

# **Up-and-Down Procedure (UDP)**

## **Peer Panel Report**

**July 25, 2000 Meeting**



## 1.0 INTRODUCTION

This report summarizes the results of the July 25, 2000 independent scientific peer review panel evaluation of the revised Up-and-Down Procedure (UDP), a method proposed as a substitute for the existing LD50 test for assessing the acute oral toxicity potential of chemicals. The meeting was organized by the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) and the National Toxicology Program (NTP) Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM), and sponsored by the National Institute of Environmental Health Science (NIEHS) and the NTP. The Peer Review Panel evaluated the usefulness of the UDP as an alternative to the conventional LD50 test method for acute oral toxicity currently accepted by regulatory authorities. *Federal Register* notices relevant to the meeting include a Request for Data and Nomination of Expert Scientists (NIEHS, 2000a) and Notice of Peer Review Meeting and Request for Comments (NIEHS, 2000b). These notices are provided in **Appendix D**.

This introduction briefly summarizes the purpose and history of acute toxicity testing and the purpose and conduct of the July 25, 2000 meeting. The remaining parts of this section summarize the UDP Peer Panel's discussions, conclusions, and recommendations from the July 25, 2000 meeting. A report on a follow-up meeting of the peer review panel on August 17, 2001 is provided in Section II. **Appendix A** provides ICCVAM Test Method Recommendations on the UDP, **Appendix B** contains the Final Revised U.S. EPA UDP Test Guideline which addresses the recommendations from both Panel, **Appendix C** contains the materials reviewed by the Panel for the August 2001 Peer Panel Meeting, and **Appendix E** provides Summary Minutes and Public Comments from the UDP meetings. **Appendix F** provides the Background Review Document on the UDP which has been revised to incorporate many of the recommendations and suggestions from the Panel at the July 2000 meeting. **Appendices G** through **P** provide additional background information about the UDP Primary Test, Limit Test, and Supplemental Test which was reviewed by the Panel in preparation

for their July 2000 meeting. **Appendix Q** summarizes the relevant U.S. Federal Regulations on Acute Oral Toxicity.

### 1.1 History and Purpose of Acute Toxicity Testing

Acute oral toxicity testing is conducted to determine the hazard potential of a single oral exposure to various chemicals and products. The acute oral toxicity test in rodents is a critical step in defining the toxicity of a test material for the purpose of hazard classification and labeling. It is designed to determine adverse effects and to estimate the dose that is expected to kill 50% of the test population (i.e., the LD50).

Four regulatory agencies in the United States, the Department of Transportation (DOT), the Consumer Product Safety Commission (CPSC), the Occupational Safety and Health Administration (OSHA), and the U.S. Environmental Protection Agency (EPA) require industry to label chemicals and products with hazard information based on LD50 estimates. DOT requires oral lethality data to determine the transportation requirements for hazardous substances (49 CFR 173). CPSC requires such information for labeling hazardous substances so as to protect consumers when such products are used in the home, the school, and recreational facilities (16 CFR 1500). OSHA requires the use of acute lethality data to implement labeling requirements for the hazard communication program to protect employees (29 CFR 1910). Certain U.S. EPA regulatory programs also require the submission or generation of acute toxicity data for hazard classification purposes (40 CFR 156). During acute toxicity testing, non-lethal endpoints may also be evaluated to identify potential target organ toxicity, toxicokinetic parameters, and/or dose-response relationships.

As shown in Table 1, the international community also uses acute oral toxicity data as the basis for hazard classification and the labeling of chemicals for their manufacture, transport, and use (OECD, 1998b; updated OECD, 2001). Other potential uses for acute toxicity testing data include:

- Establishing dosing levels for repeated-dose toxicity studies;
- Generating information on the specific organs affected;
- Providing information related to the mode of toxic action;
- Aiding in the diagnosis and treatment of toxic reactions;
- Providing information for comparison of toxicity and dose response among substances in a specific chemical or product class;
- Aiding in the standardization of biological products;
- Aiding in judging the consequences of single, high accidental exposures in the workplace, home, or from accidental release;
- Serving as a standard for evaluating alternatives to animal tests.

Table 1.1 Adapted from the Harmonised Integrated Classification System for Human Health and Environmental Hazards of Chemical Substances and Mixtures: Acute toxicity hazard categories and (approximate) LD50/LC50 values defining the respective categories (OECD 1998b; updated OECD, 2001)

Acute Toxicity Route	Toxicity Class 1	Toxicity Class 2	Toxicity Class 3	Toxicity Class 4	Toxicity Class 5
Oral LD50 Values (mg/kg) [approximate]	≤5	>5 ≤50	>50 ≤300	>300 ≤2000	>2000 ≤5000

Historically, lethality has been the primary toxicological endpoint in acute toxicity tests. Trevan (1927) was the first to attempt to standardize a method for assessing the toxicity of potent biological toxicants, the progenitor of the "lethal dose, 50% (LD50) test". The classical LD50 test procedure evolving from this innovation in the 1970s and early 1980s used from 100 to 200 animals per test substance (Galson, 2000). Although other information, such as the slope of the dose-response curve, confidence interval for the LD50, and toxic signs, could also be obtained from this test, the procedure was severely criticized for both scientific and animal welfare reasons (Zbinden and Flury-Roversi, 1981). These criticisms eventually resulted in the proposal and adoption of a new guideline (OECD TG 401; OECD, 1987) which utilized three dose groups of five rats of one sex, with confirmation in the other sex using one group of five rats. In the absence of a range-finding study, this revision reduced the minimum number of animals used in

the traditional acute oral toxicity test from 30 to 20. This method has become the most widely used for defining the acute toxicity of a chemical and a mandatory-testing requirement for new chemicals.

More recently, the acute toxicity test procedure has been modified in various ways to refine and further reduce the number of animals used to a maximum of 16 (e.g., OECD Test Guidelines 420, 423, and 425). The Globally Harmonised Scheme for Hazard Classification (OECD 1998b; updated OECD, 2001) prompted a re-assessment of all of the OECD *in vivo* test guidelines for acute toxicity (i.e., fixed dose, up-and-down procedure, acute toxic class method) to ensure that regulatory needs are met while minimizing animal usage and maximizing data quality.

Several other test designs, including the moving average (Weil, 1983), acute toxic class method (Schlede et al., 1994), and UDP (Bruce, 1985),

have been proposed. The classical experimental method for estimating the LD50 was to orally dose individual animals, in groups of five or ten per sex, with varying concentrations of the test material and to observe whether the animal lived or died over a defined period of time (generally 14 days). The method was standardized in 1981 by the international acceptance of Test Guideline (TG) 401 (OECD, 1981).

The test material is typically administered by oral gavage to fasted young adult animals. The animals are observed periodically during the first 24 hours with special attention given to the first four hours, then at least once a day for 14 days or until they recover. Clinical signs, including time of onset, duration, severity, and reversibility of toxic manifestations, are recorded at each observation period. Body weights are determined pre-treatment, weekly thereafter, and at the death of the animals or termination of the study. All surviving animals are humanely killed at 14 days or after recovery. Gross necropsies are conducted on all study animals. Variation in the results due to inter-animal variability, intra- and inter-laboratory variability, and to differences in strain, sex, estrus cycle, and species have been characterized. Based on intra- and inter-laboratory testing, the point estimate of the LD50 appears to be reliable within a factor of two or three (Griffith, 1964; Weil et al., 1966; 1967).

Although the experimental method as to dosing, handling, and observing the animals has not varied, many attempts have been made to reduce the number of animals used while maintaining the accuracy of the method for estimating the LD50. These changes in sampling technique do not involve a change in the actual treatment of the animals or in the endpoints examined.

### 1.2 Objectives of the July 25, 2000 Meeting

The meeting was convened to conduct an independent scientific peer review evaluation of the validation status of the revised UDP. This procedure is an updated version of the OECD Test Guideline 425 (OECD, 1998a). The revised UDP

was proposed as a substitute for the existing OECD Test Guideline 401 (OECD, 1987). OECD has proposed that Guideline 401 should be deleted since three alternative methods are now available. Prior to deletion of Guideline 401, U.S. agencies requested that ICCVAM conduct an independent peer review of the revised UDP to determine the validity of the method as a substitute for Guideline 401. The Independent Peer Review Panel was to (1) evaluate the extent to which established validation and acceptance criteria (ICCVAM, 1997) have been addressed, and (2) to provide conclusions and recommendations regarding the usefulness and limitations of the method as a substitute for the traditional acute oral toxicity test method (OECD, 1987). The UDP has the potential to reduce the number of animals required to classify chemicals for acute oral toxicity compared to Guideline 401.

### 1.3 Conduct of the Meeting and Reports

The UDP Peer Panel Review Meeting, which was open to the public, was conducted on July 25, 2000. The meeting began with an introduction including an overview of the peer panel review process and a summary of current Federal agency requirements. The Panel then discussed the Revised UDP Protocol, Primary Test, Limit Test, and Supplemental Test. Following the final public comment session, the Panel provided conclusions and adjourned. Following the meeting the Panel prepared this written report summarizing their discussions, conclusions, and recommendations.

In this Panel report, all references made to the background review document (BRD) refer to the April 2000 BRD which can be found at <http://iccvam.niehs.nih.gov/methods/udpdocs/AllBRDlk.pdf>. The April 2000 BRD was revised in response to recommendations of the Panel and this revised version has been provided in **Appendix F**. When possible, both the former (April 2000) and the current reference (October 2001) for appendices and other documentation have been provided.

## 2.0 GENERAL CONSIDERATIONS

A laboratory-based, practical viewpoint was taken in evaluating the U.S. EPA Revised UDP Guideline (April 2000; formerly **Appendix C**, currently **Appendix G**). Consideration was given as to whether the procedures were described unambiguously, were workable in the laboratory setting, and comprised a sound basis for obtaining the necessary acute oral toxicity information without undue increases in time and expense.

### 2.1 Revised UDP Protocol

The type of information on the test material that should be obtained and considered prior to conducting a study is appropriately described. In general, guidance concerning the selection of the appropriate species, strain, and age of animal for testing is sufficient and appropriate. However, the revised Guideline contains an impractical reference to assigning littermates randomly to test groups. At animal receipt, the laboratory does not know which animals are littermates. In addition, since the total number of animals that will be used during a study cannot be predicted, at least fifteen animals must be assigned prior to study start. Because animal use is sequential, the study design itself minimizes bias.

Unless information is available indicating that one sex is more sensitive than the other, the use of either all males or all females should be considered to allow for additional flexibility and to decrease the total number of animals that are purpose-bred for acute oral toxicity testing. Data provided in the Background Review Document (BRD) (formerly EPA Document 14, Part A, Table 1, currently **Appendix P-1**, Table 1 on page P-6) suggest, in general, a low incidence of studies with a sex-related effect. However, gender-dependent differences in xenobiotic metabolism are more pronounced in rats when compared to other rodent species. The differences primarily involve cytochrome P450s (CYP), sulfotransferases, glutathione transferases, and glucuronyl transferases (Mulder, 1986; Nelson et al., 1996). Studies of chemicals with known sex-related differences in toxicity, attributable to differences in metabolism, have shown that females are often more susceptible when

compared to males (see former U.S. EPA Document 14 in the BRD, currently **Appendix P**).

Descriptions of the accepted weight range and procedures for minimizing weight variation during the test procedure are not adequate. The age and weight ranges are not specified in the April 2000 revised Guideline (formerly **Appendix C**, currently **Appendix G**) as they are in OPPTS 870.1100, which requires rats to be between eight and 12 weeks of age at the time of dosing. In addition, individual body weights recorded on the day of dosing must be within 20% of the mean body weight for all animals dosed during the study. Similar guidance is recommended in the revised Guideline.

Guidance regarding procedures for preparing animals for study and the description of dose preparation procedures is sufficient and appropriate. Guidance regarding dose administration, including dose volumes and stability considerations (e.g., the need for appropriate stability data if a single dosing solution is used over several days) should be further refined in the U.S. EPA Revised UDP Guideline (formerly **Appendix C**, currently **Appendix G**). The use of constant concentration (in addition to constant volume) should be included as an option for at least some types of test materials. OPPTS 870.1100 requires liquids to be administered neat or at the most concentrated workable dilution, if dilution of a liquid or suspension of a solid is needed. This issue may be important in particular when testing at the limit dose (i.e., 2000 or 5000 mg/kg) to simulate accidental exposure to the undiluted product.

The notion that the test material concentration in dosing solutions might need to be supported by analytical analysis is especially burdensome, as it would greatly increase the cost. The use of constant volume dosing solutions instead of constant concentration solutions would potentially increase the analytical task and is not recommended. The cost of analytical analysis may impact the willingness of some laboratories to use the revised UDP. OPPTS does not require analytical evaluation. If it is suspected that the test material is unstable in solution, a fresh

mixture should be prepared prior to each administration. The absence of a concurrent vehicle control is justified sufficiently.

Paragraph 27 of the Revised Guideline (formerly **Appendix C**, currently **Appendix G**) provides an adequate description of appropriate observations to be recorded. The reference to Chan and Hayes (Chapter 16. Acute Toxicity and Eye Irritancy. *Principles and Methods of Toxicology*. Third Edition. A.W. Hayes, Editor. Raven Press, Ltd., New York, USA, 1994) should be removed. It may be more appropriate to include specific references in a guidance document. The first two sections of paragraphs 26 and 27 of the revised Guideline (April 2000) are repetitive and contradictory. We recommend replacement of the first sentence in paragraph 26 with the first sentence of paragraph 27. Each time the 48-hour observation interval is mentioned, as in “each animal should be observed carefully for 48 hours (unless the animal dies)”, the qualifier “but need not be rigidly fixed” should be added as delayed mortality will occur often. Also, “time of death” should be worded as “time found dead” as it is unlikely the exact time of death will be determined, unless a moribund kill has been conducted.

Appropriate endpoint(s) for humanely killing animals prior to the end of the required holding period are sufficiently and appropriately described. Frequency of body weight measurements and procedures for pathology evaluations are described appropriately.

The description of the data to be collected and reported is largely standard guideline wording and is sufficient as such. A specific rationale for the starting dose and dose progression should be provided only when it varies from the standard described in the revised UDP Guideline (formerly **Appendix C**, currently **Appendix G**), and removal of the requirement for justification of starting dose and dose progression when the defaults are used is suggested. However, one Panel member suggested that a rationale be provided for all starting doses and dose progressions even when the default is used. It would be helpful if a table of log doses from 0.1

log to 0.5 log was provided, starting at 10 mg/kg and progressing to 5000 mg/kg.

Procedures for recording and storing data, including suggested forms or formats, are described sufficiently. Descriptions of equipment, materials, and supplies needed are appropriate. However, a comprehensive, validated software package should be developed and distributed to assist in conducting all variations of the UDP protocol. Ideally, a series of data sets (testing program) should be provided for the purpose of “in-house” validation for compliance with Good Laboratory Practice (GLP) guidelines.

## 2.2 Animal Welfare Considerations (Refinement, Reduction, Replacement)

With regard to the Revised UDP Guideline (formerly **Appendix C**, currently **Appendix G**), the majority of the Panel concluded that the validation studies and simulations appear to have demonstrated that the number of animals necessary for the revised UDP Primary Test (i.e., between six and 15) and the revised UDP Limit Test (between three and five) are appropriate to obtain scientifically valid results. However, some Panel members were concerned that the optimal numbers of animals for each test had not been adequately demonstrated.

The majority of the Panel concluded that the procedures in the revised UDP addressed the potential for pain and distress issues based on the inclusion of the OECD Guidance Document on the Recognition, Assessment, and Use of Clinical Signs as Humane Endpoints for Experimental Animals used in Safety Evaluation (OECD, 2000a; formerly **Appendix B**, but no longer appended in this final report). However, the Panel concluded that only limited or no improvement was made in the area of replacement, especially for the UDP Supplemental Test. The Panel felt that additional information would be needed to adequately evaluate the UDP Supplemental Test.

The rationale for the necessity to use animals to determine acute oral toxicity is appropriate and justified, although there is an implication that the reason for not testing in humans is a legal issue rather than a moral one. The revised UDP

Guideline (formerly **Appendix C**, currently **Appendix G**) states that the primary reason for conducting animal tests is for the protection of humans from the consequences of exposure to unsafe products. However, product testing also benefits wildlife, domesticated animal, and pets.

### 2.3 Other Considerations

The procedures for the observation and reporting of clinical signs are appropriate and adequate for regulatory needs. However, the procedures for considering delayed deaths need clarification.

Based on the revised Guideline and the supporting documentation, the proposed test methods can be readily conducted in GLP-compliant laboratories. The procedures take more time and are more cumbersome than OECD TG 401 (formerly **Appendix A**, currently **Appendix I**) or OPPTS 870.1100. Explanation of the statistics in the revised UDP Primary Test and the UDP Supplemental Test accompanied by illustrative examples (perhaps in the form of flow charts in an appendix to the April 2000 Guideline) will be critical for the non-statistician to conduct these studies. As mentioned previously, a comprehensive, validated software package should be made available to assist with these calculations.

A reordering of the presentation of the three different types of studies in the revised UDP Guideline (formerly **Appendix C**, currently **Appendix G**) is recommended. The revised UDP Limit Test should be described first. Additional guidance should be included to provide for a transition from the revised UDP Limit Test to the revised UDP Primary Test, when necessary.

Personnel training and experience requirements are adequately described and reasonable. The necessary equipment, materials, and supplies (e.g., animals, and computers) should be readily obtainable.

The estimated cost of an UDP study provided in the April 2000 BRD is not realistic. The cost of conducting the revised UDP Primary Test will be greater than the traditional acute toxicity test, perhaps up to twice as much, due to the needs for

increased technical expertise, specialized statistical analysis, as well as to the difficulty associated with scheduling (animal shipments, dose preparation, dosing, necropsy) and organizing the data for reporting. For example, the challenge of scheduling multiple simultaneous UDP Primary Tests is much greater than that associated with the scheduling of the same number of OECD TG 401 (formerly **Appendix A**, currently **Appendix I**) tests. Ensuring that adequate numbers of animals in the appropriate weight range are readily available will be more difficult than would be for the traditional LD50 test. Laboratories that infrequently conduct the UDP test may be forced to humanely kill a greater number of undosed animals. As a consequence, particularly for smaller companies with limited resources, the difference in product testing costs could be significant.

Depending on study progression, it is likely that the revised UDP Primary Test will take significantly more time than the traditional acute toxicity test. Realistically speaking, it is difficult to dose more than two animals per week unless one of the treated animals dies on treatment day. If dose levels are started close to the LD50, animals generally take two to three days to show morbidity/mortality. Therefore, the revised UDP Primary Test will most likely take at least three weeks if the minimal number of animals (i.e., 6) is used and seven to eight weeks if the maximum number of animals (i.e., 15) is used. Although not recommended by the Panel, addition of the UDP Supplemental Test would increase the total duration of the study by an additional two to five weeks per test material. In contrast, the traditional acute toxicity test using three dose levels generally takes four to five weeks and yields a similar amount of information.

In reference to the revised Guideline (formerly **Appendix C**, currently **Appendix G**), the outcome of the UDP Primary Test is likely to be sensitive to differences in dose selection and progression as well as to the statistical procedures employed. This revised UDP Primary Test protocol has now become even more complicated than the current UDP (OECD, 1998; former **Appendix A**, current **Appendix H**) and the results are probably very sensitive to errors in dose level

selection. The more complicated the protocol, the more extensive the measures that must be taken to minimize the likelihood of errors in the laboratory.

in conducting all variations of the UDP protocol. Ideally, a series of data sets (testing program) should be provided for the purpose of “in-house” validation for compliance with GLP guidelines.

## **2.4 Recommendations**

1. The U.S. EPA Revised UDP Guideline (formerly **Appendix C**, currently **Appendix G**) should be re-ordered to present the revised UDP Limit Test first since this test is more likely to be used for the majority of test materials.
2. Additional guidance on the transition from the revised UDP Limit Test to the revised UDP Primary Test, when appropriate, should be provided in the revised Guideline.
3. All reference to littermates should be excluded from the revised UDP Guideline (April 2000; formerly **Appendix C**, currently **Appendix G**).
4. The use of either sex (all males or all females) in a study should be allowed unless information is available suggesting that one sex is more sensitive.
5. The use of animals of 8 to 12 weeks of age at the time of dosing should be specified in the revised Guideline.
6. The revised Guideline should state that individual animal body weights on the day of dosing must be within 20% of the mean body weight for all animals dosed.
7. The option for constant concentration in addition to constant volume solutions should be included in the revised Guideline.
8. In the U.S. EPA Revised UDP Guideline (formerly **Appendix C**, currently **Appendix G**), the Chan and Hayes (1994) reference and the first sentence in paragraph 26 should be deleted. Paragraph 27 provides an adequate description of the clinical observations to be conducted. In addition, the qualifier of “but need not be rigidly fixed” should be added to “48 hours”.
9. A table of log doses from 0.1 log to 0.5 log, starting at 10 mg/kg and progressing to 5000 mg/kg, should be included in the revised Guideline.
10. A comprehensive, validated software package should be developed and distributed to assist

### 3.0 REVISED UDP PRIMARY TEST

#### 3.1 Introduction and Rationale for the Revised UDP Primary Test

##### 3.1.1 *Scientific Basis for the UDP Primary Test*

Inadequate information on the **scientific** basis of the revised UDP Primary Test (e.g., what information is needed about acute toxicity, how the test results would be used) was provided in the U.S. EPA Revised UDP Guideline (formerly **Appendix C**, currently **Appendix G**) and in the April 2000 BRD. The **technical** basis for the revised UDP Primary Test is described in detail; however, the description is not completely understandable and requires clarification. Paragraph 10 of revised UDP Guideline [Principle of the Primary (Single Estimate) Test] and the corresponding Section 1.2 the April 2000 BRD (The Scientific Basis of Revised UDP) appear to discuss different issues; paragraph 10 provides a synopsis of the test method while Section 1.2 provides information about the philosophy behind the procedure. Consequently, it is difficult to reconcile the information provided in these two sections. Nonetheless, the technical basis for the revised UDP Primary Test is, for the most part, adequately described. The literature reference on page C25 of the April 2000 BRD is incomplete; for reference number 14, the date is 1994.

##### 3.1.2 *Intended Uses of the Revised UDP Primary Test*

In the revised Guideline (formerly **Appendix C**, currently **Appendix G**), the rationale for the revised UDP Primary Test is clearly presented. By concentrating testing around the LD50, the UDP requires fewer animals per study than OECD TG 401 (formerly **Appendix A**, currently **Appendix I**). Should the starting dose be far from the LD50, a bias may be introduced. This bias is true particularly for test materials with a shallow slope for the dose-response curve; in addition, the bias is reduced relative to OECD TG 425 (formerly **Appendix A**, currently **Appendix H**) by the increased progression factor between consecutive doses. It is stated that the revised UDP will replace the current regulations on acute

oral toxicity testing for the Consumer Product Safety Commission (CPSC), the U.S. EPA, and the U.S. Department of Transportation (DOT). However, it appears that both the U.S. EPA and the U.S. DOT already use this revised UDP Primary Test and that only the CPSC will be adopting this protocol as a new procedure. The justification provided is that the use of the revised UDP Primary Test will enhance the ability of the CPSC to use data for risk assessment purposes and for probabilistic modeling; information is not provided about the scientific basis of the test.

If the observations of animals administered a low dose demonstrate a no-observed-adverse-effect-level (NOAEL), these data may be used to estimate an acute reference dose when considering residues of highly toxic pesticides in foods. It appears that the revised UDP Primary Test (April 2000) provides a better estimate of the LD50 for classification when compared to OECD TG 401 (formerly **Appendix A**, currently **Appendix I**). A summary table comparing simulation results for the April 2000 revised UDP Primary Test with OECD TG 401 in a format similar to that on former page C-401, current page O-13 of the BRD would be helpful.

Neither the revised Guideline, the April 2000 BRD, nor the oral presentation at the July 2000 Panel meeting provided sufficient information for evaluation of how the revised UDP Primary Test will be integrated into the U.S. EPA's strategy for assessing the hazard or safety of materials. The types of materials that are amenable to the test have been delineated. The test is designed for materials that can be administered neat (without dilution) or in a solvent. The test is not restricted to materials that are water-soluble. Any solvent or vehicle can be used, but the solvent or vehicle must not add to or mask the toxicity of the test material. Although the proposal did not specifically address biopesticides, there should be little concern about testing these materials with the revised UDP Primary Test procedure. The revised Guideline stated that the LD50s of materials with shallow slopes are underestimated.

The Panel had two concerns regarding the 25 test materials used to validate the revised UDP (Bruce, 1987, Bonnyns et al., 1988, Yam et al., 1991).

First, in the Bruce (1987) validation study, eight of the 10 test materials were proprietary. As a consequence, their chemical class is unknown and some members of the Panel expressed doubt as to whether these data should have been considered for validation. Second, as each of the 25 test materials was tested in a single laboratory only, no assessment of interlaboratory reproducibility was possible. However, with the exception of mercury chloride, there was excellent concordance in the estimated LD50 between OECD TG 401 (formerly **Appendix A**, currently **Appendix I**) and the current UDP (formerly **Appendix A**, currently **Appendix H**).

### 3.2 Revised UDP Primary Test Protocol

A statement is made in the U.S. EPA Revised UDP Guideline (formerly **Appendix C**, currently **Appendix G**) that all information on the material to be tested should be considered. However, no details were provided about the nature of the information to be obtained or how such information should be considered. Thus, prior to study start, a general description of the information (e.g., *in vitro* data, physicochemical properties, etc.) for consideration should be provided; in addition, how such information should be used to predict the need for the study and/or the starting dose should be determined [for example, Spielmann et al., (1999) provides information that could be useful].

A precise description of what is meant by the “slope” of the dose-response curve should be included in the Guideline. Also, in paragraph 18 of the revised Guideline (formerly **Appendix C**, currently **Appendix G**), the sentence stating, “however, when justified by specific regulatory needs, testing up to 5000 mg/kg body weight may be considered” needs to be clarified (i.e., when is it a requirement, and if not, what would justify testing at the higher limit dose?). In the revised Guideline, a “similar” dose progression should be reworded to the “same” dose progression. The April 2000 BRD (Section 1.1.5) states that the default starting dose of 175 mg/kg was chosen based on historical data and the results of computer simulations; further justification of this starting dose is needed.

The revised Guideline should include a more comprehensive description of the information needed to select an appropriate value for the slope, of when to use the default dose progression factor, and of the methods to be used in the final analysis. Because the dose progression factor can have a large effect on bias if chosen inappropriately, it should be stated that a value other than the default should be used only if there is clear evidence that the slope of the dose-response curve is far from a value of two.

The term “half-log spacing” is more accurate than a dose spacing factor of 3.2. It should be defined and used consistently throughout. The use of half-log units appears to lead to a reasonable estimate of the LD50, although no direct comparisons with other possible values were found in the simulation study results. The relatively large value reduces the bias when the starting dose is far from the true LD50 because the testing dose approaches the LD50 rapidly. This spacing allows one to reach 2000 or 5000 mg/kg with considerably fewer animals than the original 1.2 progression factor. The disadvantage is that when testing does occur near the LD50, the final estimate of the LD50 is less precise due to the larger dose spacing. An extreme example is for materials with steep slopes (above about 4); in such studies, dose levels often exhibit 100% mortality or 100% survival. The estimated LD50 is known only to occur between the lowest fatal dose and the highest non-fatal dose. This type of data occurs also in the methods described in OECD TG 420 and OECD TG 423 (formerly **Appendix A**, but not included in this final report), which do not provide an estimate of the LD50. Any estimate of the LD50 resulting from the UDP depends on the choice of the assumed dose-response curve slope. A similar situation arises when both death and survival occur at a single dose level only. It would be interesting to know how often this finding was observed in the simulations.

In the revised Guideline and in the April 2000 BRD, the description of stopping rule #3 is not provided in sufficient detail and some aspects are confusing and/or scattered throughout the documents. The information could be consolidated and clarified. Terms like “the

number of animals after the first reversal” should be more clearly defined. A single software package allowing implementation of all three stopping rules should be developed and evaluated in an *in vivo* practicability study.

Computer simulation results show clearly that using the revised stopping decision criterion reduces the effect of an outlier on the estimate of the LD50 relative to the estimate obtained using OECD TG 425 (formerly **Appendix A**, currently **Appendix H**). There does not appear to be any specific evidence regarding reliability, though the reliability of the U.S. EPA Revised UDP Guideline (formerly **Appendix C**, currently **Appendix G**) would likely be comparable to OECD TG 401 and OECD TG 425 (formerly **Appendix A**, currently **Appendices I and H**, respectively). The Guideline should be modified to allow estimation of the LD50 by any suitable statistical method (e.g., isotonic regression).

### 3.3 Performance of the Revised UDP Primary Test

#### 3.3.1 Characterization of Materials Tested

Given that this test represents a modification of OECD TG 425 (formerly **Appendix A**, currently **Appendix H**) only, simulation studies seem to be an appropriate method of assessment. The simulation studies include materials with a full range of LD50 and slope values. However, the range of dose-response slopes is not clearly discussed in Sections 3 or 6 of the April 2000 BRD.

#### 3.3.2 Performance of the Revised UDP Primary Test

The conclusions on the usefulness of the April 2000 revised UDP Primary Test are appropriate based on computer simulations. Since no formal *in vivo* validation has been reported for the revised UDP Primary Test, at a minimum, a practicability evaluation of the revised test should be conducted. The performance of the revised UDP Primary Test has been adequately described. The revised UDP Primary Test better predicts the LD50 when compared to the traditional acute toxicity test method (OECD TG 401; formerly **Appendix A**,

currently **Appendix I**). However, although the revised test method uses fewer animals, the study duration in most cases will be longer. Costs for the revised UDP Primary Test and OECD TG 401 (formerly **Appendix A**, currently **Appendix I**) are reported in the April 2000 BRD to be similar, but in reality appear to be greater.

With regard to the revised UDP Guideline (formerly **Appendix C**, currently **Appendix G**), the primary limitation of the revised UDP Primary Test is the poor estimation of the LD50 for test materials with shallow slopes for mortality. This limitation is common to all of the proposed test methods. Since only a small number of chemicals have been evaluated in the current UDP (formerly **Appendix A**, currently **Appendix H**), the extent of this limitation cannot be defined with any degree of assurance. However, according to the April 2000 BRD, it is stated that any class of chemicals or products that can be tested using OECD TG 401 (formerly **Appendix A**, currently **Appendix I**) can be tested using the revised UDP. The April 2000 BRD further states that this test method is designed for materials that can be administered neat or in a solvent. The test method is not restricted to materials that are water-soluble; any solvent or vehicle can be used as long as the solvent or vehicle does not add to or mask the toxicity of the test material. These are logical statements, but insufficient data are available to support these assertions.

#### 3.4 Reliability (Intra-laboratory Repeatability; Intra- and Inter-laboratory Reproducibility) of the Revised UDP Primary Test

In the revised UDP Guideline (formerly **Appendix C**, currently **Appendix G**), the estimated intra- and inter-laboratory reliability of the revised UDP Primary Test appears to be acceptable and better than that for OECD TG 401 (formerly **Appendix A**, currently **Appendix I**). Although the reliability is likely to be very similar to that for OECD TG 425 (1998) and even for OECD TG 401 (1987), Section 7 of the April 2000 BRD states “there are no known *in vivo* data on the reliability and repeatability of the revised UDP.” In the limited testing that has been conducted, the UDP has been shown to perform

well when compared to OECD TG 401. A number of the test materials evaluated in the Bruce study (1987) were unidentified and only a small number of materials were examined in the Bonnyns et al. (1988) and Yam et al. (1991) studies, with no single material tested in more than one laboratory. Additional computer simulations should be conducted to assess the effect of changing response probabilities with the age and weight of the animals at the time of treatment.

### 3.5 Summary Conclusions

With regard to the revised Guideline, the revised UDP Primary Test is a suitable replacement for OECD TG 401 (formerly **Appendix A**, currently **Appendix I**). Most information obtained with OECD TG 401 is also obtained with the revised UDP Primary Test (e.g. classification, point estimate, acute toxicity characteristics). There is substantial reduction in the number of animals required, but no or little improvement in the areas of refinement or replacement.

It appears that the revised UDP Primary Test provides a better estimate of the LD50 for classification and the potential for better overall information on acute toxicity with fewer animals when compared to OECD TG 401.

### 3.6 Recommendations

1. The scientific basis for the test should be enhanced and added to the April 2000 Guideline, with greater explanation in the April 2000 BRD.
2. The revised Guideline should include a description of how historical data should be used to decide when to use the UDP Primary Test, the UDP Limit Test, or not to conduct any test.
3. Justification should be provided in the revised Guideline as to why the recommended starting dose of 175 mg/kg (in the absence of any relevant information) should be used.
4. In the Guideline, stopping rule #3 should be clearly defined and justified.
5. A single software package covering the entire procedure and including all three stopping rules should be developed.
6. In the U.S EPA revised Guideline, stopping rule #1 of the UDP Primary Test and the UDP Limit Test should be harmonized.
7. In the Guideline, the term “half-log” units should be used throughout rather than the approximate dose progression factor of 3.2.
8. A table of computer simulations comparing the revised UDP Primary Test with OECD TG 401 (formerly **Appendix A**, currently **Appendix I**) should be included in the BRD (e.g., see the table on page O-13 of **Appendix O-2** (former page C-401) comparing the original UDP with OECD TG 401). The simulations should include an assessment of the effect of changing response probabilities with the age and weight of the animals at the time of treatment.
9. Since no formal *in vivo* validation has been reported for the revised UDP Primary Test, at a minimum, a practicability evaluation of the revised test should be conducted.
10. The April 2000 BRD should include a separate section discussing how reduction, refinement, and replacement (i.e., the 3 R's) are addressed by the revised UDP Primary Test.
11. In the U.S. EPA Revised UDP Guideline, the overall usefulness of information (e.g., clinical signs, time course of effects, target organs, pathology, etc.) gained beyond the LD50 in the revised UDP Primary Test should be emphasized.
12. It is recommended that either sex can be used unless information suggests one sex is more sensitive.
13. The term “slope” should be defined in the April 2000 Guideline and BRD.
14. The revised Guideline should state that any suitable statistical LD50 estimate method (e.g., isotonic regression) might be used.

## 4.0 REVISED UDP LIMIT TEST

### 4.1 Introduction and Rationale for the Revised UDP Limit Test

With regard to the U.S. EPA Revised UDP Guideline (formerly **Appendix C**, currently **Appendix G**), the scientific basis for the revised UDP Limit Test is not adequately described in either the Guideline or the April 2000 BRD. A brief description of how to conduct the UDP Limit Test is provided, but no explanation of the scientific basis or the rationale for the revised test is reported. A scientific basis would explain why the proposed approach produces valid estimates and would provide a description of the advantages of the revised UDP Limit Test over other methods. The scientific basis should be added to the revised Guideline, with greater explanation in the BRD.

The rationale for the revised UDP Limit Test as a substitute test method for existing regulatory acute toxicity limit test methods, such as OECD TG 401 (formerly **Appendix A**, currently **Appendix I**), is not adequately described. It would be helpful to explain why the revised UDP Limit Test is a suitable replacement of the Limit Test in OECD TG 401. The rationale should describe the conclusions that could be made using the revised UDP Limit Test. The primary conclusion of the revised UDP Limit Test is that the LD50 is either above or below the limit dose used in the test. The discussion in the April 2000 BRD describes the potential uses of the revised UDP Primary Test, but not the revised UDP Limit Test. Consequently, additional discussion of the functionality of the revised UDP Limit Test in the strategy of hazard or safety assessment would significantly strengthen the revised Guideline. A flow chart with decision criteria for the entire testing scheme might be an efficient way to characterize this relationship. A chart would help also to place the revised UDP Limit Test in perspective to other tests as well as explain its relationship to the revised UDP Primary Test and any supplemental tests.

### 4.2 Revised UDP Limit Test Procedure

In the U.S. EPA Revised UDP Guideline (formerly **Appendix C**, currently **Appendix G**), the procedures for conducting the revised UDP Limit Test merit further clarification. Specifically, further explanation is needed in the Guideline regarding the scientific basis, the selection of the limit dose, the stopping rule, how the revised UDP Limit Test is integrated into the revised UDP Primary Test, and factors that may set the two tests apart. These Guideline clarifications would improve the usability of the test and reduce confusion in its implementation.

While the scientific basis and rationale for the revised UDP Limit Test should be stated in the April 2000 BRD, a short statement, similar to that for the revised UDP Primary Test, would also be helpful in the revised UDP Limit Test Guideline. The revised Guideline would be improved if a short rationale such as the following were added: “Principle of the Limit Test: When it is necessary to determine if (or confirm) that the LD50 is above a defined limit (2000 or 5000 mg/kg), the UDP Limit Test may be performed.” This or a similar statement would help explain the general purpose of the revised UDP Limit Test.

Clarification of the selection of the limit dose would be helpful in the April 2000 Guideline and BRD. The description of the revised UDP Limit Test specifies a limit dose of 2000 mg/kg with the option of using 5000 mg/kg. This option reflects the difference between European and U.S. testing. However, this difference is not discussed in the Guideline or the BRD and inclusion of such information would be helpful. Further, the Guideline and BRD state “dosing should not normally exceed 2000 mg/kg body weight.” This statement could be interpreted in several different ways and requires greater clarity. The BRD implies that 2000 mg/kg is the standard limit dose, but in some cases 5000 mg/kg may be used. However, one section of the April 2000 BRD (Section 6.3.3.2) differs from the other sections in that it mentions a lower testable dose. Discussions indicated that in some circumstances the limit dose could be less than 2000 mg/kg. The Panel is concerned that tests with lower limit doses may be inappropriate and may confuse

standardization of guidelines. The rationale for conducting a test at a limit dose lower than 2000 mg/kg should be clearly explained in the BRD.

The stopping rules are explained in the April 2000 Guideline (Paragraph 23) and in the April 2000 BRD (Section 2.1.4). The basic stopping rule in the revised UDP Limit Test is the occurrence of two additional survivors or three deaths following survival of the first animal. This rule differs from the stopping rule that would be applied when reaching the upper bounding limit during the revised UDP Primary Test, which requires that three consecutive animals survive. The two different stopping rules may cause confusion. This issue needs to be clarified in the Guideline and the BRD.

With regard to the revised Guideline, guidance was not provided as to the next action to take when the test does not demonstrate that the LD50 is above the limit dose tested. The Guideline should state clearly that, depending on the pretest question, testing either stops or the revised UDP Primary Test should be conducted. Furthermore, in Limit Test studies in which three animals fail to survive, it should be stated explicitly that the results do not provide **any** scientifically relevant information about the actual value for the LD50. Integration of the revised UDP Limit Test into the testing strategy would clarify how the testing should be approached. As recommended previously, the revised UDP Limit Test section should precede the revised UDP Primary Test section.

The April 2000 revised UDP Limit Test, which allows the conclusion that the LD50 is greater than the limit dose if three animals, including the first, survive, is much less stringent than OECD TG 425 (in which six consecutive animals, three of each sex, must survive), but slightly more stringent than OECD TG 401 (in which at least five of ten animals must survive). In the BRD, the probability calculations (formerly EPA Document 7, **Appendix C**; currently, **Appendix M**) show that the performance of the proposed sequential method is very similar to that of a method where the number of animals tested is fixed (e.g., OECD TG 401 Limit Test; formerly **Appendix A**, currently **Appendix I**). However, the reduction in

sample size results in an increased probability of misclassification for materials with an LD50 above the limit dose, especially when the LD50 is close to the limit dose. More discussion in the April 2000 BRD regarding the relative performance of alternative methods would be helpful.

**Appendix M** of the BRD (page M-5, item 2, second sentence; formerly EPA Document 7 in **Appendix C**) appears to make an incorrect statement regarding the stopping rule. This Appendix discusses the stopping rule and suggests that “n,” the number of animals, is always odd. The number of animals tested can be even (i.e., four) and may occur in three of the 11 possible testing sequences. The expression  $(n+1)/2$  is equal to 2.5 for those sequences with four animals tested. Therefore, statements involving the expression  $(n+1)/2$  are not always correct and require clarification.

The dosing section of the U.S. EPA Revised UDP Guideline (formerly **Appendix C**, currently **Appendix G**) requires clarification regarding the actual procedure to be followed. The currently proposed procedure, described in the revised Guideline Section 23, line 5, states “if [the first] animal survives, two more animals are dosed sequentially at the limit dose.” Since the Guideline requires that two more animals be tested regardless of outcome, the word “sequentially” should be deleted. Also regarding the revised Guideline, paragraph 23, line 6 states “if one or both of these two animals die, two animals are dosed sequentially at the limit dose....” However, conditions for stopping the test may be met after only one additional animal is tested. Therefore, the sentence should read, “if one or both of these two animals die, additional animals are dosed sequentially at the limit dose....” These two changes would help clarify the revised Guideline. This confusion can also be found in Appendix II, Paragraph 12 of the April 2000 Guideline, where the statement “then dose an additional two animals” is made; this statement is not always true and should be corrected. This type of statement is also mentioned in the April 2000 BRD (Section 2, 2.1.4, first paragraph). In the description of the testing scenarios in the April 2000 Guideline Appendix II, Paragraph 13, the

sequence S DD DX (in the most recent revision, O X XXU) is duplicated. There are only four sequences for this test that can end in death. Also, the parenthetical expressions can be eliminated because U would not occur in these sequences. All five of these sequences end with an S (or O in the most recent revision). Finally, in the April 2000 BRD (**Appendix C**, Tab 7, page C-184, first paragraph, third sentence), it is stated that the animals could be dosed sequentially or all at one time. The revised Guideline calls for dosing the animals sequentially--one at a time. This statement should be corrected. Consequently, the April 2000 Guideline and BRD provide a confusing and possibly contradictory description of dosing and should be corrected.

Due to the lack of clarity in the U.S. EPA Revised UDP Guideline (formerly **Appendix C**, currently **Appendix G**), there appears to be a difference between the revised UDP Primary Test and the revised UDP Limit Test in the time of observation after dosing. The revised UDP Primary Test requires that the LD50 calculation be based on all reported deaths up to 14 days after dosing. The revised UDP Limit Test Guideline implies that decisions are based on all reported deaths that occur within two days. This discordance should be clarified by discussing the observation procedure as a general procedure in the revised Guideline. Currently, the observation period is only discussed in the paragraphs describing the revised UDP Primary Test.

While some features of the revised UDP Limit Test set it apart from the revised UDP Primary Test, most of the procedural steps for the two tests are similar. Consideration should be given to reorganizing the revised Guideline to improve clarity in a manner that indicates what features of the Guideline apply to both tests (e.g., test material preparation, dosing procedure, observation period, the intended range of materials amenable to the test, and testing of biopesticides). The April 2000 Guideline (Paragraph 17, page C-18) and the April 2000 BRD (Section 2.1.2.1, second sentence) do not provide adequate information regarding consideration of other acute toxicity data prior to conducting the test. However, this deficiency is common to all acute toxicity tests. Factors that

pertain only to the revised UDP Limit Test should be clearly demarcated in the Limit Dose section of the revised Guideline. The Guideline should also state how to determine that a Limit Test and not the Primary Test is required.

### 4.3 Performance of the Revised UDP Limit Test

Information in the April 2000 BRD (such as in Sections 6.1, 6.3, and 6.5) was not helpful in determining if the revised UDP Limit Test adequately predicts whether the LD50 is above or below the limit dose. The only information identified for this task in the BRD was found formerly in EPA Document 7 in **Appendix C**, currently **Appendix M**. The performance of the revised UDP Limit Test was not tested with *in vivo* data, only with probability calculations. Based on the calculations, the procedure seems to work well and the performance characteristics may be adequate. However, it is not readily apparent how the revised UDP Limit Test was derived from these analyses. It would be helpful if the calculations were performed in a manner that allowed a clear comparison of the revised UDP Limit Test to the Limit Test described in OECD TG 401 (formerly **Appendix A**, currently **Appendix I**); instead, the calculations address the general issue of fixed versus sequential dosing.

The probability study (formerly in EPA Document 7 in **Appendix C**, currently **Appendix M**) begins with certain assumptions to be used for calculations. For example, the evaluation assumed that for all the animals tested there is the existence of a definable probit dose-response curve with a known LD50. However, if substantial variability exists in the animals during the study (e.g., in weight and age changes), there may not be a definable single slope. Weil et al. (1966) states that one of the more significant causes of laboratory-to-laboratory variability in estimates of the LD50 is the weight of the animals used. Because the April 2000 revised UDP Limit Test is a sequential procedure, the first animal tested will be younger and smaller than the last animal tested. There are no specific criteria given as to how wide the time span from the first to last animal tested can be for the test to remain valid. The primary concern is that the calculations

utilize a constant probability of death for a given level of exposure regardless of when that exposure occurs. This assumption is probably unrealistic given the sequential nature of the test and real life environmental factors that occur and can alter the probability of response during the conduct of the study.

With regard to the U.S. EPA Revised UDP Guideline (formerly **Appendix C**, currently **Appendix G**), the Panel has several concerns regarding the accuracy of the revised UDP Limit Test and the ability of the test to minimize the use of animals. As indicated in the former **Appendix C**, Document 7, Table 3, current Table 3 in **Appendix M** on Page M-9, the probability of misclassification of a 5000 mg/kg UDP Limit Test for a sigma of 0.5 is 2% if the true LD50 is 1500. If the slope is more shallow, for example with a sigma of 2, the probability of misclassification of a 5000 mg/kg UDP Limit Test is increased such that a 21% misclassification occurs if the true LD50 is above 3000 mg/kg. Thus, there is concern about the accuracy of the revised UDP Limit Test, particularly for materials with shallow slopes for mortality. The table should be recalculated to provide the estimates for doses that represent the general Hazard Classes (i.e., 5 mg/kg, 50 mg/kg, 300 mg/kg, 2000 mg/kg, and 5000 mg/kg). This table would allow the reader to understand the chance of misclassifying various classes of toxic materials as non-toxic. Furthermore, similar comparisons using OECD TG 401 (formerly **Appendix A**, currently **Appendix I**) would clarify the strength of both tests. Additionally, the calculation that results in doses above 5000 mg/kg merits clarification in the April 2000 BRD.

The value of the revised UDP Limit Test would be improved if additional calculations were conducted regarding the probability for correct classification using other decision criteria. For example, assume failure of the revised UDP Limit Test when 1) any animal death occurs out of up to three tested, or 2) death of the first animal or death of two of five animals. These criteria may also yield a reduction in the number of animals tested. Consequently, additional calculations, similar to those in the revised BRD Table 3 in **Appendix M** on Page M-9, should be completed

to determine if the expected number of animals tested is reduced.

The question of the need for additional calculations is discussed above. The April 2000 documentation did not provide *in vivo* studies to characterize the performance of the revised UDP Limit Test. It is laudable that probability calculations were used in an effort to help design a test procedure that would use fewer animals. However, it is not clear if the revised UDP Limit Test can be accepted in the absence of *in vivo* studies. Possibly, studies designed to test the practicability of the procedure, as was suggested for the revised UDP Primary Test, are needed.

The range of toxicity of the chemicals/products used to estimate the performance of the revised UDP Limit Test should be extended. The results from existing animal tests suggest it would probably help to have additional calculations using shallower slopes. It might be helpful to add results that would occur for LD50 values of 10000 and 20000 mg/kg. The additional information should provide a clearer picture of what occurs when materials with a fairly high LD50 are tested using this protocol. It would seem that materials with high LD50 values are those that would most likely be tested with the revised UDP Limit Test.

The April 2000 BRD (Section 2.5) describes the adequacy of results based on the explanation that a single experiment has been considered sufficient in the past. In general, this reasoning is not a scientifically sound justification for using only a single UDP Limit Test. The adequacy of a single experiment is not a major factor that needs to be considered since the purpose of the UDP Limit Test is to provide the same information as past testing while reducing animal use.

#### **4.4 Reliability (Intra-laboratory Repeatability, Inter-laboratory Reproducibility) of the Revised UDP Limit Test**

*In vivo* acute lethality data were not considered in the evaluation of the reliability of the revised UDP Limit Test. The only available data are based on probability calculations shown in the revised BRD Table 3 in **Appendix M**, Page M-9 of the BRD.

The problems associated with this approach are discussed above.

With regard to the revised UDP Guideline, the only scientific basis for the revised UDP Limit Test is the probability calculations. Much of the April 2000 BRD documentation does not appear to apply to the revised UDP Limit Test. Extrapolating from studies used to estimate the LD50, it appears that the revised Guideline must be specific in all aspects of study design in order to ensure adequate LD50 reproducibility. The Guideline may not be sufficiently specific to ensure reproducibility. Factors such as the age and weight of the animals used appear to be very important to ensuring adequate reproducibility, but these factors are not rigorously specified in the revised Guideline. The specific determination of whether an animal is moribund and should be humanely killed can vary from investigator to investigator. Because no more than five animals will be used, an error in a single observation can have a major influence on outcome. Only *in vivo* studies appear able to address these issues.

#### 4.5 Summary Conclusions

With regard to the U.S. EPA Revised UDP Guideline (formerly **Appendix C**, currently **Appendix G**), the Panel members reviewing the revised UDP Limit Test concluded that the test has been evaluated sufficiently. Its performance is satisfactory to support its adoption as a substitute for the Limit Test described in OECD TG 401 (formerly **Appendix A**, currently **Appendix I**) for oral acute toxicity. However, there are qualifications regarding the accuracy and reliability of the Limit Test. The revised UDP Limit Test is expected to perform as well as or better than the Limit Test in OECD TG 401, with a reduction in the number of animals. Regarding animal welfare, the Panel members also discussed whether the revised UDP Limit Test adequately considered and incorporated where scientifically feasible, procedures that refine, reduce, and/or replace animal use. The revised UDP Limit Test does not replace animal use. It was not clear to these Panel members if the procedure refined animal use, in terms of reducing pain and suffering. However, the majority of these Panel members concluded that the procedure reduced

animal usage, particularly in comparison to the Limit Test in OECD TG 401.

The Panel members noted deficiencies in the description of the revised UDP Limit Test in the April 2000 Guideline and BRD. The scientific basis for the revised UDP Limit Test is not adequately described in either document. There was no rationale provided for the method. Little justification for the UDP Limit Test is provided in the BRD, particularly regarding the starting dose (i.e., 2000 or 5000 mg/kg). The overall product was inadequately organized for review of the revised UDP Limit Test. The revised UDP Limit Test Guideline was not well written and the organization of the current document made it difficult to locate the relevant sections to address the questions in the Evaluation Guidance. The relationship of the revised UDP Limit Test to the revised UDP Primary Test is unclear in the April 2000 BRD. The probability calculations and presented data were insufficient to determine the accuracy for correct classification at shallow slopes. Other limitations of the revised UDP Limit Test are also present in the revised UDP Primary Test and in acute toxicity testing in general.

#### 4.6 Recommendations

1. The scientific basis of the revised UDP Limit Test should be included in the U.S. EPA Revised UDP Guideline (formerly **Appendix C**, currently **Appendix G**), with greater explanation in the April 2000 BRD.
2. Additional discussion in the revised Guideline of the applicability of the UDP Limit Test in hazard or safety assessment would significantly strengthen the test. A decision criteria flow chart describing the complete testing scheme might be an efficient way to achieve this goal.
3. The revised Guideline would be improved if a short rationale for the UDP Limit Test were added in a separate paragraph.
4. The revised Guideline as currently written is difficult to follow. Consideration should be given to reorganizing the Guideline to improve clarity.

5. The use of constant volume or constant concentration of the test material should be allowed.
6. In the Guideline, all reference to littermates should be excluded.
7. Animals of 8 to 12 weeks of age at the time of dosing should be used.
8. The individual animal body weights on the day of dosing must be within 20% of the mean body weight for all animals dosed.
9. Clarification of the selection of the limit dose would be helpful in the April 2000 Guideline and BRD.
10. The current organization of the BRD made adequate document evaluation difficult. Movement of some material in former **Appendix C**, Tab 7 (current **Appendix M**) to the main section of the BRD would improve the organization and address many issues of concern. Furthermore, clarification of several details in the Guideline or the BRD would improve the understanding of the test.
11. Additional calculations to justify the benefits of the revised UDP Limit Test would be helpful. The document should provide probability estimates for accuracy using criteria that compare the revised UDP Limit Test to OECD TG 401 (formerly **Appendix A**, currently **Appendix I**) to clearly delineate the benefits. The document should provide probability estimates for accuracy using more stringent criteria to determine if a further reduction in the number of animals tested is possible.
12. Table 3 in former **Appendix C**, Document 7 (current **Appendix M** on Page M-9) should be recalculated to provide dose estimates that represent the general Hazard Classes (i.e., 5 mg/kg, 50 mg/kg, 300 mg/kg, 2000 mg/kg, and 5000 mg/kg). It might be helpful to add results that would occur for LD50 values of 10000 and 20000 mg/kg.
13. The value of the revised UDP Limit Test would be improved if additional calculations were conducted regarding the probability for correct classification using other decision criteria.
14. The basic stopping rule in the revised UDP Limit Test is the occurrence of two additional survivors or three deaths following survival of the first animal. This rule differs from the stopping rule applied when reaching the upper bounding limit during the revised UDP Primary Test, which requires that three consecutive animals must survive. The two different stopping rules may cause confusion and additional explanation in the BRD is suggested to address this issue.

## 5.0 UDP SUPPLEMENTAL TEST TO ESTIMATE SLOPE AND CONFIDENCE INTERVALS

### 5.1 Introduction and Rationale for the UDP Supplemental Test

While there are several reasons why some estimate of the slope for the dose-response curve may be needed, none were articulated in the BRD. Slope information is, for example, useful in selecting doses for subsequent longer-term studies. However, determination of an exact slope is rarely necessary.

One exception is that the U.S. EPA has a legal requirement to perform wildlife risk assessments for acute toxicity. Within the 29 countries of the OECD, this exception appears to be the only regulatory requirement for a rodent acute toxicity test that generates the slope of the dose-response curve as well as an LD50 value. It is uncertain what proportion of all acute toxicity tests will be required by the U.S. EPA to provide a slope value. Will it only apply to new pesticide active ingredients or will such information also be needed for all new formulations being registered for use? Is the inclusion of the UDP Supplemental Test in the revised OECD TG 425 justified? Far fewer animals would be killed if information on slope were requested through the conduct of a non-guideline study. A non-guideline study could utilize any scientifically relevant test method, as agreed upon by the registrant and the Agency. The revised OECD TG 425 would then contain only the acceptable UDP Primary and Limit Tests and would allow the OECD to proceed with the deletion of OECD TG 401 (formerly **Appendix A**, currently **Appendix I**) and approval of a method that further reduces animal use for acute toxicity testing.

The scientific basis for the proposed UDP Supplemental Test is not adequately described or even addressed. Why and when such data would be needed is not defined. The justification for the UDP Supplemental Test presented in the BRD is discussed in statistical terms stating that the UDP proposed by Dixon and Moods (1948) centers trials around the LD50 value. This method is

appropriate for estimating the LD50, but it is not a good means of estimating the 'slope' in the probit model. The fit of the UDP Supplemental Test into a strategy for hazard or safety assessment is not adequately discussed. The lack of a description of the utility of this test in hazard assessment was a significant omission.

The BRD makes the point that more animals are needed for the generation of sound data for determining slope and confidence intervals (CI) for LD50s. This requirement is a fundamental problem with the proposed UDP Supplemental Test—too few data points. This issue makes it very questionable that the proposed UDP Supplemental Test would meet published regulatory acceptance criterion that “the method should be suitable for international acceptance.” To increase the number of animals used per test, without demonstrated and necessary improvements in precision, would not be consistent with the regulatory acceptance criterion that “the method must provide adequate consideration for the reduction, refinement, and replacement of animal use.” Compared to OECD TG 401 (formerly **Appendix A**, currently **Appendix I**), the proposed UDP Supplemental Test meets the criterion for reduction in that it provides better quality information from fewer animals.

Virtually no information was provided that would allow a determination on whether the intended range of materials, based on chemical class or physico-chemical factors, was appropriate. As noted in the Summary Conclusions, the number of agents tested, the number of chemical classes evaluated, and the range of effects expected are far fewer than what would be needed to adequately address this question. Additional background information is needed to properly evaluate any new procedure proposed to generate slope and CI information in addition to the LD50 value.

The slope is said to be equal to  $1/\sigma$  (in one place the BRD says proportional to  $1/\sigma$ ), but is never directly defined. What is  $1/\sigma$  the slope of? The definition of slope should be clearly provided in the Guideline and in the BRD upon the first mention of slope. The slope of a

probit curve is a different value at each point on the curve.

What scientific questions are being asked where the "slope" is required for determining the answers? Information of this type in the BRD is too vague. For example, in U.S. EPA Document 1, page 9, it states that, "Some authorities also use test results to perform various risk assessment functions, including determination of confidence interval and slope to make projections at the low end of the dose-response curve." The Panel was unable to discern what data need would be satisfied by the calculation of slope and CI, or how low on the dose-response curve that data points would be extracted.

If the slope is being used to estimate the LD<sub>p</sub>, where p is some toxicity rate other than 50%, then what values of p are being used and for what purposes? The BRD presents one example in which 20% of the LD<sub>50</sub> is of interest. This example is odd in that the toxicity rate associated with 0.2 LD<sub>50</sub> depends on the steepness of the probit curve and has no intrinsic meaning. Furthermore, there is a problem with the regulations and/or procedures that use criteria based on k\*LD<sub>50</sub>, such as are reported in Federal Regulation (40 CFR(129)). It needs to be emphasized that k\*LD<sub>50</sub> is not LD(k\*50). For example, 1/10\*LD<sub>50</sub> is not the dose at which the chemical is toxic for 1/10\*50=5 percent of the population. The basis for this convention of setting standards at k\*LD<sub>50</sub> is incomprehensible because the toxicity rate at this level depends entirely on the slope of the dose-response curve and does not provide a constant standard in obvious manner. Criteria for toxicity should be stated in terms of the LD<sub>p</sub>, where p is between 0 and 1, and presumably less than or equal to 0.5.

The level of precision required for the estimates of slope and CI should be stated. This information is important because a procedure that is efficient for one objective is likely to be less efficient for a different objective. A toolbox of procedures is needed to meet different objectives. For example, a good procedure for estimating the LD<sub>50</sub> and the slope will not be so helpful in estimating the LD<sub>p</sub> for p far from 50. The latter would require the correct model and extremely good precision. The

consequences of using a procedure for anything but its designed purpose need to be presented. The BRD should clarify whether a CI is for the LD<sub>50</sub>, the slope, or if both are needed. It should also be stated how the CI is to be calculated and interpreted.

Although not explicitly stated, it appeared to the Panel that there was a lack of distinction between the CI for the LD<sub>50</sub> and certain percentiles of the probit curve. These two need to be clearly defined in the Guideline to avoid confusion. In particular, if exposures were selected independently and randomly from a normal density, a 95% CI for the LD<sub>50</sub> would be the estimated LD<sub>50</sub> +/- 1.96\*sigma/sqrt(n), where n is the sample size. However, in none of the procedures (1987 OECD TG 401, OECD TG 425, or the revised UDP; **Appendices I, H, or G**, respectively) are exposures selected randomly from a normal density. Thus, the use of the constant 1.96 in establishing a CI for the LD<sub>50</sub> is arbitrary and not related in any known manner to some degree of confidence. In fact, the LD<sub>50</sub> +/- 1.96 sigma gives estimates of the LD<sub>2.5</sub> and the LD<sub>97.5</sub>. The CI for the LD<sub>50</sub> using the UDP and its revision will depend on the interval between doses as well as on sigma. The formula for the CI of the LD<sub>50</sub> also will depend on the type of estimator (e.g., Maximum Likelihood Estimate (MLE) or Modified Isotonic Estimate (MIE)) and the procedural rules that prescribe how exposures are selected.

The CI for the LD<sub>50</sub> given maximum likelihood estimation can be obtained using an expression for the variance of the estimated LD<sub>50</sub> that is given, for example, by Mats et al. (1998). It could also be obtained from replicated experiments or bootstrapping [See Stylianou (2000), for details on bootstrapping the CI of the LD<sub>50</sub>].

From the simulations, the dose progression proposal appears to be efficient for estimating the slope when it is high, but not when the slope is low. Furthermore, few animals are tested at doses far from the LD<sub>50</sub>, therefore, the efficiency level for this procedure is not maximized. In the BRD (U.S. EPA Document 8, Part D), it is shown that treating near, but not at, the optimal dose can result in significantly reduced efficiency. A slight

modification of the UDP as described in the April 2000 Guideline Appendix II (formerly **Appendix C**, currently **Appendix G**) will cluster the exposures around the optimal doses, even though they are unknown. We anticipate that other starting and stopping rules, as well as a dose progression schedule, can be developed to improve the current proposed UDP Supplemental Test, as well as the 1987 OECD TG 401 (formerly **Appendix A**, currently **Appendix I**) procedure.

## 5.2 UDP Supplemental Protocol

The general description is unclear as written. The complicated, statistically-based language is difficult to comprehend and translate into a manageable protocol, even by an experienced study director. More detail is needed and an example of the procedure (i.e., showing dose levels with response/no response) would be helpful. Potential problems exist where the Guideline makes statements such as "based on results, good judgement is required" and a possible "alternate procedure" may be appropriate. Also, an explanation for "staggered" starting doses is needed. The use of other acute toxicity information is mentioned, but is neither discussed nor is its relevance to dose setting addressed.

Computer simulations were used to consider possible outcomes of the UDP Supplemental Test and these simulations seem adequate. However, this approach is no substitute for actual laboratory studies. Comments from laboratory personnel who conduct these studies routinely should be carefully considered. Not only should the predictability of the test be considered, but also the difficulty involved in conducting the test. This procedure would require constant monitoring of responses and identification of each next dose, followed by a relatively complicated computer analysis for slope and CI.

The UDP Supplemental Test will take longer to complete as compared with a standard LD50 OECD TG 401 study (formerly **Appendix A**, currently **Appendix I**). A time of 48 hours between each dosing must be used. If dosing was performed on Monday, Wednesday, and Friday (requiring observations on Saturday and Sunday),

and 15 animals were needed, the test would take at least five weeks to complete. The UDP Supplemental Test would require at least another five weeks, for a total of at least 10 weeks. This is a relatively long time period for conducting an acute oral toxicity study. Industry is attempting to shorten development timelines for new chemicals as much as possible and an additional month of testing for an acute oral LD50 study could be significant. In addition, the need to test large numbers of chemicals, as in the High Production Volume chemicals program, will result in testing laboratories quickly reaching capacity. The time to complete these studies should be considered.

There are major concerns over the practicality of performing the UDP Supplemental Test in a standard toxicology laboratory. To ensure that the age/weight range is not exceeded late in the testing period, the number of animals required at study initiation could be quite high. Many of these could be wasted if other tests were not being conducted in the laboratory over the same period. Hence, not only does the UDP Supplemental Test procedure use no fewer animals than the OECD TG 401 procedure, it could indirectly result in the death of more animals because unused animals may have to be culled.

While, on the surface, the UDP Supplemental Test appears quite simple to conduct, the uncertainties that may be involved make it far from simple. Moreover, because the UDP Supplemental Test has never actually been conducted *in vivo*, the question of whether the general procedures are appropriate and described in sufficient detail cannot be ascertained.

## 5.3 Performance of the UDP Supplemental Test based on Computer Simulations

Based only on computer simulations, the usefulness of the UDP Supplemental Test cannot be determined without better knowledge of its intended purpose. The numbers and types of chemicals represented by the simulations were not appropriate. Reference was made to a listing of data from six pesticides, but there was no indication in the BRD as to where this information was used. The range of dose-response curves presented seemed adequate; however, very

shallow or steep dose-response curves should have been discussed in greater depth.

There was little evidence that the developers attempted to summarize the results from the large number of simulations. The description of Simulations II and III of BRD former U.S. EPA Document 8, Part D (current **Appendix N-4**), states that “for each run the computer randomly picked the appropriate number of animals from the entire population ...”. What is this population? Is it assumed that the animals are normally distributed around the LD50, with standard deviation sigma, and if so, why would this be the case? A population of very sensitive animals might be concentrated around the LD85, for example. If some other distributional assumptions were made, what are they?

#### **5.4 Reliability (Intra-Laboratory Repeatability, Inter-Laboratory Reproducibility) of the UDP Supplemental Test**

A major weakness of the proposed UDP Supplemental Test is that no confirmatory testing against conventional *in vivo* studies has been conducted. Any conclusions regarding the reliability of the UDP Supplemental Test are significantly restricted by the absence of *in vivo* data. The premise that computer simulations alone are sufficient for predicting biological events is not accepted by most scientists in the life sciences arena.

The issue of intra- and inter-laboratory variability has not been adequately addressed for the UDP Supplemental Test protocol. This failure is a major reason for a lack of confidence in this procedure. Some inter-laboratory variability is inherent in any test and information in the BRD indicates that values obtained with the standard LD50 study can vary by at least three-fold. There have been no inter-laboratory variability comparisons for the revised UDP Primary Test or for the UDP Supplemental Test. With the UDP Supplemental Test, additional variability may result from the fact that the rats tested may be of different weights/ages due to the length of testing. Also, the timeline for waiting for animal deaths to occur may add variability. Some investigators may dose animals every 48 hours to accelerate the

process, while others may wait longer between dosing to better assess for delayed deaths.

#### **5.5 Summary Conclusions**

1. The UDP Supplemental Test for slope and CI was not recommended for adoption. The Panel was unable to evaluate the utility of the test because sufficient information regarding the use of the data was not provided.
2. The revised UDP Primary Test and Limit Test adequately consider and incorporate procedures that reduce animal use. For the revised UDP Primary Test, the use of 0.5 log units for dose spacing is reasonable and appropriate based on experience and the results of computer modeling. This spacing allows the investigator to move through dose levels more quickly and thereby limits the number of animals used. In contrast, the UDP Supplemental Test, which includes the determination of slope, may use more animals than OECD TG 401 (formerly **Appendix A**, currently **Appendix I**). The UDP Supplemental Test does not replace animal use. Because the UDP Supplemental Test requires the use of starting doses below the LD50, there is a possibility that overall pain and distress may be reduced compared to OECD TG 401. At this point, there are no alternative animal species more suitable than rats for obtaining the type of information generated in acute toxicity testing.
3. The development of the UDP Supplemental Test has not followed the customary track for evaluating alternative methods in that only computer simulations were conducted. No actual *in vivo* testing was performed.
4. It is acknowledged that there has been a desire for a number of years to delete OECD TG 401, primarily for humane reasons. It is clear that the revised UDP Primary Test is an attractive replacement along with the revised UDP Limit Test, the FDP, and the ATC methods for estimating acute toxicity. While the UDP Supplemental Test was designed and proposed as a means of estimating the slope and CI, it is not clear whether this design is appropriate to address regulatory data needs. Moreover, these data needs have not been clearly presented to the Panel.

5. The BRD would be improved by closer attention to the norms of good method development and a clearer, more focused document preparation.
6. In Guideline Section 13.0 (UDP Supplemental Protocol) and in Addendum III of the Panel Report (Statistical Evaluation of the Revised UDP and the UDP Limit Test), a number of suggestions are offered that may be evaluated by the sponsors of this peer review.
7. If a procedure is needed to define points on the dose-response curve well below the median lethal dose, an alternative procedure, such as that detailed in Addendum I of this Report (Direct Estimation of a Point on the Dose-Response Curve that is far from the LD50), can be considered. Similarly, one possible alternative method for calculating the slope is presented in Addendum II of this Report (Consideration for Estimating the Slope).

#### **5.6 Recommendations**

1. Regulatory data needs currently addressed by estimation of the slope and CI derived from acute oral toxicity studies in the rat and other species need to be more clearly defined.
2. Consideration should be given as to whether the slope and CI are the most appropriate parameters for addressing regulatory data needs or if these needs can be addressed more directly. For example, an alternative procedure outlined in Addendum I of this Report may be used to estimate points on the dose-response curve well below the median lethal dose.

## 6.0 REFERENCES

- 16 CFR 1500. 2000. Title 16: Commercial Practices. Chapter II. Consumer Product Safety Commission. Part 1500: Hazardous Substances and Articles; Administration and Enforcement Regulations. Government Printing Office, Washington, DC.
- 29 CFR 1910.1200. 1998. Title 29: Department of Labor. Chapter XVII. Part 1910: Occupational Safety and Health Administration. Subpart Z: Toxic and Hazardous Substances. Section 1200: Hazard Communication. Government Printing Office, Washington, DC.
- 40 CFR 156. 2000. Title 40: Protection of Environment Agency. Code of Federal Regulations. Part 156: Labeling Requirements for Pesticides and Devices. Government Printing Office, Washington, DC.
- 49 CFR 173. 1999. Title 49: Department of Transportation. Code of Federal Regulations. Part 173: Shippers--General Requirements for Shipments and Packagings. Government Printing Office, Washington, DC.
- American Society for Testing and Materials (ASTM). 1987. Standard Test Method for Estimating Acute Oral Toxicity in Rats. ASTM E1163-87. In: Annual Book of ASTM Standards, Philadelphia.
- Barlow, R.E., D.J. Bartholomew, J.M. Brenner, and H.D. Brunk. 1972. Statistical Inference Under Order Restrictions: The theory and application of isotonic regression. John Wiley & Sons, New York. 388 pp.
- Bonnyns, E., M.P. Delcour, and A. Vral. 1988. Up-and-Down Method as an Alternative to the EC-Method for Acute Toxicity Testing. IHE Project No. 2153/88/11. Institute of Hygiene and Epidemiology, Ministry of Public Health and the Environment, Brussels. 33 pp.
- Bruce, R.D. 1987. A Confirmatory Study for the Up-and-Down Method for Acute Toxicity Testing. *Fundam. Appl. Toxicol.* 8:97-100.
- Bruce, R.D. 1985. An Up-and-Down Procedure for Acute Toxicity Testing. *Fundam. Appl. Toxicol.* 5:151-157.
- Dixon, W.J. 1991. Staircase Bioassay: The up-and-down method. *Neurosci. Biobehav. Rev.* 15:47-50.
- Dixon, W.J. 1965. The Up-and-Down Method for Small Samples. *J. Am. Stat. Assoc.* 60:967-978.
- Dixon, W.J., and A.M. Mood. 1948. A Method for Obtaining and Analyzing Sensitivity Data. *J. Am. Stat. Assoc.* 48:109-126.
- Durham, S.D., and N. Flournoy. 1995. Up-and-Down Designs I: Stationary treatment distributions. In: *Adaptive Designs*, Flournoy, N. and W.F. Rosenberger (Eds.). Hayward, California: Institute of Mathematical Sciences. pp. 139-157.
- Durham, S.D., and N. Flournoy. 1994. Random Walks for Quantile Estimation. In: *Statistical Decision Theory and Related Topics V*, Gupta, S.S., and J.O. Berger (Eds.). New York: Springer-Verlag. pp. 467-476.

- Durham, S.D., N. Flournoy, and A.A. Montazer-Haghighi. 1995. Up-and-Down Designs II: Exact treatment moments. In: Adaptive Designs, Flournoy, N., and W.F. Rosenberger (Eds.). Hayward, California: Institute of Mathematical Sciences. pp. 158-178.
- Durham, S.D., N. Flournoy, and W.F. Rosenberger. 1997. A Random Walk Rule for Phase I Clinical Trials. *Biometrics* 53:745-760.
- Flournoy, N. 1993. A Clinical Experiment in Bone Marrow Transplantation: Estimating a percentage point of a quantal response curve. In: Case Studies in Bayesian Statistics, Gatsonis, C, J.S. Hodges, R.E. Kass, and N.D. Singpurwala (Eds.). New York: Springer-Verlag. pp.324-336.
- Galson, S. 2000. Historical and current regulatory perspectives. Opening Plenary Session, ICCVAM International Workshop on *In Vitro* Methods for Assessing Acute Systemic Toxicity, October 17-20, 2000.
- Griffith, J.F. 1964. Interlaboratory Variations in the Determination of Acute Oral LD50. *Toxicol. Appl. Pharmacol.* 6:726-730.
- ICCVAM. 1997. Validation and Regulatory Acceptance of Toxicological Test Methods: A report of the *ad hoc* Interagency Coordinating Committee on the Validation of Alternative Methods. NIH Publication 97-3981. National Institute of Environmental Health Sciences, Research Triangle Park, NC. Available: <http://iccvam.niehs.nih.gov/docs/guidelines/validate.pdf> [cited October 18, 2001].
- Levitt, H. 1971. Transformed Up-Down Methods in Psychoacoustics. *J. Acoustical Soc. America* 49:467-447.
- Lipnick, R.L., J.A. Cotruvo, R.N. Hill, R.D. Bruce, K.A. Stitzel, A.P. Walker, I. Chu, M. Goddard, L. Segal, J.A. Springer, and R.C. Myers. 1995. Comparison of the Up-and-Down, Conventional LD50, and Fixed-Dose Acute Toxicity Procedures. *Food Chem. Toxicol.* 33:223-231.
- Mats, V.A., W.F. Rosenberger, and N. Flournoy. 1998. Restricted Optimality for Phase I Clinical Trials. In: New Developments and Applications in Experimental Designs, Flournoy, N., W.F. Rosenberger, and W.K. Wong (Eds.). IMS Monograph Series 34:50-61.
- Mulder, G.J. 1986. Sex Differences in Drug Conjugation and their Consequences for Drug Toxicity. Sulfation, glucuronidation and glutathione conjugation. *Chem. Biol. Interactions* 57:1-15.
- National Institute of Environmental Health Sciences (NIEHS). 2000a. National Toxicology Program: Request for Data and Nomination of Expert Scientists to Participate in the Independent Peer Review Evaluation of the Revised Up-and-Down Procedure for Assessing Acute Oral Toxicity. Evaluation of the Up-and-Down Procedure. 65 FR 8385. February 18, 2000.
- NIEHS. 2000b. National Toxicology Program: Notice of Peer Review Meeting on the Revised Up-and-Down Procedure (UDP) as an Alternative Test Method for Assessing Acute Oral Toxicity. Request for Comments. 65 FR 35109. June 1, 2000.
- Nelson, D.R., L. Koymans, T. Kamatski, J.J. Stegeman, R. Feyereisen, D.J. Waxman, M.R. Waterman, O. Gotoh, M.J. Coon, R.W. Estrabrook, I.C. Gunsalus, and D.W. Nebert. 1996. P450 Super Family: Update on new sequences, gene mapping accession numbers and nomenclature. *Pharmacogenetics* 6:1-42.

- Organisation for Economic Co-operation and Development (OECD). 2001. Harmonised Integrated Classification System for Human Health and Environmental Hazards of Chemical Substances and Mixture. OECD Chemicals Committee and the Working Party on Chemicals, Pesticides, and Biotechnology, Series on Testing and Assessment, No. 33. OECD, Paris. 247 pp. Available: <http://www.oecd.org/ehs/class/HCL6.htm>. [cited October 18, 2001].
- OECD. 2000. Guidance Document on the Recognition, Assessment, and Use of Clinical Signs as Humane Endpoints for Experimental Animals used in Safety Evaluation, OECD Environmental Health and Safety Publications, Series on Testing and Assessment, No. 19. OECD, Paris. 44 pp. Available: <http://www.oecd.org/ehs/test/monos.htm>. [cited October 18, 2001].
- OECD. 1999a. OECD Guideline for Testing Chemicals Revised 420: Acute Oral Toxicity - Fixed Dose Procedure. OECD, Paris.
- OECD. 1999b. OECD Guideline for Testing Chemicals Revised 423: Acute Oral Toxicity-Acute Toxic Class Method. OECD, Paris.
- OECD. 1998a. OECD Guideline for Testing Chemicals 425: Acute Oral Toxicity: Up-and- Down Procedure. OECD, Paris.
- OECD. 1998b. Harmonized Integrated Hazard Classification System for Human Health and Environmental Effects of Chemical Substances as Endorsed by the 28<sup>th</sup> Joint Meeting of the Chemicals Committee and the Working Party on Chemicals in November 1998, Part 2, p. 11. Available: <http://www.oecd.org/ehs/class/HCL6.htm>. [updated August 14, 2001 -- see also OECD, 2001].
- OECD. 1987. OECD Guideline for Testing Chemicals Test Guideline 401: Acute Oral Toxicity. OECD, Paris.
- Schlede, E., W. Diener, U. Mischke, and D. Kayser. 1994. Organisation for Economic Co-operation and Development expert meeting: Acute toxic class method. January 26-28, 1994, Berlin, Germany.
- Schlede, E., U. Mischke, W. Diener, and D. Kayser. 1995. The International Validation Study of the Acute Toxic Class Method (oral). *Arch. Toxicol.* 69:659-670.
- Schlede, E., U. Mischke, R. Roll, and D. Kayser. 1992. A National Validation Study of the Acute Toxic Class Method - An alternative to the LD50 test. *Arch. Toxicol.* 66:455-470.
- Sitter, R.R. and C.F.J. Wu. 1993. Optimal Designs for Binary Response Experiments: Fieller, D and A criteria. *Scandinavian J. Statistics* 20:329-341.
- Robertson, T., F.T. Wright, and R.L. Dykstra. 1988. *Order Restricted Statistical Inference*, John Wiley & Sons, New York.
- Spielmann, H., E. Genschow, M. Liebsch, and W. Halle. 1999. Determination of the Starting Dose for Acute Oral Toxicity (LD50) Testing in the Up-and-Down Procedure (UDP) from Cytotoxicity Data. *ATLA* 27:957-966.
- Stylianou, M. 2000. *A New Approach to Dose Finding for Phase I Clinical Trials*. Dissertation. American University.

- Stylianou, M., and N. Flournoy. 2000. A New Approach to Dose Finding for Phase I Clinical Trials. Technical Report Number 2000-2. Department of Mathematics and Statistics. American University.
- Trevan, J.W. 1927. The Error of Determination of Toxicity. *Proc. Royal Soc.* 101B:483-514.
- van den Heuvel, M.J., D.G. Clark, R.J. Fielder, P.P. Koundakjian, G.J.A. Oliver, D. Pelling, N.J. Tomlinson, and A.P. Walker. 1990. The International Validation of a Fixed-Dose Procedure as an Alternative to the Classical LD50 Test. *Food Chem. Toxicol.* 28:469-482.
- van den Heuvel, M.J., A.D. Dayan, and R.O. Shillaker. 1987. Evaluation of the BTS Approach to the Testing of Substances and Preparations for their Acute Toxicity. *Human Toxicol.* 6:279- 291.
- Weil, C.S. 1983. Economical LD50 and Slope Determinations. *Drug Chem. Toxicol.* 6:595-603.
- Weil, C.S. 1975. Toxicology Experimental Design and Conduct as Measured by Interlaboratory Collaborative Studies. *J. Off. Anal. Chem.* 58:683-688.
- Weil, C.S., C.P. Carpenter, and H.F. Smyth. 1953. The Median Effective Dose. *Ind. Hyg. Q.* 14:200-206.
- Weil, C.S., C.P. Carpenter, J.S. West, and H.F. Smyth. 1966. Reproducibility of Single Oral Dose Toxicity Testing. *Am. Ind. Hyg. Assoc. J.* 27:483-487.
- Weil, C.S., and G.J. Wright. 1967. Intra- and Inter-laboratory Comparative Evaluation of a Single Oral Test. *Toxicol. Appl. Pharm.* 11:378-388.
- Yam, J., P.J. Reer, and R.D. Bruce. 1991. Comparison of the Up-and-Down Method and the Fixed Dose Procedure for Acute Oral Toxicity Testing. *Food Chem. Toxicol.* 29:259-263.
- Zbinden, G., and M. Flury-Roversi. 1981. Significance of the LD50 Test for the Toxicological Evaluation of Chemical Substances. *Arch Toxicol.* 47:77-99.

**Addendum I: Direct Estimation of a Point on the Dose-Response Curve That Is Far From the LD50**

Estimating a LDp value that is near the LD50 is quite robust with respect to model assumptions; however, sensitivity increases as the LDp of interest moves away from the LD50. This increase in sensitivity is as expected because typical models (e.g., logistic, probit, Weibull) differ most in the tails. Relying on estimates of model parameters to estimate a low (high) LDp with only a few animals should and can be avoided by using a nonparametric procedure with a nonparametric estimator.

Exposures can be tailored to cluster around an unknown LDp, such as the LD16, using a slight modification of the UDP called the Biased Coin Up-and-Down Design (BCD) [Durham and Flournoy, 1994; see also Durham et al., 1997].

By using the BCD with any increasing dose-response function, such as the probit, exposures will quickly cluster around any target LDp, similar to what the standard UDP does for the LD50. To cluster points around the LD1p,  $p = 0.50$ , proceed as follows:

Use a biased coin, with probability of heads  $= [p/(1-p)]$ . If there is a toxic response, treat the next animal at the next lower dose; if there is a non-toxic response, flip the biased coin. If the coin comes up tails, treat the next sequential animal at the same dose; if the coin comes up heads, treat the next sequential animal at the next higher dose.

Note that for  $p=0.50$ , the BCD procedure reduces to Dixon and Mood's (1948) up-and-down design. For  $p>0.50$ , see Durham and Flournoy (1994).

The Modified Isotonic Estimate (MIE) of the LDp, described in Addendum IV, is an attractive alternative to the Maximum Likelihood Estimate (MLE) since it does not require a probit or other parametric model assumption. This approach is particularly important for estimating a LDp far from the LD50 where model differences are most pronounced. Stylianou and Flournoy (2000)

demonstrate that the MIE outperforms other nonparametric estimators found in the literature, and compares well with the MLE.

It appears that no one asked how accurately OECD TG 401 (formerly **Appendix A**, currently **Appendix I**) provided estimates of toxicity at low doses, using the estimation of the slope in a probit model; however, the Panel was asked to evaluate the UDP Supplemental Test protocol for estimating toxicity rates at fractions of the LD50. Finding that little thought had been given to precision, our evaluation cannot determine whether this requirement will be met. Some consideration should be given to stopping rules that take precision into account. Stylianou (2000) considered stopping rules for the BCD. A likelihood ratio test similar to Rule #3 in the revised UDP Primary Test may work well also. This approach should be evaluated.

## Addendum II: Considerations for Estimating the Slope

The "optimal design" (i.e., the procedure yielding the most information about the LD50 and the slope simultaneously, with a fixed number of animals) would be to administer the test substance to animals (*cf.* Sitter and Wu, 1993) at the:

- LD13 and LD87 if the response function is probit,
- LD18 and LD82 if the response function is logistic,
- LD10, 50, and 90 if it is double exponential, and
- LD21, 50, and 79 if it is double reciprocal.

A compromise might be to treat animals at LD16, LD50, and LD84 (if possible). If avoiding highly toxic doses is desired, the LD16 and LD50 are attractive choices. Assuming a probit dose-response function, the LD16 and LD84 are  $-1$  and  $+1$  sigma from the LD50, respectively. Thus, the estimates of sigma can be obtained from estimates of  $[\text{LD84-LD16}]/2$ ,  $[\text{LD84-LD50}]$ , and  $[\text{LD50-LD16}]$ . Differences in these estimates would indicate that the sample size is too small or that the probit model is not a good fit.

As recognized by the development team for the revised UDP, even assuming the probit model, it is impossible to implement the optimal design because the optimal values of LDp are unknown. Certainly, selecting a few dose levels (based on certain expectations as in OECD TG 401) and treating a fixed number of animals at those dose levels can be very inefficient, because even good expectations based on considerable experience can be incorrect (see, for example, Flournoy, 1993). Simulations in BRD U.S. EPA Document 8, Part D demonstrate also the decline in efficiency that can result from the use of designated points near, but not at, the optimal ones.

To deal with this efficiency issue, the UDP Supplemental Procedure incorporates several escalation-dosing series, starting at low doses. The problem with increasing the dose at every nontoxic outcome is that exposures are closer to the LD50 than to doses such as the LD16 after only a couple of animals.

Simulations in former U.S. EPA Document 8, Part D (current **Appendix N-4**) indicate that the UDP Supplemental Test procedure yields a reasonable estimate of sigma when sigma is small, but substantially underestimates sigma when sigma is large. This discrepancy could result from the dose escalation procedures when very few animals are tested at levels far from the LD50, or because of the large interval between doses. These two possibilities should be examined.

To shorten the time required for estimating the LD50 and slope together, simultaneously conducting BCD procedures to target two or three points on the dose-response curve (e.g., the LD16 and LD50, the LD16 and LD84, or the LD16, LD50, and LD84) should be considered. Clustering treatments around but not at two or three nearly optimal dose levels using simultaneous BCD is expected, on theoretical grounds, to produce more efficient estimates of the LD50 and slope when compared to the UDP Supplemental Test.

MIE (see Addendum IV of this report) of the necessary LDp values are attractive alternatives to MLE. Of course, more animals are required to estimate LDp values distant from the LD50, but at least for doses as low as the LD10, the expected increase in the number of animals is modest. In particular, the expected number of animals required is less than that required by the combined UDP Primary and Supplemental Tests for estimating both the LD50 and sigma. Additionally, targeting the LD16 and the LD50 will be less efficient for estimating sigma and the LD50 than targeting the LD16, LD50, and LD84, and also much less efficient than targeting only the LD16 and the LD84. The relative efficiency of targeting the three points versus two points on the dose-response curve should be examined. For example, it could take many more animals targeting two dose levels (instead of three) to get the same quality estimates of the LD50 and sigma. If animals should not be treated around the LD84 to avoid pain and suffering, this point is moot.

**Addendum III: Summary of the Statistical Evaluation of the Revised UDP**

Significantly more information per animal will be obtained using an up-and-down procedure for estimating the LD50 when compared to treating fixed numbers of animals at several doses. This increase in the extent of information per animal has been shown theoretically (*cf.* former references 1-6 of U.S. EPA Document 2, current **Appendix J-2**) and has been demonstrated in the simulation studies provided in the BRD. A suggestion to simplify the use of the likelihood ratio statistic as a stopping rule is offered for consideration by the development team.

It is important to recognize that the variability of the LD50 estimate increases with the step size used between sequential dose levels. The UDP is proposed for many different purposes and varying degrees of precision will be appropriate for different purposes. For example, for the crude classification of chemicals, a large dose progression factor with its associated relatively large variation in the LD50 estimate will be satisfactory. However, when considering the effect of a chemical on an endangered species, considerably greater precision is desired. One may predict that the precision expected for some purposes simply cannot be obtained with the proposed step size. To prepare for a revision (perhaps three years from now), it is recommended that the precision desired for different purposes be ascertained. This information would be used to develop rules for adjusting the step size (and perhaps the nominal sample size and stopping criteria as well) to allow the procedure to yield the desired precision.

**THE PRIMARY PROCEDURE**

With respect to generating the most information per animal, the LD50 is the most simple single summary statistic to measure on the dose-response curve. An up-and-down procedure is very efficient, in terms of the number of animals used, for obtaining this estimate. The up-and-down procedure specified in OECD TG 425 has been demonstrated to efficiently estimate the LD50, except when the step size is based on a "slope"

estimate that is very far from reality or when the initial dose is distant from the LD50. A number of reasonable suggestions are made to mitigate these problems.

1. Stopping rule #3 involves those special cases when the procedure has not stopped at or before the nominal sample size is achieved. In this case, the recommendation is to stop if the likelihood ratio statistics for testing whether the true LD50 is 2.5 times greater than the estimate or 1/2.5 less than the estimate are both greater than 2.5. Simulations show this modification yields a great improvement in the estimates, particularly, when the slope is low or the initial treatment is far from the LD50. These ideas are strongly endorsed.
2. One modification to stopping rule #3 that warrants consideration is to calculate the likelihood using MIE of the dose-response function. MIEs have the advantage of (1) being very easy to calculate (a laboratory technician can compute MIEs without need of a computer; see Addendum IV of this report) and (2) not requiring an estimate of sigma when using the null hypothesis. An estimate of the slope is required for calculating the likelihood under the alternative hypotheses used in stopping rule #3.
3. Assuming a probit response function, a crude estimate of sigma can be obtained from the MIE of the dose-response function (rather than using a default estimate). Sigma can be estimated, for example, by noting that LD50-<sub>sigma</sub> is the 31st percentile of the normal probability density and LD50+<sub>sigma</sub> is the 69th percentile. Reading off the 31st and 69th percentiles (LD31 and LD69) of the interpolated isotonic estimate of the dose-response function, an estimate of sigma is (LD69-LD31). In addition, 2\*(LD50-LD31) and 2\*(LD69-LD50) provide two estimates of sigma. If they are very close to each other, the estimate (LD68-LD32)/2 should be

reasonable. A large difference might reflect the small sample size or it might indicate that the dose-response function is not symmetric, as is assumed by the probit model. Because of the relatively large interval between doses in the revised UDP Primary Test, it might be reasonable for the purpose of stopping to estimate sigma using estimates of LD<sub>p</sub> values that are more distant from the LD50 than are the LD31 and LD69 (e.g., LD16 and LD84). Because the data are clustered around the LD50, any estimate of sigma will not be very accurate, but it is worth evaluating whether this approach is better than assuming the default when the default is not true.

4. Future work, which should not interfere with the adoption of the current proposal, includes obtaining the exact distribution of the likelihood ratio statistics. This task will permit the critical value of 2.5 to be adjusted to satisfy the accuracy required for a particular application and should not be too difficult to accomplish assuming a (probit) model.
5. It needs to be emphasized that a variable stopping rule is essential in dose-response studies, because the investigator does not know how distant the initial dose level is from the LD50 (see Flournoy, 1993, for example). The development team for the revised UDP Primary Test recognized this need in developing the revised test.
6. Another recommendation is to increase the default step size. The recommendation is to adopt this proposal at this time. However, the issue of maintaining a constant step size throughout the experiment deserves additional investigation. For example, in the psychometrics literature (*cf.* Levitt, 1970), recommendations include doubling the step size after a string of like responses and halving the step size after a string of consecutive reversals. A procedure such as this could reduce the number of animals needed to get into the region of the LD50 (due

to starting far away) and decrease the width of a confidence interval around the LD50 (when a steep dose-response curve causes many consecutive reversals).

Producing a reasonable algorithm for changing the step size is a considerable effort, in and of itself, and becomes even greater when the varied purposes for which this UDP is proposed are considered. Consequently, it is not recommended that this subject be investigated for the current proposal to OECD, but be included in future revisions.

#### MISCELLANEOUS DETAILS

The term "LD50" should not be used for both the parameter and the estimate. This wording is confusing in the BRD.

Also, there is an objection to a dose-escalation procedure being referred to as an up-and-down design. The up-and-down design with a nominal sample size of two is a simple dose-escalation procedure, as there is no decrease in exposure levels. It will have none of the nice features of the biased coin up-and-down design, such as clustering treatments around a target LD<sub>p</sub>. To refer to dose escalation as an up-and-down procedure is equivalent to treating all the animals at the same dose level, but stating that they were treated according to the normal probability density with variance equal to zero.

**Addendum IV: Modified Isotonic Estimates of the Dose-Response Function**

Reviews of isotonic estimation can be found in Barlow et al. (1972) and Robertson et al. (1988), among others. Modified isotonic estimates (MIE) of the dose-response curve were proposed by Stylianou (2000) and are reported in Stylianou and Flournoy (2000). A brief description is given below.

At each dose, the proportion of deaths observed is calculated. These proportions are reconsidered beginning at the lowest dose level. The proportion of animals that died at the lowest dose is the isotonic estimate of the probability of death at this dose. If the proportion of deaths at the next higher dose level is larger than the first proportion, it is the isotonic estimate of the probability of death at the second dose level. At successively higher doses, the proportion of animals that died is considered to be the isotonic estimate of the death rate, until a proportion is observed that is lower than the previous proportion. The dose-response function should increase with dose. When the data are inconsistent with this assumption, a weighted average of the two proportions is calculated, with weights equal to the sample sizes at the two dose levels. The weighted average replaces the observed proportions of animals that died as the isotonic estimators. The investigator continues to compare each observed proportion of animals dying at a particular dose level with the proportion at the preceding dose level and combining estimates when they fail to increase with increasing dose level. When the highest dose level has been considered, all of the isotonic estimates have been calculated.

The isotonic estimators are calculated only at the dose levels used in the experiment. An estimate of the death rate at any dose level is obtained by plotting the isotonic estimates and drawing lines between the points by hand or by computer. The curve that results from this linear interpolation is called the MIE and can be used with any acute toxicity procedure to estimate any LDp.

Up-and-down procedures cluster dose levels around target dose levels (see Addendum I of this report). If the up-and-down procedure in the revised UDP Primary Test is used, estimates of mortality at dose levels distant from the LD50 will not be very accurate; whereas, if a biased coin up-and-down procedure is used, the estimates will not be very accurate at dose levels distant from the targeted LDp. As a consequence, estimates of mortality for a specified dose level need to be generated using a procedure that is appropriate for a particular goal.

