RESPONSE TO REVIEWERS

Standard Evaluation Procedures for Submitted Developmental Neurotoxicity Data

Each reviewer's comments are copied or summarized where needed: responses are in italics.

General comments

Reviewers 2 and 4 provided no general comments

Reviewer 1

Grammar comments: "sex" not "gender", "dam" not "mother", others.

All specified changes made.

Strengthen description of effect size.

The term "effect size" has a specific statistical meaning, but in this document we were referring mostly to the magnitude of change. Presentation of the calculation and use of effect size is beyond the scope of this document, so that terminology has been changed.

Each module should have information on testing for statistical outliers

There is already text describing the need to evaluate individual animal data, and to look for obvious outliers. Such outliers could be a result of technical or instrument error, individual sensitivity or resistance, or some other factor. While there are statistical tests for outliers, there is no general agreement on their use in DNT studies. Furthermore, the different types of data generated in these tests are not always amenable to statistical outlier tests. Text has been edited to describe better an evaluation of extreme values, but that the assistance of biostatisticians would be useful for more formalized outlier tests.

Reviewer 3

Many strengths in description and organization, well-written, and useful examples.

No changes needed.

Overall charge questions

Reviewer 1 did not answer these questions

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

Reviewer 2

Add sentences to Introduction to specify the purpose of the document

Text is added to the Introduction to be more specific on the purpose and use of this document.

Reviewer 3

Introduction doesn't specify when the guidelines should be used. Requirements for the different guidelines are summarized but it is not clear when these guidelines apply or who requests them.

Text is added to specify when these different test guidelines might be used.

No clear statement as to why they should be used. Material in specific modules implies the guidelines would apply when tests are submitted for approval, but this is not explicit.

Text is added to clarify when guidance can be used, and by whom.

Locomotor activity is addressed differently in module A and B, with module A being superficial and module B quite detailed, and differences in automated procedural requirements.

> *There is a mention of automated activity measurements in module A, but since this is never conducted as part of clinical observations it is not covered here.*

Module A notes that observation protocols vary widely across laboratories and may not be age-specific. It is not clear at what point the submission is inadequate.

> *The regulator should be aware of these differences across laboratories. Determination of when a submission should be deemed unacceptable involves more considerations than can be addressed in this guidance.*

Purpose of document sometimes unclear: whether it is intended to ratchet up the standards for an improved and more consistent set of practices, or is it intended to capture the current state-of-the-art. It appears suboptimal practices have been tacitly accepted this practice.

> *Text is added to explain that the purpose is to foster better and more consistent reviews of DNT data.*

Reviewer 4

Yes, provides rationale for why and when the guidance should be consulted but details of the guidance are not entirely accurate

> *Inaccurate details listed by this reviewer are addressed in specific modules.*

2. 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?

Reviewer 2

Nothing in the draft would hinder review or interpretation of DNT data

> *No changes needed.*

Reviewer 3

Outliers are described and potential reasons and impacts are addressed, but there is little guidance what to do about them. Suggestions might include reviewing the data records for aberrant or physically impossible values, or running analyses with and without the outlier.

> *As described in the above comment on outliers, there is information on ways to consider extreme values, but the use of statistical outlier analyses on the various test measures goes beyond the scope of this document.*

Statistical approaches are described but do not specify that F-ratios and degrees of freedom should be reported, helpful for QC and a check on appropriate conduct of analysis.

> *While we agree that this information is useful, statistical analyses of DNT data could involve a number of different approaches which do not provide traditional F-ratios and degrees of freedom, e.g., non-parametric tests or mixed-model ANOVAs. On the other hand, p-values are common to all analyses. Thus, requesting more than p-values in reports could be more confusing to the reviewer than helpful.*

Inconsistent treatment of positive controls across modules, with module A providing little information on the requirements, importance, and expectations.

> *Description of positive control information has been strengthened across modules. For module A, text is added to note that, unlike other tests, there are no positive control requirements for clinical observations.*

It is stated that exposure should occur from GD0 [sic] through weaning but there is little guidance about how to determine that such exposure has actually occurred. How can the regulator consider the level of postnatal exposure in the absence of biomarkers of exposure?

> *Exposure actually starts at GD6, not GD0. It is already noted that the reviewer should consider information about biomarkers of exposure, but that those data may not exist.*

The treatment of motor deficits is minimal and mostly ancillary to other tests, potentially missing motor effects. Why not do quantitative gait assessment or rotarod?

> *The test guidelines do not require specific testing of motor deficits such as suggested by the reviewer. Since the regulator can only review the data received, there is no need to cover these tests.*

The auditory system is the only sensory system evaluated directly. The visual system is almost ignored and olfaction and somatosensory systems are completely ignored; this seems surprising.

> *As above, the guidelines do not require these sensory evaluations so the regulator will not have such data to review.*

Reviewer 4

Much of the guidance is good and appropriate, but in some areas it is incomplete, inaccurate, or needs clarification. Also, some views expressed are not consistent current knowledge and need adjustment. Specific points are listed for each module.

> *Inaccurate details listed by this reviewer are addressed in specific modules.*

Module A (Observations)

1. 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?

Reviewer 1
Some areas in this module need more specific details. Regarding the progression of tests from the least to the most invasive (p. A-6), a regulatory scientist may not be able to determine if this was done.

> *Text is added emphasizing that the laboratory's SOP should include detailed test descriptions and order of testing.*

Regarding separation of the pups from the dam ("not for very long"), an actual duration would be helpful.

> *An actual duration of separation is added.*

Reviewer 2
Guidance is mostly sufficient. Additional information or details on statistical evaluation could be helpful.

> *Several options for statistical analyses are added.*

Reviewer 3
These is confusion regarding the terms "clinical observation" and "expanded clinical observation" and whether they referred to the FOB.

> *Text is added to be more explicit about clinical observations compared to the FOB. It was apparently not clear to this reviewer that this module is not about the FOB, which is not required by guidelines.*

The test guidelines are reviewed but elsewhere notes that various procedures are applied regardless of age of the rat or chemical tested. The guidance is ambiguous in how to reconcile requirements with accepting whatever is submitted.

> *This comment is similar to the general comments by this reviewer, and is addressed above. Text is added to note that the various procedures used do generally meet the minimal guideline requirements.*

Positive control guidance is ambiguous, especially in this module. Details regarding test ages, strain, etc, are not specified. The reviewer suggests positive control data that are contemporaneous with the testing.

> *Text is added to better describe the value of positive control data in validating the observer's ability and to note that, unlike the other tests, there are no positive control requirements for clinical observations.*

Section 5.1 calls some observations "useless". There is a question of whether the regulator should ignore these, since such observations can indicate nonspecific illness.

> *The term "useless" referred to guideline terminology, not the observations themselves. Text is added to make this clearer.*

Reviewer 4
Reactivity to handling and placing are sensitive to handling, which is not similar across labs or time.

> *Text is added to note that inconsistency across laboratories makes interpretation of measures such as these difficult.*

Frequency of urination and defecation have declined over decades, and measures with low frequency raises concern about reliability.

> *While the reviewer did not provide data on these measures over decades, text is added to make the reviewer aware that data with low frequencies are difficult to analyze and interpret.*

The reviewer should understand that the ontogeny of righting reflex is very rapid, and that when tested daily, delays should be interpreted cautiously.

> *The clinical observations described in this module do not include evaluation of ontogeny of righting reflex.*

The reviewer should be skeptical of data that does not include high quality positive control data.

> *Text is added to specify that positive control data are not required for clinical observations.*

Greater emphasis should be placed on quantitative data over observational data.

> *Text is added to note greater confidence with quantitative data.*

Data should be interpreted differently if the assessment was after the daily dose compared to before the dose.

> *There is already advice that the reviewer should consider whether assessments were taken before or after daily dosing.*

This reviewer provides recommendations to drop many measures, including handling reactivity, urination, defecation, rearing, and others.

> *The purpose of this guidance is not to recommend changes or deletions of specific procedures, but to help the reviewer deal with the data that are submitted.*

2. 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?

Reviewer 1

This module should have more information on potential influences of maternal toxicity on behavior and test performance.

> *Maternal toxicity altering offspring behavior is specifically addressed in module E. The influence of developmental delays influencing test performance is described in module D. Module A is not an appropriate place for this information.*

Reviewer 2

The Data Interpretation section raises points to consider but doesn't specify how to use that information to decide whether the dose was appropriate.

> *Text is added to note the need to consider toxic effects in both the dam and pup when evaluating offspring behavior. However, whether or not the reviewer decides that the dose was "too high" does not alter interpretation of the effects at that dose.*

Reviewer 3

This module notes that various procedures are applied regardless of age of the rat or chemical tested. There is little guidance for a reviewer who feels the tests may be not suitable, and little guidance on selecting tests.

> *As noted above, the regulator should be aware of differences across laboratories. Determination of when a submission should be deemed unacceptable involves more considerations than can be addressed in this guidance. Furthermore, the reviewer does not select what tests should be used.*

Reviewer 4

No, and reasons have been described above

> *Responses described above.*

3. 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?

Reviewer 1
With the minor edits described above, it will be as complete as it can be.

Responses described above.

Reviewer 2
There could be a few more specific examples.

There are no suggestions or specifications for more examples in this comment.

Reviewer 3
There should be mention of looking for outliers when there is abnormally large SD.

Text is added to recommend evaluations of variable data for outliers and/or data errors.

The statistical analysis section provides appropriate information on main effects and interactions, and implications of large CVs.

There is no mention of main effects, interactions, and large CVs in this module.

Reviewer 4
Yes and no. Examples have been described above.

Responses described above.

Module B (Motor activity)

1. 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?

Reviewer 1
Suggest adding a reference to Figure 1 with details of number, strain and whether they are the same subjects throughout.

> *Text is added to explain that Figure 1 is only an example of the patterns across ages, not data taken from the literature.*

Information on vertical positioning (height) of photocell sensors (to detect rearing) should be required.

> *The importance of knowing the height of photocells is already addressed in section 4.3.*

Reviewer 2
Yes. This module provides useful detailed and specific guidance for the regulatory reviewer, especially on statistical evaluation and habituation.

> *No changes needed.*

Reviewer 3
Reviewer had no comments on this question

Reviewer 4
Figure 1 should indicate SD or SEM, and a description of the test apparatus.

> *Text is added to explain that Figure 1 is only an example of the patterns across ages, not data taken from the literature.*

The reviewer should be aware that video tracking systems are prone to artifacts, such as reflections, which impact the tracking accuracy. The reviewer should request and review video files to check for this. These systems should be discouraged.

> *Text is added to note tracking problems with video systems, and that reviewing video files could be useful. The purpose of this guidance is not to encourage or discourage specific systems, but to help the reviewer deal with the data that are submitted.*

The importance of session length for demonstrating habituation should be emphasized, and 60-min sessions should be encouraged.

> *Text is added to note that sessions should be long enough to show habituation in control rats. The purpose of this guidance is not to recommend or dictate experimental procedures, but to help the reviewer deal with the data that are submitted.*

Specific information is necessary on chemicals used for cleaning equipment since inadequate denaturing would not remove odors.

> *Text is added to specify that cleaning information should include chemicals used. The purpose of this guidance is not to specify chemicals that should be used for cleaning, and in addition denaturing proteins may not necessarily remove odors.*

Analyzing within-subject data across intervals is the preferred way to evaluate habituation.

> *There is already information that analysis across time blocks is necessary to demonstrate habituation.*

Standard errors should be used instead of standard deviations (SDs), and reviewers should be clear that SDs are only descriptive.

> *The guidelines require that means and SDs be listed, and furthermore SEMs are very sensitive to sample size discrepancies. Text is added to mention the use of SEMs.*

There are additional approaches for analyzing sex within litters.

> *Text is added to for additional approaches, including adding a random factor for litter.*

To sort out interactions, slice-effect ANOVAs are better than simple-effect ANOVAs. Dunnett's test is advisable for comparing groups to control. Tukey HSD is not appropriate for more than 3 groups, Tukey-Kramer should be used.

> *The use of Dunnett's test is already noted. Text is added to specify Tukey-Kramer or other tests with a note about controlling for multiple comparisons.*

In section 7.1 age should be a within-subject factor if the same animals are tested at different ages.

> *This is corrected.*

The document assumes weaning at PND21, which may not be the case in some laboratories. If testing is based on weaning day, the data would be quite different from testing on PND21.

> *The guidelines specify PND21 as the testing day, rather than being based on day of weaning. Reference to testing at weaning has been removed.*

The scale of measurement for motor activity is actually counts and therefore not strictly considered continuous data with a normal distribution.

> *Text is added to consider underlying distributions as a source of variability.*

The caption for Figure 2 should say what the data are (means? SD? SEM?)

> *Text is added to explain that the different habituation patterns in Figure 2 are examples and not real data. This figure is not taken from the literature.*

2. 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?

Reviewer 1

Total activity is typically sum of ambulatory and non-ambulatory movements, and does not include rearing.

> *Some laboratories submit rearing as part of total activity. There is already text explaining that the measures comprising total activity should be specified.*

The absolute prohibition against doing an interval-by-interval analysis and description of potential effects are excellent.

> *No changes needed.*

Reviewer 2

Yes, this module includes sufficient guidance for a regulatory scientist.

> *No changes needed.*

Reviewer 3

Table 1 specifies the age of testing +- 2 days, but that is a huge range for 13 day olds.

> *The guidance allows ±2 days for PND60 only*

Reviewer 4

No. The document falls short of providing appropriate guidance for the reasons described above.

> *Responses described above.*

3. 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?

Reviewer 1

In section 6.1, a common dependent variables is average speed, which may be provided by most automated systems.

Average speed is added to the list of common variables.

Figure 1 “shows how the variability decreases with age” but this cannot be seen without information on group sizes.

Text is added to explain that Figure 1 is only an example of the patterns across ages, not data taken from the literature.

Reviewer 2

Yes, this module details the information the reviewer should look for and provides good supportive examples

No changes needed.

In Figure 2 it is unclear whether these patterns are all from doses of a single test material.

Text is added to explain that the different habituation patterns in Figure 2 are examples that could be seen with multiple doses of a single chemical, or with different chemicals. This figure is not taken from the literature.

Reviewer 3

Reviewer made no comment to this question

Reviewer 4

Yes and no. Examples have been described above.

Responses described above.

Module C (Acoustic/auditory startle response)

1. 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?

Reviewer 1

For readability, add the red dashed line to the female graph of Figure 3.

This graph was inadvertently altered in the conversion process and is now corrected.

One important procedural detail missing is that the report must describe the size of the chambers (typically Plexiglas restrainers) used.

The size of test chamber was already listed as an important procedural detail to be reported, and more text is added to describe how the size can impact the measure.

Many laboratories include “blank” trials with no auditory stimulus presented; these are important to verify that the subjects are not “startling” without the auditory stimulus.

Text is added to describe blank trials and the need to include those data in the Data Reporting section.

Knowledge of the testing history for the subjects is essential, especially with auditory startle testing that is obviously stressful.

Testing history is already listing as an important part of Test Procedures.

Figure 2 is unclear, does not describe what solid lines represent.

This graph was inadvertently altered in the conversion process and is now corrected.

On page C-17, the text should say “simple effects” rather than “simple-simple effects”

This is corrected.

It should be noted that the initial trial typically has the highest startle response, and when averaging the first 10 trials this high value contributes to the higher values for the first block.

Text is added to the description of habituation to note the impact of including the high first-trial data in block averages.

Reviewer 2

Yes. This module provides detailed and specific guidance to help the regulatory reviewer identify the critical details regarding testing equipment, procedures, study design, statistical models, and results.

No change needed.

Reviewer 3

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. It should note instead that habituation can be detected statistically by a downward trend in the data across trials.

This section already states that habituation is a decrease in the test measure, and that the statistical result should be due to the downward trend in the data.

There is a long discussion about the influence of body weight on startle amplitude but also states that body weight should not be used as a covariate if it is influenced by exposure. While this is true statistically, failing to control for body weight could result in false negatives or positives.

This reviewer basically agrees with the authors regarding restrictions for doing ANCOVA. Text is added to lessen the absolute restriction but to examine both measures and conduct whichever analysis makes the most sense. The examples provided by the reviewer were already described in the text.

Reviewer 4

Be consistent in the use of the term “acoustic” or “auditory” startle response, and in the use of the abbreviation ASR.

> *Text is added to note that the terms are interchangeable but “auditory” is now used consistently since the guidelines use this terminology. The term “acoustic” is used to describe the stimulus, since literally “acoustic” refers to sound and “auditory” refers to hearing. The abbreviations have been deleted.*

Add more recent review references including Gomez-Nieto et al. under Test Description.

> *The papers cited are reviews; Gomez-Nieto et al. is an experimental paper. We searched but could not find more recent reviews on the underlying biology of the startle reflex and its use in toxicology.*

Revise specifics on prepulse inhibition studies: length of prepulse and how to define interval between signals.

> *The PPI section is not meant to be prescriptive. Text is added that is more general regarding these experimental parameters, with an additional supporting reference.*

This reviewer recommends the inclusion of PPI studies, and testing each animal twice on two consecutive days.

> *The purpose of this guidance is not to recommend or dictate experimental procedures, but to help the reviewer deal with the data that are submitted.*

In the bullets on pg C-7, the document should state that the proper scale is sound pressure level, that white noise is advisable over pure tones, and that the animal holder should be of appropriate size.

> *Text is added to specify details and impact of sound level, white noise delivery, and size of test chamber. It is noted that this information should be reported.*

This reviewer presents an extended discussion of the technical aspects of accelerometer vs load cell detectors. The reviewer notes that both detectors output an electrical signal (voltage) and may be calibrated to convert to other units. The reviewer also believed that the text leaned heavily in favor of load cells over accelerometers, but it should be impartial and in fact, this reviewer prefers accelerometers. The reviewer also took offense to the term “arbitrary units” for the accelerometers. These points are expressed repeatedly throughout the comments.

> *There was no intention to recommend one system over the other, only to inform the regulator about the different ways that the data are collected and reported. Advantages and disadvantages of both systems were already presented, but the text has been reviewed carefully and edited to eliminate unintentional bias. The term “arbitrary units” is removed and the nature of the outputs is better described.*

This reviewer states that 20-40 ms is too short a time frame for the entire startle response, but later in the comments says 20-40 ms is appropriate for the startle.

> *References are added to support the time frame for the full and peak response. Some differences in exact timing are due to the nature of the recording equipment and the interpretation of peaks occurring after the first major peak. These issues are addressed throughout these other comments.*

This reviewer disagreed with statements suggesting that the second peak in an accelerometer output does not represent the startle response. Over several comments, the issue arises as to the interpretation of multiple peaks, and what variable to use to represent the startle response. The reviewer states that averaging these multiple peaks is preferred.

> *There is no clear consensus on this topic, despite the claims of the reviewer. Multiple peaks may be a result of the detecting equipment, the confinement of the animal, and/or a representation of sequential muscle recruitment. Different researchers use either peak response or an integrated response*

measure over a defined window of time. Rather than go into excruciating detail on this topic, we have lessened the emphasis and added reference to a recent paper that addresses this confusion.

The startle signal onset rise time should be specified under "Test Procedures"

This is already listed in the information that should be provided, under "Data Reporting".

There is too much emphasis on latency measures (response or onset latency). The reviewer believes this measure has little scientific value and mentions several times that there should be less emphasis on this as a dependent variable.

Any emphasis on latency was unintentional, and we do agree that changes in latency are less important than amplitude, and sometimes uninterpretable. Throughout the module, latency is now described or listed after amplitude. There is already text that describes the relative value of this measure.

The document specifies both signal rise and times as variables to be reported, as well as frequency of response for the speakers used. However, the reviewer feels that fall time is not important and that speaker specifications are not needed.

Fall time and speaker specifications have been deleted from this section.

Included in data reporting are criteria for exclusions of certain non-response trials. The reviewer agrees that animals do not always respond on every trial but states that the trials should not be removed, and that averaging trials across blocks smooths out trial-to-trial variability.

We do not suggest removing non-response data, and text is added to specify that looking at individual data can be useful to pick up treatment-related increases in the number of these trials.

Table 2 is long and unnecessary, and should be replace by a much simpler table.

Table 2 presents actual terminology that has been used in study reports, and it is included here to help the reviewer understand how various terms may all mean the same thing. This was apparently unclear to reviewer 4, and text is added to be more explicit that these came from actual reports. It is also noted that reviewer 3 stated that Table 2 is very helpful. No changes were made to the table itself.

Fig 2 caption should define the mean response.

Text is added in the Figure caption to indicate mean amplitude.

In one place it is stated that Fig. 2 came from Sette 2004 but elsewhere it says Raffaele 2004, which isn't in the references: are the data from two different sources.

The Sette 2004 reference is a SOT poster (printed in The Toxicologist) which showed the graph, whereas the Raffaele 2008 paper provides individual data for each laboratory in tabular form. Raffaele 2004 was an error and is corrected.

In the "Data Analysis" section, should add that litter can be handled as a random factor

Text is added to for additional approaches, including adding a random factor for litter.

The restriction of not using body weight as a covariate in analysis of startle data may be statistically correct but this may be the only way to do the analyses. This reviewer suggests doing the analyses with ANOVA and ANCOVA, look at how different (or similar) they may be, and acknowledge the limitations of the analyses. It is noted that if there were no effects on body weight, there would be no need to even conduct an ANCOVA.

This reviewer basically agrees with the authors regarding restrictions for doing ANCOVA. Text is added to lessen the absolute restriction but to examine both measures and conduct whichever analysis makes the most sense. The use of body weight without a treatment effect can account for weight differences across animals that may exist but not be related to treatment.

The first sentences in the paragraph presenting Table 3 are confusing and not germane to the points later in the paragraph.

Much of this figure legend is deleted since the same points are already explained in the text.

Table 4 presents data that are not believable. Showing these data and then listing reasons why they might be wrong is akin to setting up a straw man and then knocking it down.

This table presents data that has been seen in submitted datasets. The purpose of this section is to show what key variables to examine, what the conditions are that could have generated those data, and what to ask for to resolve the discrepancy. It is noted that Reviewer #2 stated that this table, and the accompanying text, is very helpful.

The use of specific values for latency should be avoided.

Specific values have either been deleted or changed to reflect a biological plausible range.

2. 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?

Reviewer 1

With the minor edits suggested above, this portion should serve as sufficient guidance.

Revisions described above.

Reviewer 2

Yes, the Acoustic Startle Response module does include sufficient guidance that a regulatory scientist could interpret the data and results without reading the interpretation from the study author.

No change needed.

Reviewer 3

There is no guidance regarding the level of background noise, other than to say that it should be below the level of the stimuli. This is a critical variable and firmer guidance is warranted.

Text is added that more clearly describes the relationship of background noise to startle magnitude, and reference added. Also added is explicit text that the information on the background noise level should be reported.

Table 2 in the startle section is very helpful and it raises an important methodological point regarding Newtons and grams as units of measure. The only appropriate unit of acceleration is g, the acceleration due to gravity.

See below.

Accelerometer outputs that are in arbitrary units raise some question about the reproducibility of the data, especially across systems. Many accelerometers provide data in units of acceleration, and a properly calibrated accelerometer should not require arbitrary units. Units of “volts” are only marginally better than arbitrary units.

See below.

Units of time should be seconds or milliseconds, not “Tmax” or just “mean time,” which confer little information.

These three comments appear to refer to Table 2, which presents examples of the way data have been reported, not how they should be reported. The authors agree with the reviewer’s comments regarding units.

Reviewer 4

No. The document falls short of providing appropriate guidance for the reasons described above.

Responses described above.

3. 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?

Reviewer 1

Some reports may list startle response for the first trial separately since it is typically so much higher than even that exhibited on the second trial. The examples listed in this portion were very helpful.

No changes needed.

Reviewer 2

Yes, the Acoustic Startle Response module does describe in detail the types of information the reviewer should look for in submitted data and provides excellent examples of data to support the interpretation of treatment related changes.

No change needed.

Reviewer 3

The Y-axes in Figure 3 should not say just “ASR Vmax” when both body weight and ASR are shown.

This figure is corrected.

In Section 7.1 should note (again) that main effects should be tested directly only if there are no interactions.

Text is added to emphasize this.

The section on statistical vs biological significance in the startle section was especially thoughtful.

No change needed.

The presence of representative datasets in the startle section is much appreciated. The two sets showing data that are difficult to interpret are helpful, but both examples are from false-positive cases. A negative example would also provide useful guidance.

The authors agree but there were no such examples readily available from the literature.

Reviewer 4

Yes and no. Examples have been described above.

Reponses described above.

Module D (Learning and memory)

1. 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?

Reviewer 1
In general, this module is very specific and detailed and provides useful information. The examples are excellent (but see my comment about Figure 4 below). Only a few minor edits for this module are suggested.

All editorial changes are made. Responses to additional comments described below.

In Figure 1, it would be helpful if a small platform diagram is placed in one arm of each maze with an indication that this is the "goal" or "end" arm.

Because the "goal" arm could be any arm depending on the procedure, a platform diagram would be needed in each arm. We believe this would be confusing to the reader.

Reviewers would likely appreciate an appropriate temperature range for the water maze.

Text is added with information and a reference describing learning at different water temperatures.

It has been shown that rats can use urine trails from a previous subject to navigate in the Morris water maze so it is good practice to gently stir the water between subjects in any water maze.

This text and reference are added.

Not all laboratories place the extra-maze cues on the room walls, some use a circular curtain around the maze.

Text is revised to be less specific, since there are several ways that cues can be placed around the maze.

Figure 2 is blurry in my copy.

Figure 2 has been copied into the document again and should be clearer.

The distance from the water maze wall for the platform must be described, should it should not be too close to the wall.

Text is added regarding distance of the platform from the tank wall.

The inter-trial interval must be constant for animals run in squads, even if the subsequent trials are quicker.

There are already statements that the inter-trial interval must be consistent.

This is some confusion about Figure 4 and the text interpretation. It is not clear that there was little (if any) learning in the high dose PTU group, while learning did occur in the high dose heptachlor group but at a slower rate. Whether there were effects on swim speed should be specified.

The figure legend is now more explicit in describing the types of effects produced by these chemicals (no learning vs slower learning). Neither treatment altered swim speed, and this information is added.

Page D-23 only notes swim speed for the probe trial, but swim speed needs to be detailed for all trials.

Text is added to note that if videotracking is used, swim speed across all trials should be assessed.

I do not see the increased complexity of Path "B" in the Cincinnati maze. If it is anticipated that there will be increased use of the Cincinnati/Biel mazes, it would be helpful to know if data are typically collected via videotracking or by hand.

Text and a recent reference is added to be more explicit in describing the increased complexity of path B. There are not enough submitted studies to know whether laboratories are using video-tracking or observers to collect the data.

Use of the KM method can be recommended for all mazes for which there is a maximum cut-off time for a trial (not just passive avoidance).

The discussion of the KM methods already states it can be used for any tests with cut-off criterion. Only the figure example discusses passive avoidance.

I greatly appreciate inclusion of the statement on p. D-41 "It is also possible to have significant interactions but not have a significant effect on follow-up tests".

No change needed.

Reviewer 2

Yes. The Learning and Memory module provides detailed and specific guidance for the regulatory reviewer. The statistical evaluation procedures are sufficiently described, but there might be a suggestion for the reviewer to seek the help of a statistician if the analyses provided are unclear or inadequate.

Recommendations to consult a statistician if needed are repeated throughout the document.

Reviewer 3

This section is almost completely devoted to the Morris Water Maze, Passive Avoidance and the Biel/Cincinnati Maze. For each the discussion is impressively detailed and thorough but why are other more widely used tests not included, such as RAM and active avoidance? Why the exclusive attention on procedures that emphasize escape and avoidance?

Almost all submitted DNT studies have included the letter mazes, Morris water maze, passive avoidance and/or the Biel/Cincinnati maze, as shown in Table 1. This document focuses only on the tests that the regulator is most likely to see. Text has been added to make this reasoning clearer.

Page D-38, saying that ANOVAs require "continuous" data is too restrictive since, for example, integers, can certainly be analyzed using ANOVAS. In the case of count data, if the numbers are skewed then a transform will be beneficial.

Text is added to note that count data are not continuous but that ANOVAs may still be useful, and that transformations may be useful.

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. Nonparametric tests are usually less sensitive. With Kaplan-Maier estimators and log-rank analyses it should be noted that fairly large differences are needed to be statistically significant.

This document presents only basic information on transformations as well as nonparametric tests, but does not get into the nuances of the different approaches and relative sensitivities. We do not agree that laboratories should automatically transform the data first, since assessment of the raw data is always needed. Statements regarding how large differences should be to achieve significance would include further discussion of sample size, variability, etc, that goes beyond the scope of this document.

Excessive right-censoring at the maximum possible value can be prevented by the appropriate choice of experimental parameters such as trial length. Labs are to be encouraged to make the trial length sufficiently long that relying on KM estimators won't be necessary.

Trial length may be set based on a number of factors that this document does not address. The purpose of this document is not to encourage or discourage aspects of experimental design, but it is to provide the regulator with approaches to interpret the data they receive.

Reviewer 4

Page D-4 should say "(instrumental learning and/or operant conditioning)" instead of "(operant conditioning)".

This is changed.

Page D-4, the statement defining working memory in terms of time is not quite correct. Time is a critical variable in the definition of working memory in people, but not in animals. I'd suggest adding "trial-dependent memory" which is the typical definition.

We feel that the regulator may not understand the suggested addition, and feel that it's more important to keep descriptions more general even though there may be nuances that we do not address.

Page D-5, Table 1, under "Cincinnati Maze" remove the statement that "Intramaze cues detract from its sensitivity". In the third column change "Sequential Learning" to "Sequential/Egocentric Learning"

This is changed.

Page D-7, Table 2 under "Age of testing": while you are quoting from the OECD guideline for the P25 test age, this could be problematic if a laboratory is weaning on P28. Why not say that testing should be 1 to 2 days after weaning and not mention an age?

We cannot change the specifications from the OECD guidelines, which state PND25. Text is added to point out that testing may need to be later if weaning occurs later.

Page D-8, it states that rats have to make a binary decision to escape water but the next sentence refers to these as being either land or water mazes. I suggest taking out "escape the water".

This phrase is removed.

Page D-10, in simple mazes with only 2 choices, guiding animals to the goal can be done but is difficult in practice. This reviewer asked if data like this are really submitted, and then asked numerous questions regarding the exact practices and issues that arise.

It is true that such data have been submitted. There are already statements that detailed methods must be provided.

Page D-13, section describing problems with trials to criterion provides good advice. This reviewer describes a better procedure to determine maximum number of trials based on control performance rather than some arbitrary number decided by the experimenter to fit in a preconceived protocol.

The purpose of this document is not to describe alternative methods, but to provide the regulator with approaches to interpret the data they receive.

Page D-15, you should re-emphasize that these simple mazes (especially swimming versions) are known to be insensitive, and the Agency should discourage their use.

There are already statements regarding the relative insensitivity of these tests. The purpose of this document is not to discourage or encourage the use of specific tests, but it is to provide the regulator with approaches to interpret the data they receive.

Another problem with these mazes that is not mentioned is that rats show stronger turning biases in binary water mazes than in dry or complex water mazes and therefore it is critical to test using the non-preferred side. This reviewer provides several recommendations for correcting errors.

There is already a statement that the non-preferred side should be used for 2-arm mazes; however, comparisons to land-based or other water mazes is beyond the scope of this document. Furthermore, the purpose of this document is not to describe alternative methods, but to provide the regulator with approaches to interpret the data they receive.

Page D-15 the clause "or in tests involving a single trial . . ." should be removed since no MWM procedures use a single trial.

That sentence was not specific for the MWM and is deleted.

Page D-16, it is not necessary to make the water opaque. There are several different ways that will camouflage the platform so that it cannot be seen.

Text is added to stress that the platform should not be seen and that there are several methods by which this may be accomplished.

Page D-17, referring to guiding animals to the platform if they reach the time limit: There is no agreement concerning the best method of guiding, and no specific method is preferred.

Text is revised to be more general that the rat is guided in some manner or placed on the platform.

Figure 4, right panel: The caption needs to explain the gap in the abscissa.

This information is added.

Figure 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).

The caption has been edited to be more explicit about what the data are.

Page D-19, "Visual Performance Control" section: On cued trials, both the position of the platform and the start must be moved randomly on EVERY trial to prevent route learning.

Text is added specifying that platform and start point must be varied every trial.

Page D-21, second bullet point: This should be made more general, i.e., that the platform must be camouflaged.

This bullet is revised as suggested.

Page D-23, "Dependent Variables", including both trial and day in the ANOVA models may result in complexity, including a large number of F-tests, that doesn't help understand the spatial learning. Data should be analyzed with only day as the repeated measures factor.

There is already a statement that averaging trials across days may be appropriate, and the focus of the statistical analysis information (later in the module) is on using repeated measures, be it trials or days.

Page D-23, "Dependent Variables": Consider adding an analysis of group differences, including a separate analysis of the day-1 data only. A preexisting performance deficit that may not indicate a learning problem especially if ANOVA shows a main effect in the absence of an interaction between Group and Day.

A bullet is added in section 3.4 to include evaluation of pre-existing group differences as a result important to interpreting MWM data; however, this is not a dependent, or measured, variable. This is also already listed as a bullet in section 3.5.

Page D-24, bullet point 4: this is the point I made above.

See above response.

Page D-29, bullet point 5: should add that a test for shock sensitivity must be performed if there are group differences in latency.

Text is modified to be more specific for this point.

Page D-30, "Initial Acquisition", there ares several critical details need to be mentioned. First, testing rats in a straight channel before the CWM maze is an absolute requirement to prevent giving up during training. Also, it should be stated that the CWM is not to be used with mice. Second, it is critical that at least 5 min of rest is given before the next trial of the day to prevent fatigue and compromised problem-solving ability. Third, there will be an occasional animal that stops searching and swims in one spot until the time limit is reached. Such a rat's low error score will underestimate the true deficit. Using a corrected score that may be based on control error rats is suggested. Treatment may cause many trial failures and low errors because the animal stays in a small area of the maze; in such cases, close analysis of corrected and uncorrected errors is necessary.

Text is added to further explain the need for the straight channel training and sufficient resting time. Since rats are specified in the DNT guidelines, there is no need to mention mice in any context. Text is added under section 5.2 to describe the possibly of rats stopping their search. While reporting this *information is critical, developing correction procedures to replace the values is beyond the scope of this document.*

Page D-32, section 5.3, bullet 3, it is worth pointing out that a valid approach is testing only path B.

This bullet is modified to mention that testing could use one or more paths.

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

Text is added to include this.

Page D-38, section 8.3: count data are not continuous and underlying distributions are rarely normal. Should add 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.

Text is added to specify that count data are not continuous and may not have the same distributions.

Page D-38, 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. It should be mentioned non-parametric methods are less powerful compared with parametric methods.

Text is added to include homogeneity of variance as a consideration in statistical analysis, but there is already text stating that ANOVA is robust for some deviations from its assumptions. Non-parametric methods are not always less "powerful" compared with parametric methods.

Page D-40, where it refers to Type 2 errors: it should say Type I errors.

Correction made.

2. 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?

Reviewer 1

This was a well-written module. With the minor edits suggested above, it should provide valuable guidance.

Responses described above.

Reviewer 2

Yes, the Learning and Memory module includes sufficient guidance that a regulatory scientist could interpret the data and results without reading the interpretation from the study author.

No changes needed.

Reviewer 3

In the opening why does it say that learning and memory are theoretical "constructs," which are hypothetical, unobservable entities used in theories to explain learning, but this is immaterial to this document. Learning and memory are intertwined but they are clearly measurable and observable.

Text is added to modify these few sentences, which are intended only as a brief introduction.

(Page D-4) The definition of operant conditioning is incorrect: should be the process by which consequences influence behavior.

This section has been revised and excludes presentation of operant conditioning.

Section 2 and 2.1, it is stated that food deprivation is stressful, but properly done it is not only not stressful but healthy. This reviewer felt that food deprivation stands in contrast to the stress of water- or shock-based tasks, and in addition the consequences of an error in food-based tasks are less, not more, stressful.

Text is added to modify statements that food deprivation is stressful, and the section on error consequences has been eliminated since it is not germane to a discussion focused on water-based tasks.

Some investigators use a heating pad to help the animal recover core temperature. This should be placed around half of the chamber so the animal can thermoregulate by moving around.

Text is added to mention the use of a heating pad.

The document notes the use of a straight maze to estimate swim speed. If video tracking is used then having them swim through a straight alley won't be necessary.

Text is added to note that swim speed should be collected on every trial if possible. However, many researchers feel that an initial swim in a straight alley is still needed to familiarize the animal with the water and the means of escape.

Page D-15, in the discussion of incorporating swim speed in data interpretation, this should be noted in land-based mazes as well. Also, other variables such as error-to-criterion are dependent on speed and may lead to inaccurate interpretations. An "errors per opportunity" measure might address this issue.

This module explicitly excludes discussion of land-based mazes. The influence of speed on the dependent variables is already discussed in length, and examining errors per opportunity is already mentioned in section 2.3.

Page D-15. Instead of "tests should include some measure of retention", it should say "some measure of retention, typically 24 hours later."

This section has been revised and no longer contains that text.

Page D-16, the placement and color of distal cues should take into consideration the poor and monochromatic vision of rodents.

Text is added to address the distal cues in size and color.

Figure 3 is helpful but the text is too small and the reproduction is poor.

The important feature in Figure 3 is the tracking, not the words. The original figure could not be located so the resolution could not be improved.

Reviewer 4

No. The document falls short of providing appropriate guidance for the reasons described above.

Responses described above.

3. 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?

Reviewer 1

Most examples were excellent (see note about Figure 4 above).

Response described above.

Reviewer 2

Yes, the Learning and Memory module describes in detail the types of information the reviewer should look for in submitted data and provides very good examples of testing methods and study data.

No changes needed.

Reviewer 3

Assessment of trials procedures requires information on errors of commission and omission, not simply errors.

Text is added that both types of errors are necessary dependent variables.

The discussion of “nonlearners” is thoughtful and examples informative. The number of non-learners should be reported, and if this number is dose-related it should be taken into account in interpreting the data. An alternative approach is curve-fitting, in which learning would be expressed as a monotonically increasing curve and the magnitude of learning would be the upper asymptote.

> *Text is added to be explicit that the number of learners/non-learners per group should be reported. We are unclear what data should be used for fitting a curve, since learners/non-learners is a dichotomous value; no changes were made.*

Page D-14, it is noted that medians or modes should be used instead of means. This reviewer offered a number of suggestions regarding transformation or other manipulations of the data to normalize the distributions and stabilize the variance.

> *We feel such statistical techniques are beyond the scope of this document and would be more confusing than informative to the regulator. In such cases the regulator would need to consult a statistician for assistance.*

For position discrimination studies it should be reported whether a correction procedure is used.

> *Text is added that such experimental details should be provided to the regulator.*

The right panel of Figure 4 is said to come from a published report of heptachlor. However, the doses and the figure do not patch the published paper.

> *The data in Figure 4 are mistakenly attributed to the heptachlor report, but it is actually from a published tebuconazole paper. This error has been corrected.*

The right panel of Figure 4 looks more like an example of a motor effect since the high dose group was consistently slower. The caption says there was no significant interaction with day, also suggesting a motor deficit.

> *The caption in Figure 4 has been expanded to better describe the effect of tebuconazole, including the finding that there were no group differences on the very first trial but that looking at daily blocks it was evident that learning was slower. The finding of no differences in swim speed was also added.*

The incorporation of reversal learning and working memory tasks into the MWM testing is well-advised.

> *A brief description of these procedures is already included. However, the purpose of this document is not to discourage or encourage the use of specific tests, but to provide the regulator with approaches to interpret the data they receive.*

Page D-22, last two bullets, stable performance and control procedures are necessary, not ideals.

> *The word “Ideally” is removed from these bullets.*

Page D-26. In passive avoidance section: the document appropriately specifies that a 1 mA shock should be delivered but should also note that specifying shock in units of voltage is inappropriate.

> *The text states that 1 mA is typical in the literature, but does not give that as a recommendation. There is no statement about specifying shock in voltage.*

Page D. 37, fourth bullet should include mention that when repeated testing is performed the subject should always experience the same maze.

> *Text is added to include this.*

The specification of water temperature is vague. What temperature range should be used?

> *Text is added with information and a reference describing learning at different water temperatures.*

When data are collected by individual observation and there is any subjective component then some measure of interobserver reliability should be provided.

Text is added to include this.

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?

Reasonable variability can be very different depending on the measure, and this section of the module is not intended to give such a detailed break-down of values.

Reviewer 4

Yes and no. Examples have been described above.

Reponses described above.

Module E (Data considerations and integration)

Reviewer 1 answered the questions that were asked about the other modules, rather than the question specific to this module.

1. 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?

Reviewer 1

P. E-10-11 raises the issue of changes in exposure levels to the dam over gestation and lactation. The document should specify that where dietary exposure (food or water) is used, those data should be included in the report.

> *Text is added to note that actual dose may be easy to calculate from food and/or fluid intake and should be reported.*

2. 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?

Reviewer 1

The information on P. E-12 regarding the translating time website should specific it only compares on the basis of neurogenesis.

> *Text is added to clarify that "translating time" is based on neurogenesis.*

3. 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?

Reviewer 1

This module clearly presents the issues that must be kept in mind when evaluating DNT data.

> *No changes needed.*

WOE question

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?

Reviewer 2

Section 3.1.3, Determining Biological Significance seems inconsistent with the sections on biological significance in Modules B and C.

> *These sections are not inconsistent, they are different. The Biological Significance section in module E describes integrating data across the different endpoints, whereas the sections in the individual modules discuss the issue of biological vs statistical significance for that specific measure. However, some of the text in module C that was less specific for the startle data and has been moved to this module.*

Section 5, on Human relevance seems of little use in evaluation of the DNT data.

> *The human relevance section is brief but does address considerations in extrapolating behavioral effects. We feel that this section is useful and have not changed it.*

The last bullet in the conclusion mentions that the reviewer should take into account information from others studies that may not be readily available.

> *Text is added that these other information sources may not always be available.*

Reviewer 3

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.

> *No changes needed.*

I would add the tacit assumption that if no effect is detected then it is assumed that the chemical does not present a hazard, and that moving forward with commercialization of the chemical is acceptable.

> *The authors do not agree with any statement that negative findings are evidence of no effect, and furthermore, the reviewer's statement is a policy call that does not belong in this document.*

In section 6 there is a distinction between continuous data and rating scales. The issue of continuous data was addressed earlier.

> *There is no mention of continuous or other data types in section 6.*

Section 6 on alternative DNT testing seems out of place here.

> *This short section was intended to make the regulator aware that other tests (in vitro, alternative species) are becoming more popular. However, this does not aid in DNT data evaluation so it has been deleted.*

Reviewer 4

Page E-6, bullet point 3: rather than "stupor" use a more standard pharmacological term: "sedation."

> *This change is made.*

Page E-7, section 3.1.4, bullet point 2: should add that another way is to use litter as a random factor in the ANOVA.

> *Text is added to address using litter as a random factor in analyses.*

Page E-11, second bullet point: should add the gut-blood barrier, which also matures over weeks and determines basic uptake.

> *The gut-blood barrier is but one of many physiological systems influencing basic absorption: there is no reason to specifically address it.*