages identify traits of motor ontogeny that are associated integrally with nervous-system development (Altman and Sudarshan, 1975; Bolles and Woods, 1987; Vinay et al., 2002; Vinay et al., 2000; Clarac et al., 2004). That PND 13 motor activity levels and pattern (including statistical properties of data) differ from later ages is consistent with the age-dependent development of motor function. Although there is a basis for expecting that PND 13 activity may provide a measure of motor-system development, further work is needed to understand the extent to which PND 13 data alone will be sensitive in detecting biologically-relevant effects on motor function ontogeny. However, it is clear from this example that data should be inspected carefully before applying any statistical approaches, including those that may be well-accepted for these data (e.g., RM-ANOVA; Ruppert et al., 1985; Holson et al., 2008). In the specific case of PND 13 data, preliminary examination of a limited data set from three different laboratories suggests that using other statistical approaches, such as mixture models (Boos and Brownie, 1991; He et al., 2014), may prove useful in characterizing alterations in the ontogeny of motor development.

5. Optimizing experimental design and statistical analysis: auditory startle response (ASR) as a case study (Kathleen Raffaele, Thomas Vidmar, Edmund Lau, Abby Li)

DNT studies conducted for regulatory purposes incorporate multiple endpoints that vary in the specific dependent variables measured, as well as the metrics used to evaluate results. Regulatory guidelines provide flexibility in study design, placing responsibility on the individual investigator to precisely define the hypothesis being tested and design appropriate analyses. Available equipment can record multiple dependent variables, not all of which convey the most pertinent biological information for the purposes of the test guideline. In order to maximize the useful information obtained from each study, the experimental design and subsequent analysis of data should be optimized to test the key hypothesis of interest based on the test guidelines.

Only a brief overview and summary points are included in this manuscript. Although general guidance for statistical analysis of DNT studies is available (Holson et al., 2008), recommendations on the practical implementation of statistical approaches to DNT data sets have been limited. Differences in guideline specifications suggest that substantially different biological hypotheses can be applied within the general guidance provided. The ASR habituation test, as described above by Dr. Sheets, has been used here as a case study to illustrate specific issues regarding DNT study design and behavioral data analysis.

The US EPA DNT guideline (US EPA, 1998a) and the OECD EOGRTS (OECD, 2012a) specify that the ASR test should be designed to evaluate habituation in 10 males and 10 females drawn from 20 litters/dose, such that all 20 litters/dose group are represented (i.e., 1 male or 1 female from each litter; total of 80 animals). In contrast, the OECD DNT guideline (OECD, 2007) only states that some test of sensory and motor function should be conducted at two ages in 1 male and 1 female/litter (i.e., 20 males and 20 females/dose group; total of 160 animals); specifically, the ASR habituation is one of several tests that can be conducted (Table 2). If ASR habituation is being tested, then, as described above by Sheets, the inter-trial interval should be constant so that the control animals will habituate.

Available ASR testing equipment can gather data on many dependent variables. Two or three of these variables will capture most biologically-relevant information within the context of a regulatory DNT study. We identified three variables of primary interest for this evaluation: mean peak amplitude for each of 5 blocks of 10 trials, change in peak amplitude across blocks of trials (habituation), and latency to peak amplitude. Average peak amplitude across all 50 trials is generally not of primary interest, because it is confounded by habituation occurring across the blocks (and thus also has increased variability). By contrast, latency to peak amplitude is not expected to change across blocks, and is thus also useful as verification that the peak amplitude is measuring the startle response rather than other movement. For example, if the average latency to peak is longer than 35–40 ms, the trace recordings for individual data should be checked to determine if the peak amplitude selected by the computer system corresponds with the actual startle response or some other movement.

From a regulatory perspective, the biological question of interest is whether or not a treatment-related change occurs in the endpoints being measured. Although ASR habituation is a key endpoint of interest, treatment-related changes in ASR habituation are not always formally evaluated in submitted study reports, which generally evaluate only the absolute change in mean peak amplitude for each individual block of trials, and the average across all trials. Statistical analysis should thus be designed to evaluate treatment-related changes in all relevant endpoints, including the pattern of ASR habituation.

For behavioral data such as ASR habituation, in which each animal is tested repeatedly, the RM-ANOVA has many advantages over performance of multiple one-way ANOVAs (a commonly reported statistical analysis), because the effect of a given test article dose on all relevant factors can be evaluated using a single analysis, and in particular, the change in ASR across trial blocks (habituation) is tested directly. Separate one-way ANOVAs for each trial block do not directly test the effect of the test chemical on the pattern of habituation. Mixed model statistical approaches are available to evaluate different covariate structures, so that the homogeneity of variance and constant correlation are no longer required (Holson et al., 2008). The RMANOVA for ASR habituation should always include trial blocks and treatment group as factors; the power of the analysis can be increased further by including sex and different time points (i.e., ages) as factors within the same analysis. As described in greater detail by Dr. Bowers, statistical tests to follow-up on sex by treatment interactions should be constructed within each sex so that subsequent analysis and contrast procedures are based on error terms from the initial global ANOVA. In cases where both males and females from each litter are included within the test group, the litter also needs to be included as a factor in the analysis. Given the additional issues (see below, and also discussions by Drs. Sheets and Bowers) regarding combining data across age groups, it may be better to analyze the data for each age group separately, to limit the number of factors to no more than three.

Interactions between dose and other factors are of primary interest in understanding the relationship between treatment and the measured effects, and different decision trees for selecting a suitable statistical approach are available. Combining different factors into a single global analysis requires that the investigator ensure that the equipment setting and experimental design support an analysis that pools all the data together (Holson et al., 2008). For example, if equipment or equipment settings are varied based on the age and/or size of the animal, it may not be appropriate to combine data into one analysis; accordingly, data for each day (or size) will need to be analyzed separately. Similarly, if males and females are tested on different days or at different times during the day, then the interpretation of a sex effect or dose interaction with sex is inextricably confounded by the divergent testing conditions (Maurissen, 2010). The most practical solution is to counterbalance the time of testing across dose levels and the sex of the animals.

In summary, the statistical analysis should be guided by the primary biological hypothesis being tested and should be supported by appropriate experimental design. In the case of ASR habituation, the primary biological hypothesis being tested is the presence of treatment-related changes in startle amplitude, latency, and habituation. It is crucial to look at the raw data to optimize valid approaches for evaluating the data. The RM-ANOVA has

many advantages over other analyses commonly used in DNT studies, including the increased power of the analysis, the ability to evaluate changes in response over time, and the ability to consider multiple factors in a single analysis. However, it is important to consider features of equipment and experimental design, and the nature of the data results that may influence the types of analysis that can be performed appropriately.

6. How reported data and methods affect evaluation: learning and memory as case studies. (Virginia Moser and Angela Hofstra)

This presentation focused on tests of learning and memory used in laboratory animals in the context of testing in DNT regulatory studies. A background on testing and neurobiology of cognition was provided, along with information and considerations in the selection of tests. Two case studies, specifically with data from the Morris water maze (Morris, 1984) and a water M-maze, were provided to illustrate new approaches to looking at details to inform interpretations of DNT data sets.

Tests of cognition are the least standardized of all tests included in the DNT guidelines. They are included in the OECD 426 (OECD, 2007), US EPA 870.6300 (US EPA, 1998a), ICH S5A (ICH S5-R2, 2005), and US FDA Pediatric (2006) guidances. Notably, cognitive evaluations are not included in the OECD (2012a) 443 EOGRTS guideline. There is considerable flexibility in the selection and conduct of tests; however, the choice should take into account knowledge of the test chemical, and the rationale behind the choice of test should be detailed. Both the US EPA (1998a) and OECD (2007) test guidelines specify that DNT data should demonstrate evidence of learning and memory.

Cognitive constructs, as they apply to both humans and animals, have long been described in the field of experimental psychology (e.g., Hulse et al., 1978). When testing laboratory animals, it is important to understand that learning and memory cannot be observed directly; they can only be inferred from changes in behavior. For example, when a rat chooses to go left or right, it may be taken as evidence of learning the consequences of that choice; however, the investigator should understand that other explanations that do not involve learning may also apply. Furthermore, all of the test methods used to measure cognition in animals are influenced by the attention, perception, motivation, sensory, and motor abilities of the subject. In simple terms, learning can be considered as a change in behavior that results from experience, while memory is the retention of that learned behavior over time. There are multiple types of learning (e.g., spatial, conditioned) and memory (e.g., short- or long-term) that are assessed by different test procedures. No one test can measure all forms of cognition, and some chemicals may specifically impact different types of cognition (Tilson and Mitchell, 1984).

Cognitive processes represent the responses of neural networks integrated across brain regions and neurotransmitter systems. There is no simple one-to-one correspondence between a particular brain region and a single test. Instead, there are multiple and overlapping neuroanatomical and neurochemical substrates that influence the outcome of specific cognitive tests. Important structures with known functions related to learning and memory include the cerebellum (associative and spatial learning, working and procedural

memory), hippocampus and limbic system (associative, spatial and emotional learning, short- and long-term memory, working memory), and prefrontal cortex (long-term and working memory, inhibitory control, goal-directed learning) (see for example, Clark et al., 2000; Eichenbaum, 2000; Gold, 2003; Squire, 2004).

There are a number of considerations when deciding on approaches for cognitive testing, including positive aspects as well as limitations of the potential test methods. Test equipment varies in type, size, and shape, and data collection systems can range from a technician using a stopwatch to sophisticated video tracking systems. While many laboratories use commercially-available instruments, some are fabricated in-house; in such cases, comparisons across laboratories can be difficult. There are limitless parametric designs (e.g., interval between and number of training trials) and procedural details (e.g., water temperature, tank size). In general, more complex or difficult tasks may be more sensitive to detecting chemically-induced effects (Vorhees and Williams, 2006; Blokland et al., 2006; Jonasson, 2005; Perrot-Sinal et al., 1996; Roof and Stein, 1999; Vorhees et al., 2008). Modifications that increase the difficulty of the task could include adding long-term memory tests with longer intervals, or including a working memory assessment. Reversal learning, which requires the animal to “unlearn” the previous contingency and learn a new one, may uncover deficits that could otherwise go undetected. Finally, in all cognitive tests, performance controls are necessary to rule out sensory, motor, or motivational differences, because these could affect an animal’s ability or motivation to perform the task at hand.

Even more considerations are necessary in the selection of cognitive tests for use in DNT studies (Ehman and Moser, 2006). “While the guidelines recommend that the choice of test take into account the mode of action or anatomical target of the test chemical, in reality such information is often not known, and most laboratories do not have capabilities to develop and maintain historical control data for a large number of different types of cognitive tests. Methods should meet the specifications of the regulatory test guidelines. Test durations can range from one trial (one day) to months of testing, which must be integrated into the other concurrent DNT tests. The age and development of the animals should also be considered; for example, swimming ability that may be delayed by treatment could impact performance in a water maze. The behavioral history of the animal can also influence performance, especially if there was previous cognitive testing. Because of the potential for carry-over, or persistence, of learned behaviors, OECD Guidance Document 43 (OECD, 2008) recommends that different animals should be used for the young and adult cognitive tests. There are also practical considerations, in that some procedures are not feasible in regulatory testing. For example, some test methods require restriction of food or water, which interferes with normal growth when applied to weanling animals. Long-term training can also require an extensive level of effort over months.

Cognitive data submitted for regulatory evaluation should be summarized and presented in a way that shows a process of learning; for example, a graph or table with group data across days of training, not just the last day. There should also be clear evidence of memory; for example, the behaviors measured are similar to the last day of training, or else different from the first day of testing before training began. The control group should show a significant change in behavior (i.e., a learning curve); otherwise, the data from treated groups are

uninterpretable. The protocol should include performance controls (e.g., visual and/or motor abilities) and both sex and age differences should be evident where expected; for example, males tend to perform better than females in spatial tasks such as a water maze (Jonasson, 2005). Finally, appropriate, and often complex, statistical analyses are necessary. The statistical approach should consider factors such as the data type and distributions, repeated measures, changing variance over time, and litter representation, with sex nested within litter as appropriate (Holson et al., 2008).

An informal survey of DNT studies submitted to US EPA and Health Canada shows that the tests most often used are the following: passive avoidance (associative learning), Biel or Cincinnati water maze (sequential learning), Morris water maze (spatial learning), and “letter” (M, Y, E, T) water mazes (positional discrimination). While most of these tests are also widely used in the toxicological and neuroscience literature, there are no standardized procedures for most aspects of testing. In this presentation, further explanations and case studies for cognitive data were presented for the Morris water maze and the M-maze.

The Morris water maze (MWM) consists of a circular pool with a hidden platform positioned under the surface that the animal has to find using external spatial cues positioned on the walls surrounding the maze. Acquisition is assessed using multiple trials per day(s), with measures such as latency and path length to find the platform decreasing over trials. A reference memory test, or probe, is conducted some time later, with the platform removed, to measure the animal’s bias for the target quadrant. While some laboratories may conduct multiple memory probes during the course of training, this experimental design has not been employed in laboratories following the DNT guidelines. Typical endpoints include latency and distance swum (path length) to reach the platform for acquisition and time in quadrants for memory testing, as well as spatial swim patterns and other variables that are often derived using video tracking software. Considerations in the conduct of MWM testing include procedural differences (e.g., length of inter-trial interval, procedure if animal does not find the platform within a specified time), performance controls (for sensorimotor ability), dependent variables (often defined by software), and data review (e.g., initial group differences, potential outliers including animals showing excessive floating or diving). Optimal parameters have been described and summarized in recent reviews (Vorhees and Williams, 2006, 2014).

A case study of MWM data was presented that was based on a submitted DNT data set (but without any identifying features). In this study, the report stated that there was no difference in latency on any day of testing. A closer look at the data (Fig. 7) showed multiple issues that influence interpretation and that could obscure potential treatment effects. In this case, the issues included: (1) no apparent learning in the control group, as evidenced by latencies that did not decrease with training; (2) apparent group differences on the first day of training; and (3) high variability in one group on one day. Additional information can enlighten interpretation of such data. For example, the data for the first day actually represented an average of four trials. Review of those first four trials showed no group differences on the very first trial, but over the next three trials there was a clear separation of the treatment groups. These initial differences on the first day could reflect an impact of treatment. A review of the concurrent control data is necessary to understand why there was

no clear evidence of learning. In such instances, historical control data may be informative, as well as a close evaluation of study procedures and equipment. Finally, individual animal data that include the single suspicious data point could help identify an outlier that may have been due to factors such as the animal floating, equipment malfunction, or data error.

Another case study of cognitive data was presented involving the water M-maze. In this task, the animal is placed in the central stem of the M and must swim to a choice point, turning right or left to reach an escape platform. This may be described as a reference memory task, because the same arm is designated as the correct response throughout training (e.g., always turn right). Latency to escape and the number of errors per trial are recorded. Such positional learning depends on both proximal and internal cues. Acquisition is typically assessed using multiple trials per day(s) to reach some pre-defined criterion, or until a maximum number of trials has been administered. At some time later (e.g., one week), the subjects are re-tested for memory (i.e., retention of the task); however, only the first trial is actually a test of memory, and any subsequent testing constitutes relearning the task. Animals that did not achieve criterion performance during acquisition should not be tested for retention, since they did not learn the task as defined by the protocol; however, this practice may lead to differing sample sizes across groups, and may introduce bias by testing a selected subset of the population. As with the MWM, considerations for conducting M-maze tests include procedural details, performance controls, dependent variables, and data review. In this case, the number of errors is frequency count data (i.e., not continuous) and is not normally distributed, because there can be a large number of zero counts (left-censored since the number cannot go below zero).

A case study was presented in which the submitted report stated that there was no difference in average trials-to-criterion across treatment groups (Fig. 8). While this was the case when analyzing the data using ANOVA, when individual data for trials to criterion were plotted, it was evident that there were differences in the group distributions and that medians were more representative of the group performance than were the means. Non-parametric statistics (e.g., Kruskal-Wallis) used on the same data showed a treatment-related effect in both the mid- and high-dose groups. Furthermore, the procedural definition of criterion resulted in animals with very different learning patterns showing similar trials-to-criterion data. This was evident in several rats that were listed as “non-learners,” meaning that they did not meet criterion. However, rats that met criterion on the last day were also labeled “non-learners.” Comparing across subjects using other variables such as total errors, average errors per trial, or percent of trials with zero errors provided more meaningful information in terms of group differences.

In summary, scientifically-sound interpretations of cognitive data require that the reviewer be provided the right information. This includes full details regarding the experimental design, test parameters, specific procedures, physical equipment, monitoring system, and definitions of all dependent variables. For every endpoint, individual animal data should be provided for each trial, as well as summarized by group. Appropriate statistical analyses, while not described here in detail, are critical. Finally, historical or positive control data using the exact same procedure(s) may be informative.

7. Key issues influencing risk assessment: auditory startle response (ASR) and neuropathology endpoints as case studies (Abby A. Li, Robert H. Garman, Wolfgang Kaufmann, Roland N. Auer, and Brad Bolon)

The DNT test, as described in OECD 426 (OECD, 2007) and OPPTS 870.6300 (US EPA, 1998a), is designed to detect both structural (neuropathological) and functional (neurobehavioral) effects. For example, the DNT study results have been used to inform decisions on pesticides about safety levels for one or more risk assessment scenarios, with estimates that up to 40% of DNT results have had some impact prior to 2010 (Raffaele et al., 2010). The three following issues were selected for discussion, because in the opinion of the first author, AL, they have had an impact on risk assessment decisions regarding relative sensitivity and selection of points of departure for pesticides as well as other chemicals:

1. Maternal toxicity assessment in DNT studies
2. Selection of neurobehavioral response (effect) levels, when necessary
3. Quality and interpretation of morphometry data.

As a brief introduction and background, the DNT study has been used to inform decisions regarding the magnitude of the Pest Control Products Act (PCPA) factor in Canada or the Food Quality Protection Act (FQPA) factor in the United States for food-use pesticides. These two factors are based, in part, on evaluation of whether or not there is an increase in qualitative (e.g., subjective assessment as to level of concern for specific endpoint) or quantitative (e.g., larger magnitude of effects at the same dose level or effects obtained at lower dose level) sensitivity in offspring, compared to the adults. Other regulatory authorities responsible for pharmaceutical drugs, food additives, or industrial chemicals may make similar types of qualitative or quantitative comparisons to inform decisions about uncertainty factors or labeling requirements.

The design of the DNT study focuses on the evaluation of the offspring. This may limit detection of maternal toxicity and poor maternal care issues that can influence pup development, as will be discussed in greater detail below. In the European Union (EU) and California, if “developmental” effects are observed at doses that do not cause significant effects in dams, this can result in a labeling classification of “developmental toxicant” (European Chemical Agency ECHA, 2015; California OEHHA, 1993). These classifications can result in regulatory actions that are hazard-based and that can limit uses of the chemical even if there are adequate margins of safety. Thus, enhancing evaluations of the dams provides data that inform assessments of relative sensitivity. In addition, alterations in morphometric measures in offspring can be considered to be more severe developmental effects. This can lead to a conclusion that there is a qualitative increase in sensitivity in the offspring when these morphometric changes are observed at doses that produce maternal or adult toxicity that is considered reversible or less severe (e.g., body weight loss or transient decrease in activity). For this reason, the neuropathologist authors (RG, WK, RA and BB) have made recommendations intended to improve the quality and reliability of brain morphometry measures. These measures are described briefly below and in greater detail in Garman et al. (2016). In addition, the literature on the impact of maternal toxicity and stress

on brain development (structural and neurobehavioral consequences) should inform decisions about selection of the high dose level and provide additional perspective in evaluating the relative sensitivity based on DNT results.

Benchmark dose levels (BMDs) are routinely calculated (when appropriate) for different critical toxicity endpoints as part of the process of selecting the point of departure (US EPA, 2012). In addition, BMDs from adult and developmental endpoints have been compared for a quantitative assessment of relative sensitivity that informs decisions about the magnitude of uncertainty factors. The selection of the appropriate response level (BMR; considered minimally adverse) for continuous data has significant impact on risk assessment decisions that rely on BMD approaches.

When a BMD approach is not feasible, a NOAEL/LOAEL (no- and low-observed-adverse-effect level) approach is used in reference-value calculations (US EPA, 2012). A determination of what percent change from baseline is considered adverse is an important consideration when determining the NOAEL for an endpoint. Numerical differences from control that are large in magnitude, although not statistically significant, may be considered to be biologically-relevant differences from control. However, as described above by Dr. Sheets, the intrinsic variability can be high for some DNT behavioral endpoints, depending in part on the age of the animal, and this variability should be taken into account.

In summary, the scientific basis for risk assessment can be strengthened by improving the quality of maternal toxicity evaluations and morphometry measures on DNT studies and by making informed decisions about age-appropriate responses for different DNT endpoints. Each of these issues is discussed in greater detail below.

7.1. Maternal toxicity and dose selection

The DNT test guidelines are designed to focus on evaluating neurobehavioral and pathology endpoints related to the nervous system in the offspring, with minimal evaluation of an agent's effects on other general systemic toxicity endpoints in the progeny or any effects elicited in the dams. Thus, there is a higher probability of observing effects in the offspring than the dams, due to the experimental design of the DNT test. A key issue that influences risk assessment is that regulators may need to make judgments about qualitative or quantitative sensitivity of pups based on the data generated in the DNT study, which was not designed to address this issue.

The US EPA (1998a) DNT guideline specifies that the highest dose level "should induce some overt maternal toxicity, but should not result in a reduction in weight gain exceeding 20%," whereas OECD states that the highest dose should decrease body-weight gain, but not by > 10%. The high dose level is considered to be sufficiently high if it produces offspring toxicity, even in the absence of maternal effects. Typically, the lowest dose level is selected to produce no effect, to establish a clear NOAEL for maternal and offspring toxicity. The middle dose level may be the only dose suitable to evaluate the neurotoxic potential of chemicals without the confounding effects of maternal or offspring systemic toxicity or no effect.

Signs of overt maternal toxicity at the highest dose level make it difficult to distinguish direct effects on the developing nervous system from those related to maternal care or general toxicity. This distinction is less important when selecting a point of departure, because there should be protection against exposure to levels that cause these adverse effects, regardless of whether it is due to maternal care issues or a direct effect on the developing nervous system. However, the distinction becomes important if regulatory agencies make decisions regarding the relative sensitivity or severity of findings, or label the effect as a developmental neurotoxic effect. In the latter case, the decision to label a compound as a developmental neurotoxicant may be based solely on hazard instead of risk at an expected exposure level.

It is important to consider the crucial influence of the pregnant and lactating dam's health on the well-being and development of the nervous system in the offspring (DeSesso et al., 2009). Pups are dependent on dams for maternal care, milk quality and quantity, physical protection, and thermal homeostasis (DeSesso, 1987; Tyl et al., 2008). Although difficult to tease apart, high doses (e.g. doses that saturate exposure or maximally tolerated doses) may be of sufficient magnitude to produce effects that result in stress as a secondary effect (Everds et al., 2013). Prenatal stress and food deprivation have been linked to long-lasting abnormal cognitive, behavioral, psychosocial, and neuro-pathology outcomes in both animals and humans (Weinstock, 2008; Markham et al., 2010; Charil et al., 2010; Wahlsten and Bulman-Fleming, 1987; Barnes and Altman, 1973). These data suggest that alterations in brain measures and neurobehavioral endpoints in offspring, in conjunction with decreased dam and/or offspring body weights or maternal care issues, can be evidence that alterations could be related to a more general effect on progeny growth resulting from factors such as maternal stress and toxicity (Garman et al., 2016; Bolon et al., 2006; Charil et al., 2010; Weinstock, 2008; Markham et al., 2010). Although direct effects on the developing nervous system cannot be ruled out completely at doses that cause substantial maternal or pup toxicity, the potential contribution of maternal stress and general systemic toxicity to DNT endpoints should also be considered.

Three recommendations in the presentation were made to offset the DNT experimental design limitations and improve evaluation of findings within the proper context of maternal and offspring toxicity. The first is to consider alternative criteria for selection of high-dose toxicity, such as avoiding toxicokinetically-saturating dose levels (e.g., Saghir et al., 2013). Many toxicokinetic processes that influence absorption, distribution, elimination and metabolic activation or detoxication may become saturated at higher doses, resulting in systemic exposures to parent compound or metabolites that would not be expected to occur in the real life human exposures for which risk assessments are needed (OECD, 2012b). Second, measures of systemic toxicity and target-organ toxicity in dams and/or offspring should be added to the DNT or DNT range-finding study design, so that it is easier to relate the DNT study findings to other toxicity studies (e.g., adult neurotoxicity, standard adult toxicity studies, and developmental and reproductive toxicity [DART] studies). In addition, observations of maternal care issues within the DNT study should be formalized, so that they can be captured reliably and available to assist interpretation (Table 7).

For example, observations of maternal care should be made at consistent times and for a set duration while producing minimal disturbance to the dams and litter. The subjective maternal care observations should be made using a clearly-defined scoring system with operational definitions (Table 7), preferably by the same experienced technicians who have been trained to conduct detailed assessments such as those needed for a functional observational battery (FOB). The reasons for using technicians with formal training in conducting FOB assessments according to US EPA (1998b) and OECD (1997) adult neurotoxicity test guidelines are that they are both experienced at using a defined, semi-quantitative scoring system for observations, and they participate in periodic assessments of inter-observer reliability. Third, the DNT-related findings in pups need to be discussed within the study report in the context of maternal toxicity and any evidence of systemic toxicity in pups (e.g., body-weight changes), using a weight-of-evidence approach (Tyl et al., 2008; Garman et al., 2016). Direct and indirect effects on the developing nervous system can be difficult to distinguish; however, the level of concern regarding increased sensitivity of the offspring may decrease if the findings are present only when there is evidence of maternal toxicity and/or poor maternal care. Mild to moderate levels of maternal stress and litter size differences can result in changes in brain measures (reviewed by Charil et al., 2010; Weinstock, 2008; Everds et al., 2013). These studies suggest that morphometric measures of regional brain size (e.g., Bolon et al., 2006; Garman et al., 2016), while considered to be adverse effects, should be interpreted by taking into consideration the animal's size (e.g., consider both absolute brain weight and brain weight relative to body weight) and any evidence of maternal toxicity.

7.2. Determination of effect levels for benchmark dose (BMD) analysis

Data from DNT guideline studies should be analyzed based on the primary hypothesis that is being tested, as described by LS and KR. For example, more emphasis should be given to the overall effect of the test chemical on the ontogenic pattern across PNDs 13, 17, and 21, or on the intersession habituation data for PNDs 21 and 60, than on the highly-variable motor activity data for PND 13 alone. For the ASR habituation test, dose-related changes in the pattern of habituation are of greater importance than focusing on just the total session group mean. As described above by Dr. Raffaele, the most appropriate method of evaluating intersession habituation data is a mixed model RM-ANOVA approach (Holson et al., 2008).

As described above for the US EPA, some regulatory agencies may conduct BMD analyses. In such cases, the effect level or benchmark response level (BMR) should take into account the intrinsic variability of data based on well-conducted DNT studies, as described by Dr. Sheets. A default difference of 5% (EFSA, 2009, 2010) or recommendation of 20%–25% based on adult neurobehavioral studies (e.g., Crofton et al., 1991, an inter-laboratory study in adult rats) would be indiscernible from natural variation for many DNT behavioral endpoints. For example, the average CV for total session peak amplitude is 40% and 63% in control pups and adults (20 rats/dose level), respectively, from Dr. Sheets' DNT studies (Tables 5, 6). Similarly, but based on much smaller scale DNT-like studies (6–12 rats/dose level) from US EPA laboratories, the mean CV for total session peak amplitude was 30% and 51% for control pups and adults, respectively (Supplemental Table 1 of Sheets et al., 2016). These ASR results were generated by scientists who have specialized expertise in

ASR. Therefore, these data were considered by AL to be high-quality data that can serve as guides for determining an appropriate BMR. These results appear to be within the general range of contract lab data reported by Raffaele et al. (2008), but the data are not directly comparable because the CVs are reported for each intersession trial block.

In the absence of a biologically-based or accepted determination of a minimally-adverse change (the preferred standard for selecting a recommended BMR), US EPA (2012) may use a BMR of one SD for continuous data, which will account for intrinsic variability. In general, there have not been consensus scientific definitions of what minimal level of change in a continuous endpoint is biologically significant; thus, a BMR of 1 SD typically has been used. However, poorly-conducted studies might have very large SDs. The use of one SD would be appropriate if comparable or less than the CVs from well-conducted DNT guideline studies. These CVs reflect the inherent variability of behavioral data for animals of different ages and sex and not only are relevant for decisions about BMD analysis, but also should be considered when evaluating the biological relevance of numerical differences from concurrent controls for purposes of setting the NOAEL.

7.3. Quality and interpretation of morphometric data

Alterations in linear morphometric measurements were among the most common findings at the LOAEL in DNT studies submitted to the US EPA OPP prior to 2010 (Raffaele et al., 2010; 22 of 69 [32%] DNT studies). Structural (qualitative neuropathology and/or quantitative morphometry) changes are frequently considered to be severe and irreversible findings that clearly indicate the presence of neurotoxicity. These findings can impact the PCPA or FQPA safety factors or modifying uncertainty factors if there is concern regarding the possibility of increased qualitative or quantitative sensitivity. Thus, the quality of these morphometry measures and an appropriate understanding of how to interpret changes in these measures relative to other findings in a DNT study are of paramount importance.

Recommendations presented here for optimal brain processing and analytical conditions when conducting and analyzing DNT studies have been published in a separate detailed publication (Garman et al., 2016). Importantly, this publication also provided a rational basis for interpreting the DNT brain pathology and morphometric findings in the context of all the results from a DNT study, including offspring endpoints such as body weight, brain weight, and the pattern of findings across brain regions, doses, sexes, and ages, but also any evidence of maternal toxicity. The basic recommendations for the neuropathologic and morphometric analyses of the nervous system in DNT studies are summarized as follows (discussed in detail in Garman et al., 2016):

1. Neural tissues from animals in all dose groups ideally should be processed to block as soon as possible after fixation has been completed.
2. Every step of brain processing should be counterbalanced across dose group and sex.
3. Slicing and sectioning of brains should be anchored on reproducible neuroanatomical landmarks (external and internal) that are defined in specific detail in institutional standard operating procedures (SOPs) and/or reports, in

order to ensure that sections from all animals are highly homologous (i.e., essentially identical).

4. Production of a serial section to be stained with Luxol fast blue (LFB) to show myelinated tracts is highly recommended for quantitative morphometric measures, because it more clearly defines the boundaries of neuroanatomical regions that are to be measured quantitatively.
5. Certain morphometric measurements can be obtained more reliably from all animals, while others should be avoided.
6. Interpretation of neuropathological and morphometric data should take into consideration the consistency of the findings, as well as any evidence of maternal toxicity. This approach can include comparisons of absolute morphometric values with relative values (e.g., based on brain weight or volume) in evaluating whether or not alterations might be related to a more general effect on growth (see Bolon et al., 2006, and Garman et al., 2016, for more complete discussion).
7. Pathologists and neurobehavioral toxicologists should work closely together in interpreting DNT data sets to ensure appropriate integration of the behavior and pathology/morphometry findings, when possible.
8. In general, the neuropathologists who served as coauthors for the current presentation agreed that effects in linear brain measures indicating a selective effect on brain development (i.e., developmental neurotoxicity) should be seen in both juveniles (PND 11 and 22) and young adults (PND 70), while altered linear measures only seen in juveniles (i.e., without other neurobehavioral and neuropathological findings at PND 70), and in conjunction with decreased dam and/or pup body weights and/or maternal care issues, suggest that the juvenile brain changes might be indirectly related to generalized adverse effects on growth and development, rather than a specific neurotoxic effect (discussed further in Garman et al., 2016).
9. A decision tree and checklist for evaluating neuropathology and morphometric data facilitates a weight-of-evidence approach that can help define the risk posed by a test article. This approach is detailed in Garman et al. (2016).

In summary, collective experience has shown that DNT risk assessment can be greatly improved through the (a) systematic evaluation of maternal toxicity, (b) utilization of data from DNT-related studies to inform selection of a BMR (or similar index of risk), and (c) standardized approach to the processing and analysis of nervous tissues.

8. Discussion

At the end of both workshops, audience participation was encouraged by the discussion panelists: Francis Bailey and Angela Hofstra (both at SOT and Teratology), and Susan Makris (Teratology/NBTS [DNTS]) and Alan Hoberman (co-chair at Teratology/NBTS [DNTS]). These panelists, together with the speakers, represented government, industry, and contract research organization perspectives on regulatory DNT studies. The discussion

focused on improving methods and interpretation of neurodevelopmental evaluations. Although a worthy topic for future discussions, the workshop discussions did not consider how existing test guidelines might be changed to better support consistent DNT risk assessments. This section of the paper integrates major points of the presentations, with highlights of the discussions at both workshops and emphasis on some areas of agreement and disagreement. Further discussions stimulated by the workshops continued after the meetings and during the preparation of this manuscript.

Working within the context of the DNT and EOGRT guidelines, recommendations for improving neuropathology methods, with emphasis on linear morphometry measurements of the brain, is necessary and is available in a recent publication by the neuropathology co-authors (Garman et al., 2016). The detailed description of how to process brain tissue and what procedures to avoid contributed greatly toward understanding how to improve linear morphometric analysis. In particular, considerable care must be taken to generate highly-homologous sections (using reliable external and internal neuroanatomic landmarks) and reproducibly position anchor points for lines of morphometric measurement (which is accomplished most readily in Luxol fast blue [LFB]-stained sections). The optimal method to achieve consistency is to bring all CNS tissues from all dose groups to block at the same time post-fixation and counterbalance all phases of processing across dose groups and sexes. Consistent processing and sectioning techniques that provide the necessary section homology are available and need to be disseminated. There was general consensus among the speakers and panelists that these recommendations better ensure that the quantitative structural assessment is standardized among animals within and, ideally, across studies.

There was general agreement that DNT behavioral endpoints can be variable, even after controlling for extrinsic sources of variability within the constraints of DNT or EOGRT guidelines. During the discussion, VM pointed out that one way to minimize this factor is to use gestational age (counting from the day of mating) rather than counting from the day of birth (Palmer, 1981; Palmer, 1973). Physical and functional developmental landmarks are more closely correlated with gestational age (Ulbrich and Palmer, 1996; Tachibana et al., 1998), which may decrease variability due to assigning age based on the day of birth.

The workshop shed new light on how the variability of data may influence which statistical analysis should be applied. Specifically, statistical analysis methods for PND 13 motor activity data may need to be different from those used at older ages if the distribution of the data is substantially different (e.g., high number of zero responses at PND 13). There was consensus that data should be inspected carefully to determine which statistical methods are appropriate and further work is needed to understand whether or not PND 13 data should be combined with data from older animals (e.g., PND 17 and 21) in a single global RM-ANOVA. If PND 13 data should be analyzed separately, there were differences in opinion on the path forward. The biological hypothesis tested by the US EPA and OECD DNT guidelines for pre-weaning motor activity was to evaluate the effect of the test chemical on the ontogenic pattern of ambulatory motor activity from PND 13 to 21. One recommended statistical approach is to use RM-ANOVA to analyze the data by including PND 13, 17, 21 in the same analysis (Ruppert et al., 1985; Holson et al., 2008). However, if ambulatory activity at PND 13 is a fundamentally-different type of motor behavior from that measured

at PND 17 and 21, then direct comparisons of motor activity measures between PND 13 and 17 may not be informative (see Dr. Bowers' presentation). Moreover, it is unclear whether expending advanced statistical resources (e.g., statistical mixture models) to analyze the PND 13 data separately from PNDs 17 and 21 will increase sensitivity to detecting chemical effects on preweaning behavioral ontogeny. It is possible that motor activity data at the older ages, especially PNDs 21 and 60, may be more useful or more sensitive to assess developmental effects, although they would provide little information on behavioral ontogeny, per se. In fact, these were the two ages tested for motor activity in the collaborative behavioral study that is considered to support validation of the DNT guidelines (Buelke-Sam et al., 1985; Makris et al., 2008). Another possibility to explore is to reduce the session time from typical length of 1 h for adult animals to 30 min for PND 13, 17 and 21. Available data can be examined to determine if this might reduce data variability that may occur during the latter part of the session, when some pups are observed to sleep while others become more active. Staying within the test guidelines, the OECD DNT guideline recommends a test of behavioral ontogeny and gives several examples other than PNDs 13, 17, and 21 motor activity. Thus, for studies conducted for any OECD member country, including the US, one option might be to use another accepted approach for measuring behavioral ontogeny prior to weaning.

A common approach is to evaluate data in the context of the behavioral changes that are considered biologically relevant, independent of statistical significance. There was agreement that the magnitude of change considered to be biologically significant will depend on the behavior measured, the age of the animal, and the experimental design (e.g., adult studies vs DNT design). For example, an effect level of 25% compared to control might be considered biologically significant for motor activity in an adult study (e.g., based on Crofton et al., 1991). Data discussed from Dr. Sheets' lab might suggest a higher threshold for motor activity (e.g., at PNDs 13–21) and auditory startle for the DNT study (Tables 3, 5, and 6); however, it would be important to consider a range of CVs (as well as whether use of CVs is appropriate for certain types of data) from the literature based on guideline DNT experimental design (e.g., those analyzed in supplemental tables from Raffaele et al., 2008 and Sheets et al., 2016) to achieve consensus on these and other endpoints. In doing so, it is important to recognize that CVs for total (cumulative) session data are not directly comparable to CVs for each intrasession interval or trial block data.

There was consensus that appropriate application of RM-ANOVA approaches should be used to evaluate ASR and motor activity habituation data after weaning. In general, the results can be interpreted more easily if the number of factors is limited to three (e.g., treatment group, age, sex or treatment, intersession time, sex). RM-ANOVA approaches retain information about the effect of the chemical on habituation, which may be lost in a BMD analysis applied to total session data.

The US FDA and US EPA do not specify required tests or methods of data analysis. Recently, US FDA requested learning and memory tests that involve maze learning (e.g., Biel, Morris, Cincinnati Water mazes). A measure of learning and memory is not included in the OECD 443 EOGRTS, because this guideline is a more general evaluation of chemical effects on reproduction and development, rather than a focused evaluation of the developing

nervous system. All speakers agreed that the “Methods” and “Results” sections of reports should provide details usually included in SOPs, to help reviewers understand exactly what measures were taken to ensure that environmental conditions and the potential for experimental bias were well-controlled. There was also agreement that cognitive data should be summarized and presented in a way that shows a process of learning in control animals with training (i.e., not just performance on the last session). There was not sufficient time to discuss statistical analysis of learning and memory data, which includes repeated measures and also censored data. In other words, animals may not reach the escape platform in a maze within the trial time limit set for testing the animals, so that the latencies to escape cannot exceed this time limit, resulting in censored data.

There was general agreement that chemicals can have effects on the dam, as well as direct effects on the development of the offspring, such that it is sometimes difficult to distinguish whether an effect expressed in the offspring is a direct effect or secondary to maternal toxicity. Systematic evaluation of maternal toxicity and maternal care issues that allow correlation between individual data for offspring and the mother can be helpful, especially when hazard-based decisions on labeling (e.g., reproductive or developmental toxicant) are applied to a chemical by a regulatory authority. There was general consensus that (1) neurobehavioral or morphologic changes in a DNT study are of greater concern for qualitative or quantitative relative sensitivity when there is no detected maternal toxicity, and that (2) risk assessments should be protective of toxicity regardless of whether or not the effects are direct or indirect effects on offspring neural development.

9. Conclusion

Overall, the workshops at the 2015 SOT and Teratology/NBTS [DNTS] Society meetings met the goals of (a) articulating the primary biological hypotheses that are being tested in studies conducted by the DNT guidelines, (b) elucidating which dependent variables and potential interactions are of primary interest in addressing the priority biological questions, (c) improving understanding of “inherent” variability under conditions of guideline experimental design requirements, (d) identifying practical equipment considerations that affect the ability to test for the different biological hypotheses, and (e) illustrating approaches to evaluating DNT data that take into consideration inherent variability, historical control data, maternal toxicity, and appropriate statistical analyses.

The current overview publication resulting from the discussions undertaken in these workshops should provide new and useful approaches for data collection and analysis. The community-wide exchanges fostered by such workshops should also provide a rational consideration of existing regulatory guidelines around the globe that will improve the experimental design and conduct of DNT studies, as well as the interpretation and reporting of study data. Ultimately, the greater standardization and more consistent understanding of DNT data should enhance human risk and hazard assessment worldwide.

Funding source

Funding in support of the symposium held at the Teratology/Neurobehavioral Teratology Society annual meetings (June 2015) was provided by the U.S. Environmental Protection Agency, National Center for Environmental

Assessment, order no. EP-15-H000260. Dr. Abby Li is grateful to Crop Life America (CLA) and Crop Life Canada (CLC) for partially supporting her efforts as a participant at the workshops and co-author of this paper. CLA and CLC were not involved in reviewing or providing comments for either her presentation or any drafts of this paper.

References

- Altman J, Sudarshan K, 1975 Postnatal development of locomotion in the laboratory rat. *Anim. Behav* 23, 896–920. [PubMed: 1200422]
- Barnes D, Altman J, 1973 Effects of two levels of gestational-lactational undernutrition on the postweaning growth of the rat cerebellum. *Exp. Neurol* 38 (3), 420–428. [PubMed: 4735214]
- Björling-Poulsen M, Andersen HR, Grandjean P, 2008 Potential developmental neurotoxicity of pesticides used in Europe. *Environ. Health* 7, 50 10.1186/1476-069X-7-50. [PubMed: 18945337]
- 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. [PubMed: 16647766]
- Bolles RC, Woods PJ, 1987 The ontogeny of behaviour in the albino rat, limb of the neonatal rat. *J. Physiol. Lond* 383, 79–92. [PubMed: 3656143]
- Bolon B, Garman R, Jensen K, Krinke G, Stuart B, Ad Hoc Working Group of the STP Scientific and Regulatory Policy Committee, 2006. A “best practices” approach to neuropathologic assessment in developmental neurotoxicity testing—for today. *Toxicol. Pathol* 34, 296–313 (Erratum 34:697). [PubMed: 16698729]
- Bolon B, Garman RH, Gundersen HJ, Allan Johnson G, Kaufmann W, Krinke G, Little PB, Makris SL, Mellon RD, Sulik KK, Jensen K, 2011 Continuing education course #3: current practices and future trends in neuropathology assessment for developmental neurotoxicity testing. *Toxicol. Pathol* 39 (1), 289–293. [PubMed: 21075916]
- Boos DD, Brownie C, 1991 Mixture models for continuous data in dose-response studies when some animals are unaffected by treatment. *Biometrics* 47 (4), 1489–1504. [PubMed: 1786327]
- Buelke-Sam J, Kimmel CA, Adams J, Nelson CJ, Vorhees CV, Wright DC, St Omer V, Korol BA, Butcher RE, Geyer MA, et al., 1985 Collaborative behavioral teratology study: results. *Neurobehav. Toxicol. Teratol* 7, 591–624. [PubMed: 3835454]
- California OEHHA (Office of Environmental Health Hazard Assessment), 1993 Criteria for recommending chemicals for listing as “known to the state to cause reproductive toxicity”. http://oehha.ca.gov/prop65/policy_procedure/pdf_zip/dartCriteriaNov1993.pdf (web archive link, 10 March 2016) (Last accessed on 10 March 2016).
- Charil A, Laplante DP, Vaillancourt C, King S, 2010. Prenatal stress and brain development. *Brain Res. Rev* 65 (1), 56–79. 10.1016/j.brainresrev.2010.06.002. (Epub 2010 Jun 13). [PubMed: 20550950]
- Clarac F, Brocard F, Vinay L, 2004 The maturation of locomotor networks In: Shigemori DGS (Ed.), *Progress in Brain Research, Brain Mechanisms for the Integration of Posture and Movement*. Elsevier, pp. 57–66.
- Clark RE, Zola SM, Squire LR, 2000 Impaired recognition memory in rats after damage to the hippocampus. *J. Neurosci* 20, 8853–8860. [PubMed: 11102494]
- Crofton KM, Howard JL, Moser VC, Gill MW, Reiter LW, Tilson HA, MacPhail RC, 1991 Interlaboratory comparison of motor activity experiments: implications for neurotoxicological assessments. *Neurotoxicol. Teratol* 13 (6), 599–609. [PubMed: 1779947]
- Crofton KM, Peele DB, Stanton ME, 1993 Developmental neurotoxicity following neonatal exposure to 3,3'-iminodipropionitrile in the rat. *Neurotoxicol. Teratol* 15, 117–129. [PubMed: 8510606]
- Crofton KM, Foss JA, Hass U, Jensen KF, Levin ED, Parker SP, 2008 Undertaking positive control studies as part of developmental neurotoxicity testing. *Neurotoxicol. Teratol* 30, 266–287. [PubMed: 17681747]
- Davis M, 1980 Neurochemical modulation of sensory-motor reactivity: acoustic and tactile startle reflexes. *Neurosci. Biobehav. Rev* 4, 241–263. [PubMed: 6106916]
- Davis M, Eaton RC, 1984 The mammalian startle response In: *Neural Mechanism of Startle Behavior*. Plenum Press, New York, pp. 287–351.

- DeSesso JM, 1987 Maternal factors in developmental toxicity. *Teratog. Carcinog. Mutagen* 7 (3), 225–240. [PubMed: 2888203]
- DeSesso JM, Watson RE, Keen CL, Hazelden KP, Haws LC, Li AA, 2009. Analysis and integration of developmental neurotoxicity and ancillary data into risk assessment: a case study of dimethoate. *J. Toxicol. Environ. Health A* 72 (2), 94–109. [PubMed: 19034799]
- ECHA (European Chemical Agency), 2015 Guidance on the Application of the CLP (Classification, Labeling and Packaging) Criteria Version 4.1. https://echa.europa.eu/documents/10162/13562/clp_en.pdf/58b5dc6d-ac2a-4910-9702-e9e1f5051cc5.
- EFSA (European Food Safety Authority), 2009 Guidance of the Scientific Committee on a request from EFSA on the use of the benchmark dose approach in risk assessment. European Food Safety Authority (EFSA), Parma, Italy. *EFSA J.* 1150, 1–72.
- EFSA (European Food Safety Authority), 2010 Use of BMDS and PROAST software packages by EFSA Scientific Panels and Units for applying the benchmark dose (BMD) approach in risk assessment. EN-113. EFSA, Parma, Italy.
- 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. [PubMed: 16414243]
- Eichenbaum H, 2000 A cortical-hippocampal system for declarative memory. *Nat. Rev. Neurosci* 1, 41–50. [PubMed: 11252767]
- Everds NE, Snyder PW, Bailey KL, Bolon B, Creasy DM, Foley GL, Rosol TJ, Sellers T, 2013 Interpreting stress responses during routine toxicity studies: a review of the biology, impact, and assessment. *Toxicol. Pathol* 41 (4), 560–614. [PubMed: 23475558]
- Garman RH, Fix AS, Jortner BS, Jensen KF, Hardisty JF, Claudio L, Ferenc S, 2001 Methods to identify and characterize developmental neurotoxicity for human health risk assessment. II: Neuropathology. *Environ. Health Perspect* 109 (Suppl. 1), 93–100. [PubMed: 11250809]
- Garman RH, Li AA, Kaufmann W, Auer RN, Bolon B, 2016 Recommended methods for brain processing and quantitative analysis in rodent developmental neurotoxicity studies. *Toxicol. Pathol* 44 (1), 14–42. [PubMed: 26296631]
- Gold PE, 2003 Acetylcholine modulation of neural systems involved in learning and memory. *Neurobiol. Learn. Mem* 80, 194–210.
- Grandjean P, Landrigan PJ, 2014 Neurobehavioural effects of developmental toxicity. *Lancet Neurol.* 13, 330–338. [PubMed: 24556010]
- He H, Tang W, Wang W, Crist-Christoph P, 2014 Structural zeroes and zero-inflated models. *Shanghai Arch. Psychiatr* 26 (4), 236–242.
- Health Canada, 2008 Science Policy Note: The Application of Uncertainty Factors and the Pest Control Products Act Factor in the Human Health Risk Assessment of Pesticides. http://www.hc-sc.gc.ca/cps-spc/pubs/pest/_pol-guide/spn2008-01/index-eng.php#pcpa.
- 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 (4), 326–348. [PubMed: 17681748]
- Hulse SH, Fowler H, Honig WK, 1978 *Cognitive Processes in Animal Behavior*. John Wiley & Sons, New York.
- ICH (International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use), 2005 ICH Harmonised Tripartite Guideline, Detection of Toxicity to Reproduction for Medicinal Products & Toxicity to Male Fertility S5-R2. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Safety/S5/Step4/S5_R2_Guideline.pdf.
- 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.
- Kaufmann W, 2011 Pathology methods in nonclinical neurotoxicity studies: the developing nervous system In: Bolon B, Butt M (Eds.), *Fundamental Neuropathology for Pathologists and Toxicologists: Principles and Techniques*. John Wiley & Sons, Hoboken, NJ, pp. 339–363.

- Kaufmann W, Gröters S, 2006 Developmental neuropathology in DNT-studies—a sensitive tool for the detection and characterization of developmental neurotoxicants. *Reprod. Toxicol* 22 (2), 196–213. [PubMed: 16781841]
- Li AA, 2005 Regulatory developmental neurotoxicology testing: data evaluation for risk assessment purposes. *Environ. Toxicol. Pharmacol* 19, 727–733.
- Makris S, Raffaele K, Allen S, Bowers WJ, Hass U, Alleva E, Calamandrei G, Sheets L, Amcoff P, Delrue N, Crofton KM, 2008 A retrospective performance assessment of the developmental neurotoxicity study in support of OECD test guideline 426. *Environ. Health Perspect.* 117 (1), 17–25. [PubMed: 19165382]
- Marable BR, Maurissen JP, 2004 Validation of an auditory startle response system using chemicals or parametric modulation as positive controls. *Neurotoxicol. Teratol* 26, 231–237. [PubMed: 15019956]
- Markham JA, Taylor AR, Taylor SB, Bell DB, Koenig JI, 2010 Characterization of the cognitive impairments induced by prenatal exposure to stress in the rat. *Front. Behav. Neurosci* 4, 173. [PubMed: 21151368]
- Maurissen J, 2010 Practical considerations on the design, execution and analysis of developmental neurotoxicity studies to be published in neurotoxicology and teratology. *Neurotoxicol. Teratol* 32 (2), 121–123. 10.1016/j.ntt.2009.09.002. (Epub 2009 Sep 13). [PubMed: 19755149]
- Maurissen JP, Hoberman AM, Garman RH, Hanley TR Jr., 2000 Lack of selective developmental neurotoxicity in rat pups from dams treated by gavage with chlorpyrifos. *Toxicol. Sci* 57, 250–263. [PubMed: 11006355]
- Morris R, 1984 Developments of a water-maze procedure for studying spatial learning in the rat. *J. Neurosci. Methods* 11, 47–60. [PubMed: 6471907]
- OECD (Organisation for Economic Co-operation and Development), 1997 Test No. 424: neurotoxicity study in rodents. Web site: http://www.oecd-ilibrary.org/environment/test-no-424-neurotoxicity-study-in-rodents_9789264071025-en (web archive link, 10 March 2016) (Last accessed on 10 March 2016).
- OECD (Organisation for Economic Co-operation and Development), 2007 Test No. 426: Developmental neurotoxicity study. 2007 Web site: http://www.oecd-ilibrary.org/environment/test-no-426-developmental-neurotoxicity-study_9789264067394-en (web archive link, 06 March 2016) (Last accessed on 6 March 2016).
- OECD (Organisation for Economic Co-operation and Development), 2008 Guidance Document 43 on Mammalian Reproductive Toxicity Testing and Assessment. ENV/JM/MONO(2008)16. Web site: [http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono\(2008\)16&doclanguage=en](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2008)16&doclanguage=en).
- OECD (Organisation for Economic Co-operation and Development), 2012a Test no. 443: extended one-generation reproductive toxicity study. Web site: http://www.oecd-ilibrary.org/environment/test-no-443-extended-one-generation-reproductivetoxicity-study_9789264185371-en;jsessionid=2flknsxwv043m.x-oecd-live-02 (web archive link, 10 March 2016) (Last accessed on 10 March 2016).
- OECD (Organisation for Economic Co-operation and Development), 2012b Guidance Document 116 on the Conduct and Design of Chronic Toxicity and Carcinogenicity Studies, Supporting Test Guidelines 451, 452 and 453, 2nd edition. .
- Palmer AK, 1973 Some thoughts on reproductive studies for safety evaluation. In: *Proceedings of the European Society for the Study of Drug Toxicity*. vol. 14 pp. 79–90.
- Palmer AK, 1981 Regulatory requirement for reproductive toxicology: theory and practice. In: Kimmel CA, Buelke-Sam J (Eds.), *Developmental Toxicology* Raven Press, New York, pp. 274.
- Perrot-Sinal TS, Kostenuik MA, Ossenkopp KP, 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. [PubMed: 8986334]
- Raffaele KC, Sette W, Makris SL, Moser V, Crofton KM, 2003 Motor activity in developmental neurotoxicity testing: a cross-laboratory comparison of control data, *Society of Toxicology, Salt Lake City, UT Toxicologist* 72 (S1), 123.

- 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. [PubMed: 18280700]
- Raffaele KC, Rowland J, May B, Makris SL, Schumacher K, Scarano LJ, 2010 The use of developmental neurotoxicity data in pesticide risk assessments. *Neurotoxicol. Teratol* 32, 563–572. [PubMed: 20398750]
- Roof RL, Stein DG, 1999 Gender differences in Morris water maze performance depend on task parameters. *Physiol. Behav* 68, 81–86. [PubMed: 10627065]
- Ruppert PH, Dean KF, Reiter LW, 1983 Developmental and behavioral toxicity following acute postnatal exposure of rat pups to trimethyltin. *Neurobehav. Toxicol. Teratol* 5, 421–429. [PubMed: 6646316]
- Ruppert PH, Dean KF, Reiter LW, 1985 Development of locomotor activity of rat pups in figure-eight mazes. *Dev. Psychobiol* 18, 247–260. [PubMed: 3987971]
- Saghir SA, Marty MS, Zablony CL, Passage JK, Perala AW, Neal BH, Hammond L, Bus JS, 2013 Life-stage-, sex-, and dose-dependent dietary toxicokinetics and relationship to toxicity of 2,4-dichlorophenoxyacetic acid (2,4-D) in rats: implications for toxicity test dose selection, design, and interpretation. *Toxicol. Sci* 136, 294–307. [PubMed: 24105888]
- Sheets LP, Li AA, Minnema DJ, Collier RH, Creek MR, Peffer RC, 2016 A critical review of neonicotinoid insecticides for developmental neurotoxicity. *Crit. Rev. Toxicol* 46 (2), 153–190. [PubMed: 26513508]
- Slikker W Jr., 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, Sobotka TJ, 2005 Behavioral test methods workshop: a summary. *Neurotoxicol. Teratol* 27, 417–427. [PubMed: 15939202]
- Squire LR, 2004 Memory systems of the brain: a brief history and current perspective. *Neurobiol. Learn. Mem* 82, 171–177. [PubMed: 15464402]
- Stump DG, Beck MJ, Radovsky A, Garman RH, Freshwater LL, Sheets LP, Marty MS, Waechter JM Jr., Dimond SS, Van Miller JP, Shiotsuka RN, Beyer D, Chappelle AH, Hentges SG, 2010 Developmental neurotoxicity study of dietary bisphenol A in Sprague-Dawley rats. *Toxicol. Sci* 115 (1), 167–182. 10.1093/toxsci/kfq025. [PubMed: 20164145]
- Tachibana T, Narita H, Ogawa T, Tanimura T, 1998 Using postnatal age to determine test dates leads to misinterpretation when treatments alter gestation length, results from a collaborative behavioral teratology study in Japan. *Neurotoxicol. Teratol* 20, 449–457. [PubMed: 9697971]
- Tilson HA, Mitchell CL, 1984 Neurobehavioral techniques to assess the effects of chemicals on the nervous system. *Annu. Rev. Pharmacol. Toxicol* 24, 425–450. [PubMed: 6375549]
- Tyl RW, Crofton K, Moretto A, Moser V, 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 (4), 349–381. [PubMed: 17826946]
- Ulbrich B, Palmer AK, 1996 Neurobehavioral aspects of developmental toxicity testing. *Environ. Health Perspect.* 104 (2), 407–412. [PubMed: 9182049]
- US EPA (U.S. Environmental Protection Agency), 1998a Office of Chemical Safety and Pollution Prevention (OCSPP) Harmonized Test Guidelines, Group E—870.6300 Developmental Neurotoxicity Test Guidelines. <http://www.regulations.gov/#!documentDetail;D=EPA-HQ-OPPT-2009-0156-0042> (web archive link, 10 March 2016) (Last accessed on 10 March 2016).
- US EPA (U.S. Environmental Protection Agency), 1998b Office of Chemical Safety and Pollution Prevention (OCSPP) Harmonized Test Guidelines, Group E—870.6200 Neurotoxicity Screening Battery. <http://www.regulations.gov/#!documentDetail;D=EPA-HQ-OPPT-2009-0156-0041> (web archive link, 10 March 2016) (Last accessed on 10 March 2016).
- US EPA (U.S. Environmental Protection Agency), 2012 Benchmark Dose Technical Guidance EPA/100/R-12/001. Risk Assessment Forum, Washington, DC <https://www.epa.gov/risk/benchmark-dose-technical-guidance> (web archive link, 10 March 2016) (Last accessed on 10 March 2016).

- US EPA (U.S. Environmental Protection Agency), 2015 Consideration of the FQPA Safety Factor and Other Uncertainty Factors in Cumulative Risk Assessment of Chemicals Sharing a Common Mechanism of Toxicity Office of Pesticide Programs, Washington, DC <http://www.epa.gov/sites/production/files/2015-07/documents/apps-10x-sffor-cra.pdf> (web archive link, 10 March 2016) (Last accessed on 10 March 2016).
- US FDA (U.S. Food and Drug Administration), 2006 Guidance for industry. Nonclinical safety evaluation of pediatric drug products. <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM079247.pdf>.
- Vinay L, Brocard F, Pflieger J, Simeoni-Alias J, Clarac F, 2000 Perinatal development of lumbar motoneurons and their inputs in the rat. *Brain Res. Bull* 53, 635–647. [PubMed: 11165799]
- Vinay L, Brocard F, Clarac F, Norreel JC, Pearlstein E, Pflieger JF, 2002 Development of posture and locomotion: an interplay of endogenously generated activities and neurotrophic actions by descending pathways. *Brain Res. Rev* 40, 118–129. [PubMed: 12589911]
- Vorhees CV, Williams MT, 2006 Morris water maze: procedures for assessing spatial and related forms of learning and memory. *Nat. Protoc* 1, 848–858. [PubMed: 17406317]
- Vorhees CV, Williams MT, 2014 Assessing spatial learning and memory in rodents. *ILAR J.* 55, 310–332. [PubMed: 25225309]
- 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. [PubMed: 18502078]
- Wahlsten D, Bulman-Fleming B, 1987 The magnitudes of litter size and sex effects on brain growth of BALB/c mice. *Growth* 51 (2), 240–248. [PubMed: 3678936]
- Weinstock M, 2008 The long-term behavioural consequences of prenatal stress. *Neurosci. Biobehav. Rev* 32 (6), 1073–1086. [PubMed: 18423592]

Fig. 1.

Diagram of EPA Developmental Neurotoxicity Study—OCSPP 870.6300 (adapted from Raffaele et al., 2010). Detailed clinical observations and body weights are performed periodically in the F1 generation throughout the study, before weaning and after treatment is discontinued, until around postnatal day (PND) 70 (PND 0 is defined as the day of delivery). Neurobehavioral tests for the F1 generation include automated measures of activity, auditory startle habituation, and measures of cognition (US EPA, 1998a). The numbers in the table including the total time needed to conduct motor activity in the last row are from an actual study from Dr. Sheets' lab. The last row in bold is the approximate time it takes to run all the motor activity sessions. As illustrated in this example, some labs add extra pups distributed across groups to meet guideline requirements for 20 offspring/dose group (1 pup/sex/litter from 20 litters/dose group) in case of mortalities, logistical errors and equipment malfunction. The US EPA and OECD DNT study design requirements result in a particularly high workload at ~PND 21 (US EPA, 1998a; OECD, 2007). In cases where the 80–100 females are mated over a period of 3–5 days (the duration of the estrous cycle in rats), this may result in litters born over a similar period. Using this study design, having a range of delivery days helps to reduce certain aspects of the workload at specific days of age. On the other hand, this factor increases the logistical complexity to manage the daily workload, because the various study endpoint assessments overlap. PND 21 is a particularly busy day, with the need to measure body weight, obtain detailed clinical observations, and evaluate motor activity, auditory startle, and cognition for F1 animals from 80 to 100 litters. One consequence of the test guideline requirements is that animals must be tested over a

long test day and across different days, increasing variability due to circadian rhythms and other environmental factors that can vary despite best attempts to control them.

Fig. 2.
Representative intrasession motor activity (mean $\pm$ SEM) for control male Wistar rats at PNDs 13, 17, 21, and 60 from a guideline DNT study from Dr. Sheet's lab.

Fig. 3.

Theoretical data (mean) illustrating test session motor activity data patterns that reflect normal, as well as altered, habituation. Normal habituation in adult rats is shown in Vehicle Control and Treatment A. The specific temporal activity pattern in Control animals for a study will be determined by a range of factors, including the test equipment, test conditions, age, species and strain, etc. Regardless, habituation in control animals will be reflected by a decrease in activity measures over the course of the test session. While Treatment A shows decreased ambulatory activity relative to Control animals, Treatment A animals exhibit habituation rates comparable to Control (note parallel lines). In contrast, Treatment D 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. Treatment B

and C animals show reduced habituation, as indicated by the reduced rate of ambulation decrease in the second half of the test session. Note that Treatment C and Control will have comparable total session ambulation counts despite session activity patterns that are not comparable, illustrating the potential for total session data to provide misleading information. In the absence of time block data, it would not be possible to determine that Treatment C alters habituation. Similarly, while total session activity is comparable between Control and Treatment C and D, using only total session activity would mask the fact that Treatment C and D exhibit opposite alterations in habituation pattern. Note that Treatment A through D refer to different arbitrary treatments to illustrate issues to consider.

Fig. 4.

Frequency histograms for PND 13 (top), PND 17 and PND 21 (bottom) motor activity data of control animals from the laboratory of Dr. Moser. Data were collected in the figure-8 chamber and include data for males and females combined. X-axis values are photocell counts. Each chart within each age contains data for 10-min time blocks for a total of 30 min of data. The X-axis indicates the number of photocell counts, and the Y-axis indicates the frequency of counts (number of animals). For PND 13 animals, there is a high proportion of animals with little or no detected movement (frequency of 0 counts), especially after the first time block. The data set is highly skewed at PND 13. In contrast, at PND 17 and 21, there are few animals with no movement, and data are normally distributed (n 's = 43).

Fig. 5.

Frequency histograms for PND 13 (top, $n = 6$), PND 17 (middle, $n = 20$) and PND 21 (bottom, $n = 20$) motor activity data from Charles River Laboratories. Data were collected using Coulbourn Instruments Tru Scan system, and values are ambulatory counts from combined male and female control animals. Each chart within each age contains data for 10-min time blocks for a total of 30 min of data. The X-axis indicates the ambulatory counts and the Y-axis indicates the frequency of counts (number of animals). For PND 13 animals, there is a high proportion of animals with no movement (frequency of 0 counts), and the data set is highly skewed. Unlike the figure-8 chamber data, PND 17 activity data are also highly skewed, with a high proportion of animals with little movement.

Fig. 6.

Frequency histograms for PND 17 (top, n = 104) and 35 (bottom, n = 80) motor activity data of control animals (males and females combined) from Dr. Bowers' laboratory. Data were collected using Med Associates activity system, and values are ambulatory counts from control animals from multiple studies. Each chart within each age contains data for 10-min time blocks, for a total of 30 min of data. The X-axis indicates the number of ambulatory counts, and the Y-axis indicates the frequency of counts (number of animals). For PND 17 animals, the data distribution is skewed at 10 min but approaches a normal distribution at later time points. At PND 35, there is a low frequency of 0 counts, and the data set is normally distributed at most time blocks. The increase in proportion of animals exhibiting little movement (increase in proportion of 0 counts) at 30 min reflects the onset of the expected reduction in activity associated with habituation at this age.

Fig. 7.
Latency data (mean $\pm$ SEM) for chemical X over eight days of training, four trials per day. Each day is the average of four trials. Inset shows individual trial data from day 1 only (circled).

Fig. 8.

Trials-to-criterion data for chemical X. Average data expressed as mean $\pm$ SEM. Individual data range from 4 (minimum possible) to 12 (maximum number of trials).

Table 1

Contributors to the SOT and TS/NBTS [DNSTS] workshops (in order of presentation).

Name (abbreviation)	Areas of expertise	Affiliation at time abstracts to SOT were submitted
Angela Hofstra (AH)	Regulatory Toxicology	Syngenta Canada
Francis Bailey (FB)	Regulatory Toxicology	Health Canada Pest Management Regulatory Agency
Larry Sheets (LS)	Neurotoxicology, Regulatory Toxicology	Bayer CropScience
Wayne Bowers ^a (WB)	Neurotoxicology	Health Canada (currently at Dept. Neuroscience, Carleton University)
Virginia Moser (VM)	Neurotoxicology	US EPA NHEERL ORD
Kathleen Raffaele (KR)	Neurotoxicology, Risk Assessment and Management	US EPA OSWER
Thomas Vidmar (TV)	Statistics	BioSTAT Consultants
Edmund Lau (EL)	Statistics	Exponent Health Sciences
Abby Li ^a (AL)	Neurotoxicology, Risk Assessment	Exponent Health Sciences
Robert Garman (RG)	Neuropathology	Consultants in Veterinary Pathology, Inc.
Brad Bolon (BB)	Neuropathology and Developmental Neurobiology	The Ohio State University, College of Veterinary Medicine (currently at GEMPath, Inc.)
Wolfgang Kaufmann (WK)	Neuropathology and Reproductive Toxicology	Merck KGaA, Germany
Roland Auer (RA)	Neuropathology and Medicine	Hopital Ste-Justine, Departement de Pathologie, Montreal, Quebec (currently at University of Saskatchewan, Canada)
Alan Hoberman ^a (AMH)	Developmental, Reproductive and Juvenile Toxicology (DART)	Charles River Laboratory
Susan Makris (SM)	Developmental and Reproductive Toxicology, Risk Assessment	US EPA NCEA ORD

^a Chairpersons for Society of Toxicology or Teratology Association Meetings.

Table 2

EPA and OECD guidelines for developmental neurotoxicity studies (DNT) and extended one-generation reproductive toxicity studies (EOGRTS).

	EPA 870.6300 (1998a)	OECD 426 (2007)	OECD 443 (2012a)
Test species	Rat	Rat	Rat
Exposure parent/F1	GD6 to PND11 (EPA OPP currently specifies exposure to weaning)	GD6 to weaning	2 weeks pre-mating to weaning/adult
Motor activity	Pre-weaning ontogeny (PND 13, 17, 21) and 60 ($\pm$ 2)	Pre-weaning and adult	Adult
Behavioral ontogeny	None (other than motor activity)	2 measures ("e.g., PND 13, 17, 21" motor activity)	None
Clinical observations, body weight	Throughout	Throughout	Adult
Motor and sensory function	Auditory startle habituation specified; weaning and ~day 60	PND 25 ($\pm$ 2) and adult	Auditory startle habituation specified; weaning
Sexual maturation	As appropriate	As appropriate	As appropriate
Learning and memory	Weaning and ~day 60	PND 25 ($\pm$ 2) and (PND 60–70)	None
Brain weight, neuropathology, brain morphometric analysis	Weaning ^a (brain only) and adult (brain, spinal cord and peripheral nerves)	PND 22 ^b (brain only) and adult (brain, spinal cord and peripheral nerves)	Weaning (brain only) and adult (brain, spinal cord and peripheral nerves)

^aEPA guidelines state PND 11 but EPA accepts OECD guidance.

^bEPA guidelines state PND 11 but EPA accepts PND 21.

Table 3

Control figure-8 chamber motor activity from the Bayer laboratory (mean $\pm$ SD; coefficient of variation [CV]) from nine DNT studies using Wistar male rats (20/group).

Study number	PND 13	PND 17	PND 21	PND 60				
1	79 ± 109	138%	263 ± 153	58%	341 ± 92	27%	573 ± 128	22%
2	65 ± 62	95%	257 ± 128	50%	282 ± 117	41%	434 ± 113	26%
3	80 ± 76	95%	129 ± 80	62%	270 ± 84	31%	586 ± 149	25%
4	64 ± 65	102%	240 ± 160	67%	334 ± 89	27%	531 ± 200	38%
5	90 ± 81	90%	169 ± 103	61%	344 ± 134	39%	536 ± 91	17%
6	77 ± 76	99%	253 ± 143	57%	343 ± 180	52%	572 ± 142	25%
7	49 ± 37	76%	217 ± 115	53%	307 ± 117	38%	521 ± 104	20%
S	46 ± 61	133%	181 ± 129	71%	314 ± 126	40%	472 ± 98	21%
9	49 ± 44	90%	127 ± 85	67%	331 ± 77	23%	529 ± 157	30%

Table 4

Control data (10 DNT studies) from the Bayer laboratory: Interession Figure-8 Maze Motor Activity for PND 60 Wistar Rat Offspring (N = 20 male rats/group).

Interval 1	Interval 2	Interval 3	Interval 4	Interval 5	Interval 6
Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD
116 ± 32	105 ± 36	103 ± 35	94 ± 31	80 ± 23	76 ± 29
98 ± 19	72 ± 21	71 ± 32	76 ± 27	68 ± 31	50 ± 27
116 ± 35	96 ± 28	111 ± 40	98 ± 26	87 ± 22	78 ± 31
119 ± 47	84 ± 34	89 ± 38	91 ± 43	81 ± 37	67 ± 35
107 ± 20	92 ± 23	85 ± 22	92 ± 32	89 ± 23	71 ± 32
106 ± 27	94 ± 27	102 ± 38	105 ± 40	86 ± 30	79 ± 25
97 ± 15	93 ± 30	91 ± 31	85 ± 27	83 ± 32	73 ± 31
103 ± 17	84 ± 22	77 ± 24	72 ± 24	68 ± 31	68 ± 24
99 ± 24	75 ± 26	92 ± 51	89 ± 33	88 ± 37	86 ± 35
101 ± 17	88 ± 25	89 ± 22	79 ± 29	70 ± 37	76 ± 41

Table 5

Control data (9 DNT studies) from the Bayer Lab: Intersession auditory startle response (ASR) data (grams) for PND 22 Wistar Male Rats (N = 20 male pups/group).

Trials 1–10		Trials 11–20		Trials 21–30		Trials 31–40		Trials 41–50		Summary session mean ± SD (CV)
Mean peak ± S.D		Mean peak ± S.D		Mean peak ± S.D		Mean peak ± S.D		Mean peak ± S.D		
40 ± 20		39 ± 21		34 ± 17		32 ± 15		31 ± 18		35 ± 17 (49%)
43 ± 16		36 ± 16		31 ± 12		30 ± 14		28 ± 12		34 ± 13 (38%)
45 ± 17		44 ± 21		41 ± 20		39 ± 22		36 ± 15		41 ± 18 (44%)
54 ± 22		53 ± 20		49 ± 26		41 ± 23		38 ± 25		47 ± 22 (47%)
53 ± 18		52 ± 17		52 ± 22		46 ± 20		41 ± 15		49 ± 16 (33%)
42 ± 10		42 ± 16		42 ± 17		35 ± 14		35 ± 15		39 ± 13 (33%)
48 ± 21		44 ± 17		44 ± 21		44 ± 21		42 ± 24		45 ± 19 (42%)
53 ± 21		54 ± 17		54 ± 15		52 ± 14		50 ± 20		53 ± 16 (30%)
50 ± 23		44 ± 23		38 ± 19		36 ± 17		33 ± 14		40 ± 17 (43%)

Table 6

Control data (9 DNT studies) from the Bayer Lab: Intersession auditory startle response (ASR) data (grams) for PND 60 Wistar Male Rats (N = 20 male pups/group).

Trials 1–10		Trials 11–20		Trials 21–30		Trials 31–40		Trials 41–50		Summary session mean ± SD (CV)
Mean peak ± S.D		Mean peak ± S.D		Mean peak ± S.D		Mean peak ± S.D		Mean peak ± S.D		
278 ± 117		233 ± 165		215 ± 128		169 ± 94		141 ± 58		207 ± 102 (49%)
303 ± 205		254 ± 196		213 ± 168		177 ± 138		146 ± 88		218 ± 152 (70%)
299 ± 238		231 ± 164		188 ± 146		175 ± 127		149 ± 142		208 ± 156 (75%)
383 ± 167		278 ± 160		197 ± 123		179 ± 94		122 ± 46		232 ± 100 (43%)
249 ± 195		220 ± 173		180 ± 124		162 ± 115		120 ± 74		186 ± 128 (69%)
171 ± 94		155 ± 148		124 ± 137		116 ± 103		105 ± 89		134 ± 111 (83%)
255 ± 215		194 ± 159		162 ± 119		149 ± 105		113 ± 61		175 ± 119 (68%)
303 ± 157		299 ± 135		234 ± 128		148 ± 91		164 ± 120		230 ± 115 (50%)
239 ± 138		251 ± 194		155 ± 113		143 ± 88		111 ± 69		180 ± 112 (62%)

Table 7

Example of explicit, operationally defined scales and scoring criteria for assessing maternal behaviors.

Amount of litter scatter	Location of dam relative to litter
1. Litter was not scattered; all pups were in the nest	1. Dam was on the pups
2. Litter was scattered; 1–3 pups were outside the nest	2. Dam was away from the nest but actively tending to litter (e.g., retrieving pups or building the nest)
3. Litter was scattered; 4 + pups were outside the nest	3. Dam was away from the nest but actively eating, drinking, or grooming
	4. Dam was away from the nest but not actively eating, drinking, or grooming, and not tending to the litter (i.e., sleeping, resting on food jar)