Overall Charge Questions

First, I wish to observe the many strengths of this document. Each module is structured similarly, with sections that include background information about the tests covered under that module, statistical issues, examples of data (especially helpful), and other topics that will be of value. Overall the document was thorough and should be of assistance in guiding the reviewers. The literature reviews struck a good balance between being thorough and concise. The writing was free of typos and the prose was generally quite clear and well organized. Another significant strength is that the guidelines specify that data on individual subjects be presented. This can be very helpful, especially when attempting to interpret aberrant values, in understanding variability and sometimes in guiding an understanding of the robustness of effects seen.

The choice of subsections was well thought out. The subsections that are consistent across all broader sections are the introduction, test description, guideline requirements (although this terminology was a bit confusing since guidelines are frequently not thought of as requirements, and requirements are not suggestions), a table containing EPA and OECD requirements, test procedures, data recording, dependent variables, representative data, statistical analyses, biological controls, and interpretation. Especially helpful for interpretation is the presence of examples of different types of datasets that might be encountered. These offer important background for reviewers and concrete items to look for in data that have been submitted. The comments below are offered in response to the questions provided as well as in an effort to identify possible ambiguities and areas that might be strengthened.

Does the document provide enough information on why and when the guidance should be used? If not, how could it be improved?

- The document doesn’t really specify when the guidance should be used, at least not in the introduction where I would expect them to appear. The guidance is offered and sets of requirements (OECD, OECD extended one-generation, and USEPA) are summarized but I did not see a clear statement as to when these guidelines apply and who requests that a particular set of guidelines be applied. This can be important because of the differing standards among these three approaches.

- I also did not see a clear statement as to why they should be used. Regarding this question, the “when” and “why” question seems to be implicit. They are applied when data pertinent to a chemical’s neurotoxicity are submitted for review.

- The material in the specific modules contains the implication that the guidelines would apply when the tests are submitted for approval but I did not see this made explicitly. There were some areas of ambiguity. For example, locomotor activity is described in Module 1 (FOB) and, of course, in Module 2 (Locomotor Activity). The guidelines in Module 1 are superficial while those in Module 2 are quite detailed. Module 2 requires automated procedures while Module 1 is agnostic as to whether observations or locomotor activities are necessary.

- Another issue about when and why to apply the guidelines is in Module 1. This module seems to give considerable latitude to the testing laboratories to select the tests to use and to what extent a FOB is required. It notes that testing laboratories “often have one protocol of clinical observations that is used regardless of the specific requirements (i.e., for adult and DNT studies).” This seems to accept a one-size-fits-all approach that applies not only to the tests selected for use but also, somewhat incredibly, to whether it is developmental or adult testing that is under review. It is not clear at what point the submission is inadequate. \

- One question that I had at a few points in the document was what its purpose was. Is this intended to ratchet up the standards and provided a generally improved and more consistent set of practices? That was my assumption going in and the discussion of best-practices and the presence of relatively high-quality data as examples supports this assumption. Alternatively, is it intended to capture the state-of-the-art as it is now? I raise this possibility because in places it is noted that suboptimal practices have been used in the past and seems to tacitly accept this practice.

What limitations, if any, do you find in the document that would hinder data review and interpretation of DNT studies conducted using the EPA or OECD DNT Guidelines?

- Outliers are an inevitable of behavioral testing. How will they be detected and how will they be treated? The section on startle testing has some good examples of how aberrant data points could arise, for example, by using the maximum peak rather than the first peak. This section also has a section discussing the importance of studying the raw data for unusual individual scores when the coefficient of variability is very high. All modules discuss the possibility of outliers and section E notes that they could be due to sensitive (or insensitive) subjects. However, there is little guidance about what to do. Suggestions might including investigating the data records for an aberrancy, checking the data to ensure that the numbers are physically possible, and running statistical analyses with and without the outlier to determine the sensitivity of the analysis to that data point.

- The statistical sections clearly describe the statistical approach but do not specify that F ratios and, especially, degrees of freedom should be reported. Having these numbers would be helpful from a quality-control perspective, to serve as a check on the appropriate conduct of statistical analyses.

- The treatment of positive controls is inconsistently address across modules. The first module, on FOB, is rather vague, saying that such data could be useful but that the positive control data that have been submitted are of little value. In a different section in Module A, it is noted that "though positive-control data should encompass all test ages . . . this practice has not been put into practice." What is the reviewer to do with this? It sounds like the guidance is that this is the way that tests have been conducted so it is the way that they will continue to be conducted. In contrast, other modules, B and especially C (startle) are explicit about the requirements of positive control data, their importance, and contains considerable detail about what is expected.

- It is stated that exposure should occur from GD0 through weaning but there is little guidance about how to determine that such exposure has actually occurred. The possibility that there may not be post-natal exposure is raised at one point and the reviewer is instructed to consider this, but in the absence of biomarkers of exposure how would the reviewer know about this issue?

- The treatment of motor deficits is minimal and mostly ancillary to other tests. Compounds that produce subtle disruption of gait, strength, coordination could be missed. Why not do quantitative gait assessment or rotarod?

- The auditory system is the only sensory system evaluated directly. The visual system is almost ignored and olfaction and somatosensory systems are completely ignored. Omitting especially the visual system seems surprising given that system's sensitivity to neurotoxicants and its general importance.

Module Specific Charge Questions

<u>Modules 1-4</u>

Does the document provide sufficient guidance to assist regulatory scientists in reviewing reports to determine whether critical details regarding procedure, study design, results (including summary and individual data for all relevant parameters), and statistical evaluation are included in the reports for studies conducted under the EPA or OECD DNT Guidelines? If not, why not?

- FOB:
  - This is a minor thing, but is the word "clinical observation" intended to refer to unstructured observations or to the FOB? Page A-3 described "expanded clinical observations" but then goes on to say that there are no guidelines or published protocols for this. Later, clinical observation seems to refer to FOB.

    Section 3 (page A-4) reviews the various test guidelines (under the title "Guideline *Requirements*,") but elsewhere it notes that testing labs have their own set of procedures and that the procedures are applied regardless of the chemical under review or the age of the subjects. This implies that an a la carte approach or even a one-size-fits-all approach is acceptable. How is a reviewer to reconcile the presence of "requirements" with the tacit acceptance of whatever tests the testing lab submits? This guidance is ambiguous.

  - Guidelines about positive control inclusion are ambiguous. "Theoretically" they should encompass all test ages but this has not been put into practice. So, some "positive control" is required but what is it? What strain? What age? What chemical? What dose? The guidance on positive controls in this section is less specific than in other sections.

  - Section 5.1 calls some observations "useless" like "changes in skin, fur, mucous membranes" or "unusual or abnormal behaviors." Is this intended to be a guideline that the reviewer should ignore these observations? Such a suggestion seems at odds with their use in veterinary settings where they are often indicators of illness, even if they are nonspecific.

  - Section 5.2. Positive Control Data. It sounds like the guidelines are suggesting eliminating their use, even though earlier it says that they are useful for showing the sensitivity of the technician. It says that they are of little value since they involve such high-dose effects. So, why not recommend positive control data that are contemporaneous with the testing (to validate the technician, for example), or make them more pertinent?

- ACOUSTIC STARTLE

  - It is said (Page C-12) that habituation can be detected statistically by a main effect of trial block but such a statistical result effect is inadequate, as noted later in section 7.1. It might be helpful to note instead that habituation can be detected statistically by a downward trend in the data across trials. A main effect of trial block could occur from monotonically increasing data or even "W"-shaped data in which there is no consistent trend. This is clarified, almost, in Section 7.1 where it says that a decrease in startle responses is required.

  - Page C-14. There is a long discussion about the influence of body weight on startle amplitude. Body weight is an influence because of what is measured, some variant of a downward force on a load cell device or accelerometer, which will be influenced by body weight since a heavy animal will exert greater downward force than a light one. The flinching by a large animal will produce a greater change on a load cell than will the flinching by a small animal. It is then said that body weight should not be used as a covariate if it is influenced by exposure. True, a covariate must not be affected by treatment but failing to control for body weight could result in false negatives. To refer to the top panel of figure 3, if a small animal's startle is of the same magnitude as a larger one's then it would seem as though the smaller one is more sensitive to the stimulus that provokes the startle-the relative startle is greater. It might even result in a false positive in the case of an obesogen that increases body weight, producing greater downward force on the measurement device but not necessarily greater startle.

- LEARNING AND MEMORY

  - This section is almost completely devoted to the Morris Water Maze, Passive Avoidance and the Biel/Cincinnati Maze. For each of these the discussion is impressively detailed and thorough but why do they receive such privileged attention at the expense of other tests? Why, for example, is the RAM not included despite its widespread use in neuroscience?. Why isn't active avoidance used in conjunction with passive avoidance, which is a variable measure, since it is more apical and in ways more difficult. Why the exclusive attention on procedures that emphasize escape and avoidance, which all three do?

  - Page D-38. It is said that ANOVAs require "continuous" data. That is too restrictive. Integers, can certainly be analyzed using ANOVAS. Examples would be the number errors, lever-presses, or correct responses. ANOVAs do, however, require a normal distribution, stability of variance across groups, and are supposed to be performed on ratio- or interval-based scales, but are frequently performed on ordinal scales, such as data from Likert type questionnaires. In the case of count data, if the numbers are skewed toward one or zero then distributional assumptions of normality may not obtain and a transform will be beneficial.

  - The document places transformations of the data on an equal footing with nonparametric tests but they are quite different on a number of dimensions, not least in statistical power. The labs should be encourage to transform the data first (and there are many alternatives to the square root transform) and the reviewer should be on the lookout for such transforms as they can greatly decrease the likelihood of error. Nonparametric tests, in contrast, are usually less sensitive and could yield false negatives. The advice to use Kaplan-Maier estimators and log-rank analyses is a good one and the example is appropriate but it should be noted, too, that with the sample sizes commonly used it can take a fairly large different to be statistically significant. This is evident in the example where the difference between low and high is that of 90% vs 30% of the sample meeting criterion.

- Excessive right-censoring at the maximum possible value would amplify the problem of sensitivity with "survival" analyses. Too much right censoring can be prevented by the appropriate choice of experimental parameters such as trial length; if the length is too short, for example, then there will be an excessive amount of right censoring. Labs are to be encouraged to make the trial length sufficiently long that relying on KM estimators won't be necessary.

Given that regulatory reviews are conducted independent of any review or interpretation presented by the study authors: does the document provide sufficient guidance to assist regulatory scientists in interpreting the data and results from regulatory studies conducted under the EPA or OECD DNT Guidelines? If not, why not?

- FOB
  - This modules notes that test labs typically have a single set of tests and apply this set to all chemicals regardless of the type of chemical or whether exposure is developmental or chronic. Thus, it provides some guidance as to what tests must be used but leaves much up to the testing lab. The implication is that this one-size-fits-all approach is acceptable. There is little guidance for a reviewer who has concerns that the tests selected by the testing lab may not be suitable for the chemical or exposure regimen being tested. There is little guidance as to how tests are to be selected, other than what the testing lab typically does.
- LOCOMOTOR ACTIVITY
  - Table 1 specifies that the age of testing +- 2 days, but that is a huge range for 13 day olds. The differences in activity for, say, a 11 and a 15 day-old pup could make interpretation difficult.
- ACOUSTIC STARTLE
  - It is said that the startle test should be done in a sound-attenuating chamber with background noise, but there is little guidance about the background noise, other than to say that it should be below the level of the stimuli to be used. This is a critical variable that can significantly influence the results. Some firmer guidance seems warranted.
  - Table 2 in the startle section is very helpful and it raises an important methodological point. The only appropriate measure of force is Newtons although grams is an acceptable proxy (strictly speaking, grams are a measure of mass; force is grams

multiplied by acceleration from gravity). The only appropriate unit of acceleration is “g,” the acceleration due to gravity, or meters/sec/sec. The use of a physically meaningful unit can be helpful from a quality-control perspective since they would help determine if the amplitude is reasonable.

It is said that the accelerometer outputs are in arbitrary units, and this raises some question about the reproducibility of the data, especially if more than one system is used. Many accelerometers provide data in units of acceleration, sometime in fraction of *g,* the acceleration due to gravity. A properly calibrated accelerometer should not require arbitrary units. The importance of this issue arises in 8.1 when reviewing properties of startle control data. Age-appropriate amplitude measurements cannot be properly assessed if the units are arbitrary. Units of “volts” are only marginally better than arbitrary units since they give the reviewer no way to anchor the magnitude of the dependent measure and determine whether the values obtained are reasonable.. Similarly, uits of time should be seconds or milliseconds, not “Tmax” or just “mean time,” which confer little information.

- LEARNING AND MEMORY

  - In the opening why does it say that learning and memory are theoretical “constructs,” which are hypothetical, unobservable entities, such as associations, used in theories to explain learning, but this is immaterial to this document. Learning, as noted in the next sentences is a change in behavior, something that is clearly measureable, and memory, as is the retention of that change, also readily measurable and observed. Yes, they are intertwined but they are clearly measurable and observable.

  - (Page D-4) The definition of operant conditioning is incorrect. Operant behavior is behavior that is sensitive to its consequences, not the association between behavior and some stimulus, and operant conditioning the process by which consequences influence behavior.

  - It is noted in section 2 that food deprivation is stressful but properly done it is not only not stressful but healthy. For developing animals, mild food deprivation like removing food for eight hours or so before the test procedure is quite acceptable. For adult animals, caloric restriction prevents the accumulation of abdominal fat. Not only is that healthy but fat accumulates many lipophilic neurotoxic substances, interfere with the toxicokinetics, and limit the generality of the tests. It might be noted that mild food deprivation stands in contrast to water-based tests in which an animal may be placed in room temperature water, which can cause core temperature to fall, and experiences a possibility of drowning, raising markers of stress. It is not my intent to say that water-mazes are inappropriate but rather that singling out mild food restriction as stressful is misleading. And to call missing a reward on a trial as severely stressful is perplexing in light of the enthusiasm in the document for shock-based passive avoidance and water-based mazes. For example, on page D-11 it is noted that the consequences of an incorrect trial in a food-based task are severe but in a water-based task it is relatively benign. First, it is not necessary that a trial go unrewarded—the animal could certainly be guided to the correct option--but this is rarely if ever done because missing the small food pellet is not particularly stressful so rodents rarely persist in an alley as in water-based mazes. Acquisition in the food-based procedures is because of the strengthening effects of the food reinforcement, not the averseness that follows from making a mistake. Acquisition may be faster in the water tasks because acquisition of escape/avoidance responding is usually faster, and the motivation in the water is to escape the water.

- Sometimes investigators provide a heating pad to the animal to help recover core temperature. This would be in a chamber with the pad over half of it, so the animal can thermoregulate behaviorally by positioning itself over the pad or over the section of the chamber without the pad.

- The document notes the use of a straight maze to estimate swim speed and certainly swim- or running speed should be considered in assessing performance in mazes. Most software programs do this calculation so if video tracking is used then having them swim through a straight alley won't be necessary.

- Page D-15. There is discussion on the incorporation of swim speed in the interpretation of data. First, this should be applied to any maze-based test, including land-based mazes. Speed can also be an issue in any trials-based task in which a time limit is imposed. The report notes that where there are effects on speed (which, itself, could be a point-of-departure for safety assessment) then an error-to-criterion or number of subjects succeeding will serve as dependent measures, but these alternatives do not exhaust the possibilities and, in fact, lead to a false negative. For example, if an exposed animal fails to finish trials because of a lower swim speed then an errors-to-criterion measure could actually be lower than controls simply because there were fewer opportunities to make errors. An "errors per opportunity" measure might address this issue. The number of animals reaching criterion could also contain the same confound if reaching criterion is driven both by speed and accuracy, and such a measure will have relatively low statistical power.

- Page D-15. "tests should include some measure of retention." Could this be more specific? Would a one minute delay be sufficient or is the intent to re-conduct the test after 24 hours or so? Later, on page D-19 it says that the probe trial should be 24 hours later. The sentence should be modified to say "some measure of retention, typically 24 hours later."

- Page D-16. The placement of distal cues should take the poor vision of rodents into consideration. Objects or pictures should be large and differences among them should not be based on color since rodents have only monochromatic vision.

- Figure 3 is helpful but the text is too small to read and the quality of the reproduction is poor so even when expanded the resolution of the text is too pixilated to read.

Does the document provide the correct summary of the kinds of information to look for in submitted data, provide relevant examples, and assist in interpretation of any treatment-related changes?

- FOB.

  Regarding the data that are reported, some mention should be made here and elsewhere that when there is an abnormally large SD that there should be a search for outliers. This can affect both type 1 and especially type 2 errors.'

  - The statistical analysis section is thorough. The recommendation that an interaction that main effects can be evaluated only in the context of interactions that might be present is on correct and frequently overlooked.

  - It is said that CVs can be as high as 100%. If the CV is 100% then clearly the data are not normally distributed around the mean and a transform should be considered. Such a large CV would imply that the left end of the distribution is truncated at 0 at -1 SD.

- ACOUSTIC STARTLE

  - A relatively minor point, but the axes in Figure 3 are not quite right. The Y axis label should be BW or ASR, % control since both BW and ASR are shown in the figures.

  - Section 7.1 (Statistical methods) describes main effects first and then interactions. Since elsewhere it notes (correctly) that main effects should be tested directly only if there are no interactions, it would be appropriate to describe interactions first in this section. I can see where that might interfere with clarity in writing so perhaps a comment could be repeated in the statistical methods section of the sections about the importance of interpreting main effects in light of interactions.

  - The section on statistical vs biological significance in the startle section was especially thoughtful. Clearly, as noted in this discussion, if the variability is so high that the detection of an effect would be unlikely then the raw data should be examined or that portion of the experiment should be redone.

  - The presence of representative datasets in the startle section is much appreciated, as it is where it is presented elsewhere. The two sets showing data that are difficult to interpret are helpful, as in other sections, but both examples are from false-positive cases. To assist with the review process it would seem that a false negative example showing cases where a closer scrutiny of the data revealed a chemical effect would also provide useful guidance.

- LEARNING AND MEMORY

  - For many trials procedures, simply reporting errors is inadequate. A comprehensive assessment requires knowing about both errors of commission (selecting the incorrect alternative) and of omission (failing to complete a trial).

  - The treatment of “nonlearners” in section 2.3 on page D-13 is thoughtful and the three examples provided are good examples of different types of nonlearners. The number of non-learners should be reported. Since this is a dichotomous measure its statistical power will be limited but nonetheless if the number is dose-related, or appears at the highest level of exposure then this should be taken into account in interpreting the test. This issue is related to the problem with trials to criterion in general, and that problem is the arbitrariness of defining a criterion, as alluded to in this discussion. A criterion that is too high will not be reached by any subject but one that is too low will be reached by all subjects. Neither will discriminate toxic from nontoxic doses. An alternative approach is curve-fitting, as is routine in determining habituation to startle stimuli. Learning would be expressed as a monotonically increasing curve and the magnitude of learning would be the upper asymptote.

- On page D-14 it is noted that if data are not normally distributed then medians or modes should be used. These are not the best options for a number of reasons, but one is that the median is difficult to change and median tests are relatively insensitive. The mode can be highly unstable. A better option is to transform the data. Log transform, inverse transform, or, if percent-correct measures are used, trigonometric transforms can often stabilize the variance or normalize the data. Frequently a ratio of corrects to incorrects can be used which, when log-transformed, often has excellent statistical properties. If there is a zero in the data then a recommendation is to add 0.1 to both the numerator and the denominator (or one tenth of the lowest possible value).

- For position discrimination studies it should be reported whether a correction procedure is used. This is important because chance performance may be 67% in a correction procedure if an animals adopts a win-stay-lose-shift strategy.

- The left panel of Figure 4 illustrates very nicely a learning deficit on the MWM but I have some questions about the right panel. It is said to come from a published report (Moser et al., 2001, Toxicol. Sci, 60, 315) but this figure does not match the published figures in that paper. The doses are different (6 to 60 mg/kg/dy in the DNT document and 0.03 to 3 mg/kg/day in Moser et al., 2001), the pattern of effects is different (no effects in Moser, 2001) and the results are different (no effects of heptachlor in Moser, 2001). Also, I wonder about the interpretation of the figure in the guidelines document. I thought it was going to be an example of a motor effect in the absence of a learning effect. I raise this question because the high-dose group was consistently slower than the other groups but the slope of the learning curve appears to be indistinguishable from the other groups. The interpretation of this figure, however, holds that there is an effect of learning “supported by ANOVA,” as said in the caption. However, the caption goes on to say that ther was no significant interaction with day, indicating that the treated rats were slower across all days. Isn’t this a motor deficit? This apparent error suggests that this section should be carefully reviewed to ensure that figures from the published literature are included correctly.

- The incorporation of reversal learning and working memory tasks into the MWM testing is well-advised since these procedures tap very different behavioral functions and neural processes, especially reversal learning tasks.

- Page D-22. The last two bullets say that “Ideally” stable performance and control procedures should be included. These aren’t ideals, they are necessary.

- Page D-26. In the section on passive avoidance the document appropriately specifies that a 1 mA shock should be delivered. It might also be noted that specifying shock in units of voltage is inappropriate. The shock is delivered by current (charge/unit time) and the amount of current actually delivered depends not only on voltage but also resistance, which can vary with type of skin, cleanliness of the metal floor, and moisture. Thus, the same voltage can deliver varying amounts of current depending on environmental conditions.

- Page D. 37. Standard procedures for all learning procedures.
  - The fourth bullet notes the importance of counterbalancing if several mazes are used. To avoid possible misinterpretation it might be noticed that the counterbalancing should occur across groups, when repeated testing is performed the subject should always experience the same maze.
  - The specification of water temperature is vague. What temperature range should be used?
  - When data are collected by individual observation and there is any subjective component then some measure of interobserver reliability should be provided.
  - Page D-37 specifies that control data should be within a reasonable range. Some specification of what is reasonable could be helpful. A coefficient of variation less than 20? 15?

<u>WOE Module</u>

Is this weight-of-evidence chapter consistent with the presentations from the rest of the document? Does it present a logical approach to integrating data from different behavioral endpoints to make scientifically justified conclusions?

This section is an excellent review of consistent themes that run throughout the different modules, and in that it provides useful summary of points to consider in a review. It is generally consistent with the other modules, except where the modules have different perspectives on issues, such as with the role of positive controls. Additional sections include the list of principles of neurotoxicological effects and the list is quite helpful. One thing that I would add is the tacit assumption that if no effect is detected then it is assumed that the chemical under investigation does not present a hazard at the exposure levels and using the exposure regimen in the set of tests, or at least that moving forward with a commercialization of the chemical is acceptable. Another additional section is the guidance on interpretation and overall data synthesis, and these sections, too, are thoughtfully and concisely constructed.

Some of the issues raised above could be raised again in this module. For example, in section six it draws a distinction between rating scales and “continuous” data, implying that only continuous data can be used in simple data analyses. The issue of continuous data was addressed earlier.

Section six on alternative DNT testing seems out of place here, and it is difficult to know what a reviewer would do with this section. It is said that experts should be consulted, and that is certainly true, if not understated. If a high-throughput, in vitro assay shows toxicity that would lead to a reviewer’s deciding not to approve the chemical then there is an error on the side of safety, even if the chemical may be acceptable, but it is difficult to see a commercial enterprise’s motivation for submitting such information for review. If, on the other hand, such an assay finds no toxicity at all then would a reviewer conclude that there would be no sensory damage, no motor damage, no effects on learning on memory, no effects on locomotor activity, or no effects on startle? The field of in-vitro testing is not sufficiently advanced to ensure confidence in such a conclusion.

Overall, the bullets in this section contain a helpful checklist of considerations that the reviewer should take into account. It summarizes accurately the key points from the modules and presents it in a format that will be helpful.