EPA's Interagency DNT Study Review Guidance

Charge Questions

Background Information and Goals:

The developing nervous system is known to be especially vulnerable to many environmental contaminants (Grandjean and Landrigan 2006; NRC 1993; Rodier 1995; Spyker 1975), and exposures may result in altered neural development at lower doses or with consequences that may be quite unlike the chemical's effects in an adult nervous system (Grandjean and Landrigan 2006; NRC 1993; Rodier 1995; Spyker 1975). For these reasons, regulatory agencies (OECD 2007; U.S.EPA 1998a) have promulgated testing guidelines for developmental neurotoxicity (DNT). DNT refers to any adverse effect of exposure to a toxic substance on the normal development of nervous system structures and/or functions (U.S.EPA 1998b). The basic purpose of DNT guideline testing is to act as an initial assessment and screen for the potential of chemicals to cause adverse neurodevelopmental outcomes.

The full history of the development, validation, acceptance and use of DNT testing has been reviewed previously (Makris et al. 2009; Raffaele et al. 2010; Tsuji and Crofton 2012). Briefly, the design and test specifics of the US EPA test guidelines were developed at a workshop held in 1989, following which the specific guideline was developed and eventually finalized in 1998 (U.S.EPA 1998a). The OECD updated this guideline (OECD 2007) to include enhancements developed through discussion and international agreement. More recently, OECD included a limited number of DNT endpoints in the Extended One Generation Reproductive Toxicity Study Guideline (OECD 2011). A number of papers have compared these guidelines (Hass 2006; Ladics et al. 2005; Makris et al. 2009; Makris and Vorhees 2015; Piersma et al. 2012; Tsuji and Crofton 2012).

As with all guideline-based testing, data interpretation is first done by the submitting company/organization as part of the final study report submission. In addition to data summaries and interpretation by the study authors, regulatory submissions include detailed procedural information and all study data, including both summary and individual animal data for all measured parameters. Upon receipt of the study report, regulatory agencies conduct their own review of the summary and individual data. *Important to note is that regulatory reviews are conducted independent of any review or interpretation of the data presented by the study authors.* Interpretation of results by Agency reviewers may, or may not, agree with the study submitter's conclusions. Over the past two decades a number of reports have been written to provide assistance in the interpretation of the data resulting from DNT studies (Cory-Slechta et al. 2001; Elsner et al. 1986; Francis et al. 1990; Holson et al. 2008; Li 2005; Makris et al. 2009; Slikker et al. 2005; Tilson and Wright 1985; Tyl et al. 2008; U.S.EPA 1998b; Vorhees and Makris 2015). Recent interactions between international regulatory agencies have highlighted a need for procedures to support consistent interpretation of the results from DNT for use in risk decisions. The interpretation of the behavioral data is the most inconsistent between agencies, and brought into question why different agencies were deriving different interpretations from the same datasets. As a result of these international concerns, Health Canada and the US EPA developed guidance on the review and interpretation of submitted DNT data. **Thus, the focus of this document is to provide guidance on how to evaluate the quality, the conduct, and resulting data derived from the behavioral methods employed in the OECD and EPA DNT Guidelines.**

This guidance provides information for regulatory agency scientists who perform internal reviews of the behavioral test data that result from the use of the EPA and/or OECD DNT Guidelines studies, especially those who may not be experts in neurotoxicity or developmental neurotoxicity. The guidance was generated by an international collaboration between Health Canada and the US EPA. The overall goal of the guidance is to foster better and more consistent consensus-based reviews of DNT behavioral data between these two countries. This guidance may also be useful for other international regulatory agencies.

Notes:

1) This review is restricted to evaluation of the guidance provided on the interpretation of submitted behavioral data from studies conducted under Good Laboratory Practices (GLP) by sponsoring companies or contract laboratories. *This document is in no way intended to review the test methods recommended in the Guidelines, or to suggest alternative methods*.
2) This document is divided into separate modules for each of the specific behavioral tests included in the test guidelines (motor activity, acoustic startle, learning and memory, and functional observations). In order to be most useful for the regulatory reviewer, the document focuses on and describes only those methods that are most often used by industry in submitted regulatory Guideline studies. This document does not include all the potential experimental approaches to assess these behaviors.

Charge Questions:

In your review of this document, please provide written responses to the following questions. Additional comments and recommendations for improving this document are also welcome.

Overall Charge Questions

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

RESPONSE: Yes, it provides appropriate rationale for why and when the guidance should be consulted but the details of the guidance is not entirely accurate (see below).

- 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?

RESPONSE: Much of the guidance is good and appropriate, but there are areas where it is incomplete, inaccurate, or needs clarification. There are also views expressed that are not consistent current knowledge and are in need of adjustment. I have listed the points that need attention below. Let me preface my comments by saying that I assume that Modules 1-4 referred here in the Charge Questions document are the same ones called Modules A-D in the guidance document. I further assume that Module WOE here refers to Module E in the guidance document. My comments are organized according to these assumptions.

Module Specific Charge Questions

Modules 1-4

- 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?

RESPONSE: No. The current document has shortcomings which I itemize below.

Module A – Observations

Page A-8: bullet point "Reactivity to handling, placing" It is important to note that these outcomes are sensitive to handling and handling is not similar across labs or time. When the Irwin and FOB were developed, animals were housed in single, wire-bottom, barren cages; these housing conditions came to be called 'isolation' housing in the scientific literature. In the intervening years, animal welfare concerns have driven changes in veterinary care practices such that today rats are housed in pairs or groups; they are housed in solid-bottom cages with bedding; and are required to have some type of at least minimal enrichment within each cage (and what form this takes varies widely across labs). Moreover, most vivaria today are ultra-clean, so-called barrier facilities whereas when the observational batteries were developed animals were housed in conventional housing. There are published studies showing that animals housed in barrier facilities are not the same as those housed conventionally in response to drugs, chemicals or infectious agents. These changes have occurred gradually over the last 20-25 years, but their net impact is profound. Rats housed as they are today are not stressed in the way they were in the past. This affects observations such as reactivity to handling, placing, urination, defecation and other measures in observational batteries making it unclear whether these measures have the same meaning as they once did, or if they have any meaning at all. When a measure depends on the animal being in a stressful environment and environments are changed to reduce stress, then the foundation of the test is gone. While within an experiment, treated animals will always be compared to contemporaneously prepared controls, it is important to realize that these handling/housing outcomes are different than in the past and their validity is very much open to question. I know of no studies that have attempted to revalidate these housing/handling sensitive measures. As an example, the frequency of urination and defecation have declined over several decades to the point that the baseline rate today in most labs is low. Any measurement that occurs at low frequencies raises concerns about reliability. Therefore, outcomes for reactivity to handling, placing, urination, defecation and others (e.g., posture) should be given less weight than measures less affected by housing/handling factors or these should be recommend to be dropped from observational batteries altogether since they are suspect.

Same page: Bullet point "Activity and/or rearing in an open field or test arena" since DNTS studies require an automated assessment of motor activity the value of a short observational and subjective scoring of movement and rearing should be deleted from the battery or at least discounted in comparison to the data from automated systems. In fact, the value of activity by brief observation is questionable and it is redundant to assess motor activity observationally when we know that valid assessments require 40-60 minutes and cannot be obtained in a few seconds by observer rating; the sample of behavior is too short to be meaningful.

Bullet point: "Urination, defecation" these two outcomes are questionable and should be stated as such in guidance to evaluators. Not only are these intrinsically suspect, they were developed more than

50 years ago when rats were housed in what has since been described as “isolated” or “deprivation” conditions. Naïve, unhandled, barren, wire bottom cage, single-housed rats are known to exhibit significant urination and defecation rates when suddenly thrust into an open arena under bright light. But in modern vivaria, the basis for the use of these measures no longer exists (see above). Scientifically, these two outcomes should be dropped and if a reviewer receives such data, he/she should not place weight on effects reported on these indices.

Page A-9: Bullet point “Changes in righting reflex – except during development” I would be appropriate to remind reviewers that for developmental reflexes the distinction between a delay and a lasting change is important. Also, it is prudent to remember that since righting is assessed only once/per day and the reflex develops rapidly, a few hours delay in development can show up as a full-day delay because the test-retest interval is too long. One needs to be cautious in interpreting righting delays because not only can daily testing exaggerate (or even miss) differences, transient delays with catch-up may or may not have any biological significance. For example, pediatricians routinely plot growth and milestone development for children. However growth, crawling, standing, walking, running, first words, etc. show wide inter-individual differences in normal healthy children. Any pediatrician will tell parents not to worry about small lags and that there is nothing to worry about unless they are not present at a point beyond where 95% of children have met these landmarks. We tend to forget in the formal world of rodent testing that the same principles apply to all of mammalian development. This should be factored into the interpretation of minor delays in reflex development if the test is done over enough days that catch up can be shown. In general, a one-day delay (which could be only a few hours) should not be regarded as a serious manifestation of toxicity.

Page A-9: Section 5.2 “Positive control data”: If reviewers are reviewing a study from a laboratory that has not submitted high quality positive control data, the results of the study under review should be viewed with skepticism. Absence of good positive control data is a red flag with regard to the skills and knowledge of the submitting laboratory and suggests the lab lacks good practices. Personally, if I were a reviewer of a study coming from a lab with poor positive control data I’d reject the study (assuming that reviewers have the discretion to do so).

Page A-9: Section 7 “Interpretation” Greater emphasis in interpreting data should be placed on quantitative data (motor activity, ASR, L&M) over observational data. Even within observational batteries, quantitative measures such as grip strength measured with strain gauges and core body temperature measured with rectal or other appropriate thermometers should be given more weight than subjective measures of grossly observed behaviors rated on scales such as “normal, mild, moderate, or severe.” Such scales are often unreliable and the underlying data difficult to assess in terms of validity. Moreover, some observations lack nervous system foundations. For example, handling reactivity has no known nervous system substrate, therefore, its meaning is unknown. While a number of observations in these batteries involve the nervous system to some degree, there is little evidence that placing, touch, clicker-induced ear twitch, etc. are reliable measures of peripheral or central nervous system dysfunction. While some possess face validity (pupil reflex), they have never been shown to have construct or predictive validity and should be viewed with caution.

Page A-10: Par-2: When dose administration is by a bolus method (e.g., gavage) it is important that reviewers interpret data differently if the behavioral assessment was after the daily dose or before. Tests should be done before the daily dose to avoid acute effects of the exposure especially if the treatment is being given directly to the pups. If the treatment is being given to the dams, this concern is lessened but it is still preferable to test the pups before the dam is gavaged since removing her disrupts the litter.

Module B - Evaluation of Motor Activity Data

Figure 1: add that the data shown are Mean ± SEM (or SD) to the caption. Also, add a brief description of the apparatus (shape and dimensions).

Page B-9, first full paragraph: A common issue with video tracking systems is accuracy problems as these systems are prone to artifacts. All tracking systems are based on object-to-ground contrast. If the floor is shiny or the animals defecate and urinate, the tracker may pick up reflections off of the urine and feces and bounce back and forth between the animal and the reflection. We had this problem when we tried using video tacking for open-field and we abandoned it and returned to photocell systems because we could not solve the reflection problem. Once an animal urinated or defecated in a spot that was highly reflective to the camera lens there was no way to prevent the tracker from shifting back and forth from this point to the animal creating major artifacts in the data. If video tracking data are submitted, the reviewer should request access to the video files and randomly select tracking files from 20-30 animals and play them back in real-time and watch for sudden movements of the centroid that are faster than an animal could walk from one place to another. Such artifacts generally appear as jagged line between two points. If such artifacts are seen, the motor activity results should be rejected. In this reviewer's view, video tracking systems are not good for open-field applications and should be discouraged.

Page B-9, first bullet point: Session length needs to be emphasized. Although most strains of rats habituate within 30-40 minutes to square, round, or figure-8 arenas, 60 minute test sessions are preferable because shorter sessions do not show that asymptotic performance has been reached. Moreover, even if controls reach baseline by 40 minutes, if the treatment changes the habituation curve, especially if it lengthens it, the full habituation curve will not be seen in the experimental group. This will make interpretation of the data difficult. For this reason, 60 minute activity sessions should be strongly encouraged as good practice.

Page B-10: First bullet point on this page about cleaning equipment between subjects. This is a very important point that tends to receive little attention. Use of soaps, diluted ethanol (I see papers with anywhere from 10-70% ethanol solutions are used), or diluted bleach are not recommended. There are EPA-approved cleaners that can be obtained that are excellent denaturing agents and pose no toxic risk to animals or the personnel using them and are excellent antimicrobial agents. For example, Process NPD: Germicidal detergent (germicide, fungicide, verucide, detergent, deodorizer), 5 oz./gallon of water (EPA Reg. No. 1043-90) with active ingredients Octyl decyl dimethyl ammonium chloride (4.6%), Dioctyl dimethyl ammonium chloride (2.3%), Didecyl dimethyl ammonium chloride (2.3%), Alky (50% C14, 40% C12, 10% C16) dimethyl benzyl ammonium chloride (6.14%), Inert ingredients (84.66%); Steris Corp., St. Louis, MO 63133. This is NOT a product endorsement but an example of what reviewers should be looking for. 10% ethanol solutions are not effective denaturing agents; 70% ethanol may be effective but evidence is largely lacking whereas products such as Process NPD have been tested and have documentation to back up their effectiveness. Since rodents are sensitive to odors, cleaning equipment between animals with effective denaturing agents is important.

Page B-11, Par-1: While comparing early blocks or intervals to late ones can be done as a way of showing habituation, it would be helpful to indicate that the preferred way is to analyze the data in 5 or 10 minute intervals using ANOVA with interval as a repeated measure factor.

Page B-11: Section 6.2 “Reporting Data”, second bullet point: When "describing" data, I agree with the document that presenting Mean ± SD is appropriate. However, reviewers should be clear that SDs are only for descriptive proposes. When analyzing data by ANOVA and presenting inferential results in tables and figures the Mean ± SEM should be used since SDs show dispersion only in the sample and have no inferential utility, whereas SEMs are estimates of what the population variability is likely to be if many samples were tested.

Page B-11: Section “Data Analysis”: here and many times in the document litter as the sampling using for statistical analysis is emphasized, as it should be, because this is essential in developmental studies in multiparous species. However, how litter is handled in this document is too proscriptive. One way of handling sex is to have sex nested as the document sates. But there are several ways to do this. One way is with a hierarchical ANOVA (which are seldom used), treat sex within litters as a matching factor then use it as a within variable in the ANOVA model. But there is a third way: treat litter as a random factor in the ANOVA. When litter is a random factor then sex is a fixed or between-subject factor in the ANOVA since litter is already accounted for by using it as a random factor. In principle, using sex as a random factor is ideal because it does not contain the assumptions required for treating sex as a matching factor.

Page B-14: In this part on how to handle interactions from an ANOVA, I agree with what is being said that interactions need to be sorted. Hence, when there are group main effects in the absence of interactions then things are relatively straightforward. But when there are group interactions with other factors (time or sex), then interactions need to be sorted using simple effect tests or slice-effect ANOVAs, but it should be pointed out that slice-effect ANOVAs are better than simple-effect ANOVAs. This is because a slice-effect ANOVA uses the mean-square error term from the omnibus ANOVA and controls for doing multiple ANOVAs, whereas simple-effect ANOVAs calculate an error terms from each of the factors being sorted and also do not control for multiple ANOVAs. Also, it should be explained that once interactions are shorted, then within those slices that are significant (or for significant main effects), pairwise group comparisons should be performed. This is talked about later but I think should be mentioned here or say “see below in section x or y”. For DNT studies, the most important comparisons are from each treated group compared with controls. For this, Dunnett's test is advisable. If other *a posteriori* methods are used they must be ones that control for multiple comparisons, such as FDR (False Discovery Rate), Hochberg, Tukey-Kramer, step-down Bonferroni, etc. but not LSD, PLSD, Duncan, Tukey, Dunn, or any of the older methods that provide inadequate protection of alpha for multiple comparisons. Although the latter appear in older statistical text books, most statisticians are quick to point out that even the Tukey HSD test does not control alpha properly when there are more than 3 groups which is why the Tukey-Kramer was developed. The FDR is theoretically better than most of the other methods but some statistical packages do not offer it. It can be obtained in SAS through the MULTTEST program. Fortunately, the Tukey-Kramer is easily obtained directly within SAS GLM, Mixed, or Glimmix programs and is so close to the FDR in outcome that it is satisfactory. One odd fact about SAS is that when one wants the Tukey-Kramer test the command for it is “Tukey” but the actual output is the Tukey-Kramer even though the SAS output file does not indicate this (the SAS Manual indicates that it is).

Page B-14: Last paragraph before section 8: The part about age comparisons needs clarification. A two-way ANOVA of total activity by age where age is a between-subject factor is only appropriate if different animals are tested at the two ages. If the same animals are tested twice, then age would be a within-subject factor.

Page B-14: Bullet point 3 on Habituation in animals tested “at or after weaning.” The document assumes that rats are weaned at P21. No doubt this is often the case but this is changing. Attention needs to be given to when each laboratory weans. For decades rodents were weaned at P21 despite the fact that this age has no biological basis other than it is the youngest age at which laboratory rats and mice can be weaned and most of the offspring survive. It was “invented” by veterinarians and is a practical rather than optimal procedure. Rats and mice naturally wean their pups (even in laboratory settings) around P25. Because of this, a change in practice is occurring and more and more academic labs are weaning at P25 or P28. If a lab does this then a case may be made for testing animals for motor activity shortly after their specific weaning day. Rats weaned on P28 would logically be tested on P29 or P30; rats weaned on P25 could be tested on P26 or P27. This will result in different labs having different age-related activity profiles compared with those labs that wean on P21. Reviewers should recognize that such data will look different and take this into account. Also, it is not good practice to test activity exactly on the day of weaning. Separation of offspring from their littermates and the dam is a stressor. Giving 24-48 h for offspring to acclimate to separation is good practice.

Page B-15: Section 8.3 "Variability in motor activity data." Most of this advice is very good, but I'd be careful about what kind of scale of measurement one assumes activity data are. Activity data are counts and are not continuous (interval) scale data, so it is not a good idea to say they are in this document. While such data may be analyzed by ANOVA using the assumption of interval scale of measurement, since the underlying distribution is not interval and may not be normal, some statisticians recommend using different distributions, such as a Poisson distribution. Furthermore, a repeated measure ANOVA assumes data are related to one another in a continuous fashion if an interval scale is assumed, but this assumptions can be a problem sometimes. An example might be (in humans) how long it takes someone to tie their shoes measured repeatedly. A Poisson distribution might work betters because it does not assume the underlying distribution is continuous, it can discrete such as show lace tying or beam breaks in an activity monitor. There is no absolutely right answer to this issue. I've talked to statisticians who disagree about how activity counts across intervals should be analyzed and they don't all agree, but I think that the best advice for reviewers would be that there are different assumptions about the scale of measurement that can be made for activity but whatever method/distribution is assumed it should be explained so the reviewer is fully aware of how the data were analyzed.

Page B-16: Section 8.3, Par-2: Here again, I would not endorse the Tukey test, but rather the Tukey-Kramer test.

Page B-18: Figure 2 caption: The caption should say what the data are. In the left panel, I presume these are Means, but are the error bars SEMs or SDs? For the right panel, are these Means? I presume so, but error bars are missing and are needed. If this is to be an effective document one needs always state what the index of central tendency is and what variance measure was used and show the Mean and error for all data in any figure or table.

Module C – Evaluation of Acoustic Startle Response Data

Page C-3: Section 2 "Test Description": I suggest not jumping around with different terms, i.e., acoustic vs. auditory startle response. I'd suggest indicating upfront that these are equivalent terms and then use only one in the remainder of the document. I'd further suggest abbreviating Acoustic Startle Response as ASR (which you do at one point but then stop using it and spell it out every time). Rather than that, once the abbreviation is introduced then use it consistently thereafter. PS—The startle literature is now dominated by the term "Acoustic" rather than "Auditory" Startle Response, so I'd recommend following this terminology.

Page C-3, Section 2, Par-2: References at the end of the paragraph to review papers on startle: These references are old. A newer, and better, review is: Gomez-Nieto et al. (2014). Front. Neurosci., 8: 216. You might want to add it.

Page C-3, last line of last paragraph on the page: Two items are worth mentioning here. The document says that the prepulse typically lasts ~50 ms. Actually that is not the case in most of the PPI published literature. More typical is 20 ms. So, why not say "20-50 ms". Then it says that the prepulse precedes the pulse by 80-100 ms. Actually, the range most common in the literature is 60-120 ms or according to Mary Geyer in one of his reviews can be anything from 30-500 ms. Finally, in the last sentence on Page C-3 that runs over to Page C-4, it says that the prepulse is presented prior to the pulse, which is correct, of course, but it does not state how this interval between signals is measured. This interval is measured from prepulse onset to pulse onset. This should be made clear because I've seen mistakes where people think it is from prepulse offset to pulse onset and this is definitely wrong.

Page C-6 (continuation of Table 1 from previous page), part on "Study design." I indorse the document's recommendation for using PPI. There is no doubt that this procedure is underutilized in DNT studies but should be. With little additional time and effort PPI could provide much more data, and more valuable data since PPI taps higher brain centers. In addition, I'd recommend testing animals for ASR and/or PPI more than once (our experience shows that 2 days is better). I'd recommend testing each animal twice on two consecutive days. We find that 2 days sometimes turns up effects not seen with 1 day of testing. Since this is an automated test the added time is minimal. For example, with 50 to 100 trials the test only takes 23-46 minutes. And doing this is getting more out of the test with little additional effort. Given how expensive and time-consuming DNT studies are to begin with, this small change (testing for 2 days and including PPI) would represent a significant return on investment.

Page C-6: First bullet point: I agree that it is not appropriate to use the A scale on sound level meters to set the acoustic signal, but rather than stopping at what should not be used, it would be more helpful to state what should be used, i.e., the proper index is the SPL scale (sound pressure level). Bullet point 2: Should add that the use of pure tones is not advisable. Instead, a broad band signal (white noise) is better. This avoids problems with pure tones where if the right frequency is not used one can get weak responses. Mixed frequencies generally elicit a more robust response too since they recruit multiple nerves rather than just a few as with pure tones. Besides, ASR is not a hearing test. If one wants to test hearing there are better methods (brainstem auditory evoked potentials). Bullet point 3: the usual standard for the animal holder should be that it is scaled to the size of the animal but is large enough that the animal can turn around; this ensures that the animal is not so confined that it cannot flinch and also that it is not be too big. When it is too big the animal can position itself in places where the pulse is less intense since all test chambers bounce the signal around created high and low nodes. This is avoidable by confining the animal to a limited space directly beneath the speaker.

Page C-6: Last paragraph: This section on load cells versus accelerometers is one of the oddest things I've ever read about ASR. Let me start with what these devices are so we are all on the same page. I will focus on load cells and accelerometers and skip other (older) devices that are no longer in common use. What I present below can easily be found on the internet so anyone who wants to crosscheck what I'm saying can easily do so. For example, Wikipedia and many engineering sites can be found with this information.

"A load cell is a transducer that is used to create an **electrical signal** whose magnitude is directly proportional to the force being measured [emphasis added]. The various types of load cells include hydraulic load cells, pneumatic load cells and strain gauge load cells." [Startle systems that used load cells use strain gauges; but no commercial vendor of startle equipment that I was able to find that still uses load cells; they are passé.]

"An accelerometer is an electromechanical device used to measure acceleration forces. Such forces may be static, like the continuous force of gravity or, as is the case with many mobile devices, dynamic to sense movement or vibrations. Acceleration is the measurement of the change in velocity, or speed divided by time."

What are load cells primarily used for? Balances and scales. They are great at measuring static loads and their major applications are for measuring mass. They are good at this precisely because they output a steady current (change in voltage) in response to a constant force. In labs they are used in analytical balances to measure reagents and in vivaria to weigh animals. In recent years live animal balances even take readings every 1 second and average them over 10 seconds to give a more accurate measure of body weight as animals move around. So for the right applications, load cells are great. Load cells are used to weigh things all the way from tiny amounts in analytical balances all the weigh up to 18-wheeler trucks, train cars (loaded versus empty) and thousands of other applications in industry and science. But they all have range limits; one must select a load cell that is sensitive within the range of weight one wants to measure and one that can respond quickly enough to change and the latter is an issue with load cells. But notice above in the description that what they output is an electrical signal (change in voltage). They do not output grams or Newtons. They can be calibrated to translate the electrical signal to a scale of grams or Newtons, but that is not what they do at a fundamental level. Calibrating load cells to grams or Newtons is common and one can buy load cells off the shelf that are already calibrated to give readouts in grams or Newtons, but what they output is a change in voltage, specifically, mV; there is no device that measures gravity per se. Because they put out a constant voltage in response to a constant force, they vary in their ability to detect rapidly changing force. Engineers have gradually made load cells more and more sensitive in their response rates so that some can detect changes fairly rapidly but that is not what they are designed for so using them to measure rate of change is somewhat of a misapplication.

I have no idea who wrote this section of the document but with all due respect they need to explain why they believe accelerometers are inferior to load cells. There's not a single reference cited to support this assertion and it contradicts everything I know about startle. And not just me, it contradicts what Michael Davis, Jim Ison, Mark Geyer (3 of the leaders in studying startle) and others say about startle and how it is measured. Most startle papers report Vmax (the largest response peak) as the index of startle amplitude. If the EPA has authoritative data (citable) that shows that they (and the field of startle neuroscience) is wrong, please present it. Otherwise, this section needs revision.

As I read it, I was struck by how the document tilts in favor of load cell technology. Let me mention disadvantages of load cells that are not mentioned but should be. Because load cells are designed measure static loads, the animal's weight sets the output signal above zero just by placing the animal in the startle apparatus. This body mass signal must be subtracted from the change in output generated by the startle response. Load cells never start at zero one the animal is present in these systems; since every animal has a different starting point, each animal's body weight must be subtracted from the change in load when the animal flinches to the pulse. This leads to several things: (1) it requires that body weight be subtracted from the change induced by the startle response for each animal individually and a secondary dataset created; in effect one is not analyzing original data using this type of detector, and (2) the animal's body weight takes up a significant portion of the dynamic range the load cell. This has the effect of limiting the range of sensitivity available for detecting a change from the startle reflex. Because load cells are designed to measure static load, they also tend to be sluggish in their response to a change compared with accelerometers. This is why load cell response curves do not show every peak generated by the animal's different muscle groups but merges them together, in effect obscuring aspects of the full response. (3) Body mass has inertia (Newton's second law), i.e., an object in motion tends to stay in motion and an object at rest tends to stay a rest. For load cells, when the animal contracts its muscles its static body weight inhibits movement in the platform and hence force on the load cell. Once the contractions end the mass has to return and when it does it causes the load cell to receive added force. For example, if a person stands on a bathroom scale and jumps, the scale will read a higher number first from the force the muscles in the legs jump pressing downward; then the scale reads less that the person's body weight while they are in the air; then when they land the force of landing again causes the scale to read above their static body weight. Hence, load cells tend to show positive and negative wave forms. This is an issue since oscillations obscure some of the waveforms generated by the startle movement. While it makes output curves look nice it does so my blending out some of the movements while the mass of the animal is shifting.

What the report describes as a disadvantage of accelerometers is in fact an advantage. Because they don't respond to static load, there is no need to subtract body weight and hence no loss of dynamic range (and hence no loss of sensitivity) as occurs for load cells. Accelerometers were developed to detect vibration primarily, such as, in engines, high speed pumps, motors, jet engines, and high velocity impacts (such as crash testing). These phenomena are dangerous and cannot be detected with load cells which is why engineers abandoned load cells decades ago for the detection of vibrations and impacts (acceleration type events). To accurately detect high rates of change requires detection devices capable of rapid response. Before accelerometers, aircraft engines under development flew apart in midair with catastrophic results; load cells failed in these settings which is part of why accelerometers were developed. If one looks at load cell startle output waveforms one sees positive and negative defections from the rest resting state (the animal's body weight). In accelerometer systems one never sees negative deflections because it only records acceleration (rate of voltage change) so no matter which way the movement is. While an acceleration can be increasing or decreasing (deceleration) it is never negative. Hence, accelerometers are ideally suited for ASR.

The document makes a point that load cells provide data in terms of grams of force. This is true if the load cell has been calibrated for this, but accelerometers can do this too if one cares to calibrate them to do this, but that's not the issue, the real issue is why would want grams or Newtons? Some load cells can be purchased pre-calibrated that output voltages and convert it to grams, accelerometers usually are sold without such converters but one can obtain these if it is worthwhile to do so. The central question is: What scientific purpose is gained from knowing grams of force? One may be able to measure responses in grams but it is of no value for understanding the results why does the Agency think this is worth knowing? One cannot compare results across experiments or ages unless body weights are equal, which they seldom are, so grams of force provides no interpretive value. So what does one do with grams of force? If it is captured it doesn't change the fact that with load cells one still has to use statistical methods to adjust body weight, whereas with accelerometers this does not have to be done. I suggest that this section be re-written so that it is neutral about load cells vs. accelerometers; or better yet, don't discuss detectors at all. Instead, state that there are different types of detectors and that the type used and how it is calibrated must be described. The Agency should not be promoting one detector over another; and the Agency should stop saying that one type of detector outputs grams and the other voltage: they both output voltage, there is no other way an electrical device can respond. Also, The Agency should stop saying that voltage is 'arbitrary.' Voltage is not arbitrary; it is a very precise physical entity (electrons). An arbitrary scale is one with no absolute reading; typically is applies when a measurement has no zero value. Last time I checked voltage can be zero. If voltage can be zero then it is not arbitrary. Further, on arbitrary scales there is no assurance that it is linear. Voltage is very linear. 50 volts is exactly twice 25 volts, so in these terms mV is not an arbitrary scale of measurement. I found this section problematic.

Page C-8: Par-2: I'm not clear what source was used to support the view that ASRs are over in 20-40 ms. Just on the face of it this is suspect. The response takes ~14 ms to begin, and does not reach its first peak until 20-30 ms and since there are multiple peaks (caused by the rostral-to-caudal expression of the response along the length of the animal's body axis as efferent signals reach different neuromuscular junctions) it can't possibly be over in 40 ms. If the response is over in 40 ms, do all experts in this field record ASR data across 100-200 ms response windows? It is not because these investigators record useless data, but rather because the data show that the responses 70 ms or more. I suggest this range be changed to 20-70 ms.

Page C-8: Figure 1 caption, where it says that the second peak is inappropriately used as the startle response: Where does this idea come from? I don't see this in the ASR literature. The second peak in accelerometer systems is Vmax. The first peak is presumed to arise from contraction of the head and neck, whereas the second peak is thought to be caused by contraction of forelimbs, diaphragm, and abdominal muscles which results in the hunched posture seen in high speed videos of startle responses. This is the largest set of muscles recruited by the pulse. There is also usually a third peak (which is cutoff in this figure) that is thought to represent hindlimb contractions. I think arguing against the larger peak is invalid and contradicts the startle literature including such experts as Michael Davis and Mary Geyer.

Page C-9: Section 5 “Test Procedures”: last bullet point: I think you should add that in addition to all the characteristics that need to be reported, that the startle signal onset rise time should be specified. Most systems are in the ranges of 1-1.5 ms, but it is good to have this information.

Page C-9, Section “Data Reporting” Third bullet point and further down: I think you make too much out of response latency and response onset latency. First, no one uses response onset latency so I have no idea where this comes from or why anyone would even want this. Since it is not in the scientific literature how would the Agency interpret it? Reviewers would be trying to interpret it in a scientific vacuum. I’d also reverse the order and list the variables to be defined in this order: “peak response amplitude, average response amplitude, and latency to peak response” since this is the way the published literature is reported.

Same section, last bullet point: Here the document gets around to asking for pulse rise time, which is good, but it needs to be above too (or both places). Fall time is essentially useless because it has no impact on the response. The Agency can leave this or delete it but either way it is a factoid of no scientific value. Same bullet point where it mentions “frequency of response for speakers used” this is trivia. If the lab measures the pulse using a high quality sound level meter on the SPL scale, then one knows what the signal is. Details of the speaker is of no additional value since the only thing that matters is the nature of the signal reaching the rat, not of the generating device. If the speakers can’t output a 110-120 dB SPL white noise burst then the speakers are inadequate; if they can, who cares what the speaker specs are?

Page C-10, first bullet point: This bullet point needs to be redacted. It is inaccurate, incorrect, and could cause the Agency to get misleading data when the real data the agency should get is the peak response amplitude which is usually the second peak in accelerometer systems (see above).

Page C-10, second bullet point beginning “Note that animals do not . . .” this is a true statement but where is says that the experimenter should address this is unclear to me. The question becomes how should or could it be addressed? Seems vague. In practice this phenomenon is automatically ‘addressed’ in the sense that ASR data are analyzed in blocks to smooth out trial-to-trial variability. If the Agency is suggesting something else, such as that non-response trials be removed, this would be problematic. This is because there are few trials that are zero mV; rather there are weak responses. If ones looks at ASR raw data, ones sees that zeros are rare but weak responses are not rare. Given this, what threshold would one use to determine which trials to exclude? I can assure you that every possible value one can imagine will appear in ASR data. I think asking this to be ‘addressed’ is not a good idea since it is unclear how it could be addressed by anyone. The document recommends 10 trial blocks. This is the most reasonable way to deal with ASR variability which is why everyone who uses ASR does this or something similar (some use 5-trial blocks). I don’t think it is advisable to ask for things even startle experts don’t do. It will only cause consternation for reviewers.

Page C-10, Paragraph starting “The terminology used in reporting these measures . . .” where is says “The average response magnitude does not provide an accurate measure of the maximum (or peak) response magnitude”. Really? Who says so? In my lab we’ve compared Vmax and Vavg across multiple experiments and the two show correlation coefficients of 0.96-0.98. So who says that these are not nearly identical? Mark Geyer in fact argues that Vavg is the preferred measure (certainly he prefers it) and he’s one of the leaders in startle research. His argument is that it captures the integrated full response waveform. He’s correct about that. I would only add that since it is highly correlated to Vmax it doesn’t add anything to the peak response so in practice it appears to make little difference of one reports Vmax or Vavg. But please do not denigrate Vavg as inferior; there are no data to support this claim. We use Vmax but to say there is something wrong with Vavg would require extensive evidence and I don't know of such evidence anywhere in the literature. Same paragraph where it refers to latency to response onset: either cite solid references to support the unique value of this measure or drop it. Again, no one publishing in the ASR field is reporting this measure.

Page C-10: Table 2: This long table is unnecessary. First, Vmax is measured in mV, so just because a lab fails to indicate this in their reports does not mean that Vmax and mV data are different, they aren't different. Also, data in Volts is obviously a shorthand since ASR instruments measure in mV. If you asked the lab reporting this to check their owner's manual I'm sure they'd correct it to mV. The same applies to Tmax. Tmax is measured in ms, so the fact that some decide to use one term and some another doesn't mean they are different. I think a simple table should replace this one and include only the following:

Vmax (mV) is a measure of peak amplitude
Vavg (mV) is average amplitude across the recording window
Tmax (ms) latency to Vmax
Grams of force
Newtons: formula for obtaining Newtons from mass

$$F = m \cdot a$$

$$1 \text{ N} = 1 \text{ kg} \cdot \text{m/s}^2$$

In this this formula F=force, m=mass, a=acceleration; hence to get Newtons from mass, the formula is that 1 N is equal to 1 kg x mass divided by seconds-squared (should anyone want to know). My point is that it is trivial to get Newtons once on has grams of force but the bigger point is why does one want either of these? There is no reason scientifically.

Page C-11: section 6.1 "Dependent Variables", first bullet point: I think (for the reasons outlined above) this bullet point should be removed.

Page C-11: third bullet point: Why is this referred to as 'arbitrary' units? As stated above, measuring electrical current in volts is one of the most accurate measurements possible in physics and there is nothing arbitrary about it. Voltage is used as a measure by the most precise instruments in the world, including cyclotrons such as the Large Hadron Collider at CERN. My house current is measured as 110 V. If I overload a house circuit rated at 110 "arbitrary" units I'll cause a short circuit. I can't plug a 110 "arbitrary" unit device into a 220 "arbitrary" unit device or I'll burn the device up. I'm being facetious so I'll stop. If you want you could call it "relative" but why do this at all? Why not be clear and direct and say that all detectors record in mV. Please stop claiming that load cells record in grams it is simply not true. All detectors measure a change in the deformation of a multilayered sandwich of materials with ionic potential. When these materials have pressure exerted on them they deform. This deformation induces an electric signal and it is this analogue signal that is digitized and recorded.

Page C-12: Figure 2 caption: It says the data are "mean response". Mean of what? Mean Vmax? Mean of something else?

Page C-12: last paragraph on the page, 4th line up from the bottom: In Fig. 2 it says that the data in Fig. 2 are from Sette et al., 2004 but in this paragraph it says that the data in Fig. 2 are from Raffaele et al., 2004. How can that be from both? In References, the Raffaele et al. paper is 2008 rather than 2004 but the real issue is how is it that the data in Fig. 2 can come from 2 difference sources? Did both papers use the same underlying dataset? Please clarify.

Page C-13: The first paragraph (which is a continuation from the previous page) is full of inaccurate statements that I've commented on above. I think it is a mistake to say the first response is the startle response and then compound the error by making up consequences of using peak amplitude. Both reasons given are bogus. It is fascinating that all the inaccurate statements in this document are unreferenced. Either these need to be documented from multiple labs so that this peculiar view can be shown to be accepted in the startle community or it should be stripped out of the report. I've been using startle and reading startle papers for 40 years and have NEVER seen what is being asserted here. I've known Kevin Crofton for ~30 years. Kevin devoted many years of research to using and analyzing startle data. Kevin's startle system was custom made because when he started there were no commercial startle systems. Kevin's system used load cell technology. I don't know if Kevin feels that load cells have advantages over accelerometers but the field as a whole does not think so. However, at least that debate has two sides to it, but this part about the first peak is an outlier. Does Kevin indorse this? I have never heard Kevin make such an argument nor have I seen Kevin indicate in papers that a second peak is false. I suppose that with a load cell perhaps Kevin found that the second peak reflects a rebound or oscillation that might not be part of the reflex, but that doesn't apply to accelerometer systems. But if this assertion comes from Kevin then he should support it with data and references, including references from other labs that find that their systems gyrate too and produce trailing readouts that not part of the animals startle response. Some early ASR systems were 'springy' and those designs could produce gyrations that I could imagine were not the direct startle response but I can't find anyone using such systems any longer. All the newer systems use a rigid platform so there is essentially no, or very little, reverberation after the startle response is over. Since Kevin in an in-house expert, he should read what I've said and either agree or disagree (doesn't matter to me if he disagree with me but if he does then he needs to support is view with references because I can support my view with references if need be).

Page C-13, section 7 "Data Analysis" first paragraph, add that litter can also be handled as a random factor in ANOVA models.

Page C-14: second paragraph stating "However, the use of body weight . . ." Technically, this is true in the strict sense of the assumptions of ANCOVA, i.e., that the covariate must be orthogonal to the independent variable, but most statisticians will in fact recommend trying ANCOVA (cautiously) even if the independent variable is correlated to the covariate. This is because ANCOVA is about the only way one can approach a problem such as this statistically. The proviso is that if one does this and the ANCOVA comes out different than the ANOVA, one has to acknowledge the limitations of the analysis, but such an analysis should not be forbidden. The same forbidding view is expressed in the last sentence too and should be removed.

Page C-17, last bullet point: Here it says that ANCOVA can be used when there are no treatment effects on body weight. If there are no treatment effects on body weight then why would one even consider doing an ANCOVA? I don't think this makes any sense.

Page C-18: Section 8, subsection 8.1 "Properties of startle control data", first bullet point: Throughout this section latency is mentioned over and over, usually first when in fact is it is much less important than amplitude. If one looks at the literature on ASR and PPI, amplitude is universally reported, whereas latency is rarely mentioned. Why is this? (1) Because latency is affected by amplitude; and (2) because amplitude is the response of interest whereas latency is not; therefore, most people don't bother reporting latency. While there are datasets where latency is changed and amplitude is not, what does this mean? One basically doesn't see such cases reported. Why? Because no one knows how to interpret such a strange effect and the usual suspicion when one occurs is that it is a Type I error. When amplitude is affected, however, latency often is too. When one examines the data the reason for this is clear. Higher (or lower) peaks require more (or less time) to reach their maximum, therefore, changes in latency that accompany changes in amplitude are a simple byproduct of the change in amplitude and provide no additional information. I suggest that if the Agency wants latency data, that's fine, but emphasize amplitude first and foremost and make latency secondary.

Page C-18, section 8.1, paragraph beginning "Assuming that the testing system . . ." The first two sentences of this paragraph make little sense. First, it says that "if" the system is in absolute units then one should see an age effect, then the example used doesn't present data recorded in absolute units, but rather in relative units (probably mV which if this is correct should be stated as such). These sentences are confusing and not germane to the 4 points made later in the paragraph. Why not start with a reference to Table 3 showing startle data for two ages and then make the four points and leave it at that?

Page C-19: Paragraph beginning "Table 4 illustrates control data that have some inconsistencies with the known biology of the startle response in terms of age and sex." This and what follows is hard to believe and I in fact don't believe it. While I'm not privy to the system used to generate these data, the data don't look like they are apple to apple comparisons. In ASR systems, you would not normally put a P23 and P61 rat in the same animal holder. In addition, in these systems one sets the sensitivity according to the size of the animal and animal holder, so you can't compare absolute amplitude across ages, nor is there any reason to do so. Perhaps these data are flawed as you say, I really don't know, but what I can say is that I can generate data like these by setting the conditions a certain way in my startle equipment. So why are you holding these data up as a "whipping boy"? There may not be anything wrong with them but in order to know you'd have to provide details, which you don't. Instead, you're pulling data out of context in order to criticize them while depriving the reader of the details to know if what is being said is correct or not. I would delete all this; it makes no sense based on what I know about startle.

Page C-20, paragraph beneath Table 4 that begins "The issues noted in Table 4 should . . ." This paragraph does make sense. So, rather than setting up a straw man on the previous page so you can knock it down (albeit in a flimsy way), just take the smoke and mirrors out of Page C-19 and keep this part on Page C-20.

Page C-25, section 8.6, third line up from bottom of paragraph the word "animal" should be "animals"

Page C-28, second bullet point about Vavg being inaccurate. Again, this is contradicted by mountains of published data. If the Agency wants to go against the grain, this assertion will certainly do that but at the risk of making the Agency's look bad.

Page C-28 and throughout the whole C section on startle, I advise taking out the annoying reference to accelerometers recording in "arbitrary" units (see above). Refer to them as measuring voltage. Three is nothing arbitrary about voltage (see above), if that were true and I plugged my toaster into an outlet and the voltage was arbitrary I would have no clue what would happen from one use to the next. I think it would come as a shock to physicists, engineers and the electric power industry that electricity was measured in "arbitrary" units. A characteristic of an arbitrary unit is that it has no meaningful zero value, yet voltage can be zero, so how can one assert that voltage is "arbitrary"?. And the last sentence of this paragraph is equally problematic. Force transducers output voltage change, not grams of force.

Page C-29, first bullet point, I recommend against saying that the peak response occurs at ~20 ms. Using a range would be better (20-40 ms). Further along I'd delete anything about what constitutes a problematic latency using specific numbers. It would be better to simply state that unusually long latencies may be a sign of a problem and such data should be scrutinized to ensure that control animals are startling appropriately.

Comment: The Startle section was challenging. It has many problems. I can't imagine I caught them all.

Section D on Learning and Memory

Page D-4, second paragraph, about midway, rather than say "(operant conditioning)" it would be more accurate to say "(instrumental learning and/or operant conditioning)".

Page D-4, next to last line on the page: Actually this statement is not quite correct. Time is a critical variable in the definition of working memory in people, but not in animals. In animals the typical definition is "trial-dependent memory". I'd suggest adding this.

Page D-5, Table 1, under "Cincinnati Maze" in the right-hand column it says "Intramaze cues detract from its sensitivity", I've never said this and I invented this maze. Proximal cues are essential for this maze. Please remove this. Also, in the third column please change "Sequential Learning" to "Sequential/Egocentric Learning"

Page D-7, Table 2 under "Age of testing" while I realize that you are quoting from the OECD guideline for the P25 test age, I think this is a bit problematic if a lab is weaning on P28. I can't say when most CROs wean but if they wean on P28 then testing on P25 or P27 would not be appropriate. Why not say that testing should be 1 to 2 days after weaning and not mention an age?

Page D-8, first paragraph, line 5-6: It's a bit inconsistent to say on line 5 that rats have to make a binary decision to escape water then in the next sentence refer to these as being either land or water mazes. I'd suggest taking out "escape the water" in line 5.

Page D-10: first paragraph: In simple mazes with only 2 choices, guiding animals to the goal can be done but is difficult in practice. Are labs actually submitting data from letter water mazes stating that they guide animals that make a wrong turn? If so then perhaps some mention of this should be kept, but I'd point out that guiding animals creates its own issues; do all experimenters guide animals exactly the same way? How is guidance done? Is it done with a stick, by blocking off incorrect alleys, by putting a barrier behind them, or some other way? How far from the animal's nose do they hold the guiding object if they use a stick or similar item? What do they do for animals that don't follow the object? Do they confine animals to the wrong alley before guiding them to the correct side? In a simple letter maze do animal really need to be guided? Can't they find it on their own after they go the wrong way? After all, rats don't stay in the wrong alley if the goal isn't present which leaves only 2 choices left: back to the start or to the correct arm. It doesn't take rats long to figure this out, so why bother guiding them and introducing experimenter effects?

Page D-13, paragraph starting "Typically, animals that do not . . ." This section gives good advice on problems with trials to criterion methods but it applies most to cases where the number of trials allowed is fixed at a relatively low number. It might be worth pointing out that some of these problems can be rectified if the lab first determines how controls perform. If the lab establishes that controls learn the task in (Mean ± SE trials) 17 trials ± 3 (or if they find that the worst performing animal reaches criterion in, say, 19 trials), then the lab should use the upper boundary as the number of trials for all animals rather than some arbitrary number decided by the experimenter to fit in a preconceived protocol.

Page D-15, last bullet point: I think you should re-emphasize that these simple mazes (especially swimming versions) are not used in academic labs because they are known to be insensitive. Given this I think the Agency should discourage there use. Another problem with these mazes that is not mentioned is that rats show stronger turning biases in binary water mazes than in dry mazes or in complex water mazes (for reasons no one has ever figured out; although I have a theory about it). Because of this, it is critical that animals be tested to their non-preferred side to help offset this, but even this does not solve the problem. Also, it is better to use procedures similar to those used in dry T-mazes, i.e., confine the animal in the incorrect arm if it makes an error. I'd recommend against guiding animals or letting them self-correct; instead confine them to the incorrect side and then remove them immediately. This should have the effect of flattening the learning curve slightly making the test a little bit more sensitive since these tests tend to have learning curves that are too steep to begin with. What I'm recommending will make the test a little more difficult and maybe a little more sensitive (in theory at least). But the fundamental problem is that these rudimentary mazes are not very good and that is something that cannot be fixed.

Page D-15, section 3, point #2, the clause "or in tests involving a single trial . . ." should be removed as there are no MWM procedures that only use a single trial.

Page D-16, Par-3, sentence about making the water opaque: This is a myth. This is not required and has to do with the history of the test and is an historical relic. What is required is that the platform be camouflaged such that the animal cannot see it. This can be done different ways and all are acceptable: (1) make the water opaque (as you say), (2) make the maze and platform the same color, such as black (a black platform cannot be seen against a black background of the tank (good for testing albino animals) or make the platform white and the tank white (good for testing C57BL mice), (3) make the platform clear, which in water looks like water and is invisible to the rat or mouse, (4) make platform so it can be raised from the bottom by remote control. Make this sentence general and encompass all methods or just say it must be satisfactorily camouflaged (this can be proven by a simple experiment that I can describe if need be).

Page D-17, last paragraph, where it refers to guiding animals out if they reach the time limit: There is no agreement among MWM experts as to whether assisted or unassisted escape is better. I'd recommend being agnostic on this point and say that which method is used must be specified. Don't go out on a limb and say that one method is preferred.

Page D-18, Fig. 4, right panel: The caption needs to explain the gap in the abscissa.

Page D-18, Fig. 4 caption: Captions should begin by stating what the data are, such as: Ordinate shows Mean ± SEM latency (or whatever it is) per day with 2 trials per day (left) or 4 trails per day (right).

Page D-19, paragraph with heading "b) Visual Performance Control" about moving the platform: There is a critical missing aspect to how this must be done. On cued trials, both the position of the platform and the start must be moved randomly on EVERY trial to prevent route learning. Without this the data are largely worthless.

Page D-21, section 3.2, second bullet point: Here again do not stop at mentioning making the water opaque; make it more general, i.e., that the platform must be camouflaged. That's all that need be said.

Page D-23, "Dependent Variables" Do you really want both trial and day in these ANOVA models? If you do, the ANOVAs get complex with no evidence of any gain in understanding of spatial learning. I'd suggest you indicate that the data be analyzed with day as the repeated measure factor because this provides the information that is of greatest importance. If you insist on having labs do 2-between, 2-within ANOVAs it is going to generate complex interactions that may make the data harder to interpret; and by doing more F-tests, increases the chance of Type I errors. Furthermore, multiple significant interaction terms are difficult for both the experimenter and the reviewer to sort and understand. Just to be clear a 2-between, 1-within model will generate the following F-tests: Group, Sex, Day, Group x Sex, Group x Day, and Group x Sex x Day, i.e., 6 F-tests, but you insist on models of 2-between, 2-within design it will generate the following F-tests: Group, Sex, Day, Trial, Group x Sex, Group x Day, Group x Trial, Sex x Day, Sex x Trial, Day x Trial, Group x Sex x Day, Group x Day x Trial, Group x Sex x Trial, Sex x Day x Trial, and Group x Sex x Day x Trial, i.e., 12 F-tests. Why introduce this much complexity? I'd recommend that labs submit data analyzed using 2-between, 1-within models with day as the within factor and if graphical representation by day and by day and trial suggest that there are or may be within day effects, then they (on their own initiative) or with feedback from the reviewer can then do a 2-between, 2-within ANOVA as a follow-up.

Page D-23, last bullet point under "Dependent Variables": Consider adding one more thing. If in an analysis of acquisition there are group differences, including on day-1 (which will be obvious if the data are graphed), then it is important for the lab to do a separate analysis of the day-1 data only with trial as the repeated measure factor. One needs to know how the groups start out. If they start out similar to controls on the first and/or second trial and then separate, that's fine. This pattern suggests a learning impairment. But if the groups are different even on trial-1, that's a different story. A pattern like that suggests a preexisting performance deficit that may not indicate a learning problem especially of the by-day graph shows that the groups improved in parallel and the ANOVA shows a main effect in the absence of an interaction between Group and Day. Reviewers need to understand this key distinction.

Page D-23 under "Reported Data" I think it advisable to rephrase the first 3 bullet points to state that the data should be provided per treatment group on each "day" (not on each "trial"). If the agency wants it per trial then ask for it both ways but the by-day data should be the focus during review.

Page D-24, section 3.5, bullet point 4, this is the point I made above; I'm glad it is made here but it should be mentioned above too.

Page D-29, bullet point 5: I suggest adding that if group differences are found, then it is essential that a test of shock sensitivity (for pain and reactivity) be performed, otherwise the data are impossible to interpret since effects may be the result of a different pain threshold or a difference in reactivity (which is not the same as pain).

Page D-30, section 5.1, "Initial Acquisition": Several critical details need to be mentioned that are missing, as follows:

First, testing rats in a straight channel before testing in the CWM maze is an absolute requirement. If this is not done the test will yield compromised results because the maze is so complex (especially path-B) that it will induce giving-up behavior in many rats and they will stop searching. This behavior can be largely eliminated by giving straight channel trials first because it teaches rats that escape is possible. Also, it should be stated that the CWM is not to be used with mice. We have not found a strain that can reliably learn a maze of this complexity.

Second, if a rat reaches the time limit on trial-1 of a given day, it is critical that it be given at least a 5 min. rest before being given its second trial of the day. Otherwise, fatigue will compromise its problem-solving ability and distort the results.

Third, there will be an occasional animal that stops searching and swims in one spot until the time limit is reached. Such rats should have a corrected score used for errors; otherwise such a rat's low error score will underestimate the true deficit. What we use is the number of errors made by a control animal that made the most errors. This is a conservative estimate of the number of errors a failing experimental animal would have made had it continued searching. It is important to report how many animals received corrected scores. And we analyze our data both ways, with uncorrected and corrected error scores. Most of the time these turn out to show the same thing. In cases where the treatment causes many trial failures and many errors there is no need for a corrected error score. But once in a while a treatment will cause many trial failures and very low errors because the experimental animals stop searching and treaded water in a small area of the maze making almost no errors; this is rare, but in 30 years of experience with this maze we have seen it. These are the cases that require close examination and analysis of the data with both corrected and uncorrected errors to get a grasp on what effect the experimental treatment is really doing. My advice is that if such a weird case is ever seen it would be prudent to have a water maze learning expert review the data.

Page D-32, section 5.3, bullet point 3: It is worth pointing out to reviewers that an alternative to testing rats in path A and then path B (as in the associated figure), to instead use the perfectly valid approach of testing in path B only. This is because path B is the procedure that is most sensitive to neurotoxic agents, it avoids transfer of training effect, and we find it sufficient by itself. We no longer use Path-A.

Page D-37, section 8, line-7: again, litter can be handled as a nested factor or a random factor.

Page D-38, section 8.3: To be precise, count data are not continuous. All count data of anything are not continuous. While such data can be analyzed as if they were continuous and doing so can often works well (and what most people do), statisticians note that the underlying distribution of count data is rarely normal (and is not interval). Rather such data may fit other distributions (there are many distribution types, such as Poisson). Some care should be exercised in analyzing count data whether generated from mazes or activity (beam breaks) as they can present problems sometimes. I'd suggest adjusting this line and perhaps adding that count data are not continuous and may not be normally distributed, therefore, care should be exercised in how such data are analyzed; and if analyzed using non-interval distribution assumptions then this must be described.

Page D-38, Par-1: Non-normality and tests for them are mentioned on this page which is appropriate, but the other assumption of ANOVA is homogeneity of variance and this is not mentioned. However, the robustness of ANOVA for departures from normality and from homogeneity of variance should also be noted as warnings, not prohibitions against ANOVAs since most departure (as long as not too large) usually don't invalidate such statistical methods. This is because ANOVAs are robust and work well even then these assumptions are not fully met. In noting this it should be mentioned that moving to non-parametric methods is no panacea; as these methods have their own problems including that they are less powerful compared with parametric methods.

Page D-40, section 8.4, about midway through the first paragraph where it refers to Type 2 errors. This is a mistake. It should say Type I errors.

Section E – Weight of Evidence

Page E-6, bullet point 3: rather than "stupor" I'd suggest using the more standard pharmacological term: "sedation."

Page E-7, section 3.1.4, bullet point 2, here again I'd add that another way is to use litter as a random factor in the ANOVA.

Page E-11, second bullet point: I agree that the BBB and transporters are important but I'd add one more: the gut-blood barrier which also matures over weeks and determines basic uptake.

- 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?

RESPONSE: No. I have outlined the reasons the document falls short of providing appropriate guidance to reviewers above.

- 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?

RESPONSE: In some places yes, and in some places no. For instance, some of the examples provided are apt and some are not optimal; in other cases the figure may be satisfactory but the information about the data depicted in the figure are inadequately described in the text or figure caption. These issues are described above.

WOE Module

- 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?

RESPONSE: In general the WOE section is fine. I made a couple of suggestions in this section.

References

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

Elsner J, Suter KE, Ulbrich B, Schreiner G. 1986. Testing strategies in behavioral teratology: Iv. Review and general conclusions. Neurobehav Toxicol Teratol 8:585-590.

Francis EZ, Kimmel CA, Rees DC. 1990. Workshop on the qualitative and quantitative comparability of human and animal developmental neurotoxicity: Summary and implications. Neurotoxicol Teratol 12:285-292.

Grandjean P, Landrigan PJ. 2006. Developmental neurotoxicity of industrial chemicals. The Lancet 368:2167-2178.

Hass U. 2006. The need for developmental neurotoxicity studies in risk assessment for developmental toxicity. Reproductive toxicology 22:148-156.

Holson RR, Freshwater L, Maurissen JP, Moser V, Phang W. 2008. Statistical issues and techniques appropriate for developmental neurotoxicity testing. NeurobehavToxicol doi:10.1016/j.ntt.2007.06.001

Ladics GS, Chapin RE, Hastings KL, Holsapple MP, Makris SL, Sheets LP, et al. 2005. Developmental toxicology evaluations—issues with including neurotoxicology and immunotoxicology assessments in reproductive toxicology studies. Toxicological Sciences 88:24-29.

Li AA. 2005. Regulatory developmental neurotoxicology testing: Data evaluation for risk assessment purposes. Environmental toxicology and pharmacology 19:727-733.

Makris SL, Raffaele K, Allen S, Bowers WJ, Hass U, Alleva E, et al. 2009. A retrospective performance assessment of the developmental neurotoxicity study in support of oecd test guideline 426. Environmental health perspectives 117:17-25.

Makris SL, Vorhees CV. 2015. Assessment of learning, memory and attention in developmental neurotoxicity regulatory studies: Introduction. Neurotoxicology and teratology 52:62-67.

NRC. 1993. National research council. Pesticides in the diets of infants and children. Committee on pesticides in the diets of infants and children. National academy press, washington, dc.

OECD. 2007. Test no. 426: Developmental neurotoxicity study. Oecd publishing. Http://dx.Doi.Org/10.1787/9789264067394-en.

OECD. 2011. Test no. 443: Extended one-generation reproductive toxicity study. Oecd publishing. Http://dx.Doi.Org/10.1787/9789264122550-en.

Piersma AH, Tonk EC, Makris SL, Crofton KM, Dietert RR, van Loveren H. 2012. Juvenile toxicity testing protocols for chemicals. Reproductive Toxicology 34:482-486.

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.

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

Slikker W, Jr., Acuff K, Boyes WK, Chelonis J, Crofton KM, Dearlove GE, et al. 2005. Behavioral test methods workshop. NeurotoxicolTeratol 27:417-427.

Spyker JM. 1975. Assessing the impact of low level chemicals on development: Behavioral and latent effects. Federation proceedings 34:1835-1844.

Tilson HA, Wright DC. 1985. Interpretation of behavioral teratology data. Neurobehav Toxicol Teratol 7:667-668.

Tsuji R, Crofton KM. 2012. The developmental neurotoxicity guideline study: Issues with methodology, evaluation and regulation. Congenital Anomolies (in press).

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 30:349-381.

U.S.EPA. 1998a. Health effects guidelines oppts 870.6300 developmental neurotoxicity study, epa/ 712/c-98/239.Office of Prevention Pesticides and Toxic Substances.

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

Vorhees CV, Makris SL. 2015. Assessment of learning, memory, and attention in developmental neurotoxicity regulatory studies: Synthesis, commentary, and recommendations. Neurotoxicol Teratol 52:109-115.