CVERVIEW OF METHODS FOR EVALUATING EFFECTS OF PESTICIDES ON REPRODUCTION IN BIRDS

by

Richard S. Bennett
USEPA Environmental Research Laboratory
200 SW 35th Street
Corvallis, Oregon 97333

and

Lisa M. Ganio
Mantech Environmental Technology, Inc.
USEPA Environmental Research Laboratory
200 SW 35th Street
Corvallis, Oregon 97333

ENVIRONMENTAL RESEARCH LABORATORY-CORVALLIS
OFFICE OF RESEARCH AND DEVELOPMENT
U. S. ENVIRONMENTAL PROTECTION AGENCY
CORVALLIS, OREGON 97333

OVERVIEW OF METHODS FOR EVALUATING EFFECTS OF PESTICIDES ON REPRODUCTION IN BIRDS

by

Richard S. Bennett
USEPA Environmental Research Laboratory
200 SW 35th Street
Corvallis, Oregon 97333

and

Lisa M. Ganio
Mantech Environmental Technology, Inc.
USEPA Environmental Research Laboratory
200 SW 35th Street
Corvallis, Oregon 97333

ENVIRONMENTAL RESEARCH LABORATORY-CORVALLIS
OFFICE OF RESEARCH AND DEVELOPMENT
U. S. ENVIRONMENTAL PROTECTION AGENCY
CORVALLIS, OREGON 97333

FOREWORD

This report provides an overview of methods for laboratory evaluations of the effects of pesticides on avian reproduction, including a review of current guidelines for conducting an avian reproduction test and possible alternative methods for improving hazard assessment. Field methods for assessing reproductive effects are not covered in this report. This report does not establish Agency policy for conducting avian reproduction tests. The methods discussed in this report that are at variance with the existing guidelines should not be taken as establishing new policy or superseding existing policy, however, they are presented for consideration in future policy decisions. Guidelines for the avian reproduction test are scheduled for revision in the near future.

This report is intended to address several issues raised by persons outside and within the Agency concerning the adequacy and utility of the current avian reproduction test. The primary audience for the report is persons involved with conducting and evaluating avian reproduction tests, but it is hoped that the discussion of these issues will lead to further research by persons interested in the effects of chemicals on avian populations.

ABSTRACT

The standard test methods used by the U. S Environmental Protection Agency for evaluating the effects of pesticides on avian reproduction are critiqued. The intent of the report is to review several concerns that have been raised about the adequacy of the test methods, discuss ways to improve the regulatory utility of the test, and present alternative methods that discuss aspects of avian reproduction that are not adressed by current test methods. The overview of current test methods includes the selection of measurement variables, experimental design, selection of test animals, testing environment and husbandry methods, egg collection and incubation, observations of progeny, and data analysis and interpretation. The emphasis is on reducing variability in test data that is unrelated to pesticide treatment to improve the ability of the test to detect pesticide effects if they exist. Alternative test methods are discussed that determine a dose-response relationship of pesticide effects, employ a shorter exposure period for less persistent pesticides, or utilize parental incubation to determine pesticide-induced behavioral anomalies. Alternative measurement endpoints and data analysis methods also are presented.

ACKNOWLEDGMENTS

The authors wish to thank all those individuals that reviewed and commented on various drafts of this report, including (in alphabetical order) Jewel Bennett, Brian Collins, David Farrar, Ed Fite, Mary Frankenberry, Gary Heinz, Ron Kendall, Brad Marden, Dennis McLane, Pierre Mineau, Al Nebeker, and Robert Ringer. Their excellent comments and recommendations were extremely helpful in developing this report, although some of the recommendations will have to be the subject of future reports. We owe a special thanks to Jewel Bennett, whose earlier collaboration on several research projects provided the catalyst for this effort, and Anne Fairbrother, whose continuing support made this report possible. Also, we owe special thanks to Anne Barton and James Akerman for providing the senior author with the opportunity for a rotational assignment in the Office of Pesticide Programs to experience the regulatory process. The advice and example set by James Akerman will not be forgotten.

TABLE OF CONTENTS

SECTION 1

INTRODUCTION

The Office of Pesticide Programs (OPP) of the U. S. Environmental Protection Agency (EPA) is charged under the Federal Insecticide, Fungicide and Rodenticide Act (FIFRA) with determining whether a pesticide can be registered for a particular use. Under FIFRA, the EPA Administrator shall register a pesticide if it is determined that "when used in accordance with widespread and commonly recognized practice it will not generally cause unreasonable adverse effects on the environment" (P.L.95396, Sec 3 (c)(5)(D)). Additionally, FIFRA states that "the Administrator may conditionally amend the registration of such pesticide....if the Administrator determines that ...amending the registration....would not significantly increase the risk of any unreasonable adverse effects on the environment" (ibid. Sec. 3(c)(7)(B)). The determination of whether adverse effects are "unreasonable" is a risk management function that requires the integration of the benefits to society of the use of pesticides with the risks posed to human health and the environment.

The extent of ecological risk from a proposed use of a pesticide is estimated by developing an ecological risk assessment, which takes into account laboratory data, and in some instances field data, on the toxicity and effects of the pesticide (hazard) and the potential for exposure to nontarget organisms in the environment. OPP has developed a Standard Evaluation Procedure detailing the state-of-the-art for ecological risk assessment (USEPA 1986). For consistency and comparability of hazard data generated for the development of an ecological risk assessment, OPP has established Pesticide

Assessment Guidelines for conducting laboratory and field testing of pesticides in support of registration. The evaluation of pesticide hazards to nontarget wildlife and aquatic organisms are stated in Subdivision E of these guidelines (USEPA 1982a) (See Appendix A.). This report will focus on the avian reproduction test, section 71-4 of the Subdivision E guidelines, a laboratory test in the avian and mammalian testing series.

The avian reproduction test was first developed to evaluate the reproductive effects of the persistent organochlorine pesticides. These pesticides exist chronically in the environment and have been found to produce effects on avian reproduction, principally by effects on eggshell thickness (Anderson and Hickey 1972, Ratcliffe 1970) and embryonic development (Haegele and Hudson 1973, Heath et al. 1969, Longcore and Samson 1973). Avian reproduction tests were also used to evaluate potential effects of other persistent chemicals, such as polychlorinated biphenyls (Dahlgren and Linder 1971, Peakall et al. 1972) and methyl mercury (Heinz 1974, 1976a, Peakall and Lincer 1972).

As stated in Subdivision E, the avian reproduction test is required "when basic data and environmental conditions suggest possible problems" (USEPA 1982a). The stated purpose for the test is that "these tests are used principally:

-To estimate the potential for chronic impacts, taking into account the measured or estimated residues in the environment; and

-To determine if additional field or laboratory data are necessary to further evaluate hazards" (USEPA 1982a).

The test simulates a chronic dietary exposure to the pesticide with a goal of determining the highest dietary level (concentration) that produces no observable adverse effect (NOAEL) on a suite of reproduction parameters and the lowest dietary level that produces an observable adverse effect (LOAEL). The results of this test are then evaluated to determine if observed adverse effects occur at concentrations expected to exist in the environment from the proposed uses of the pesticide. If the ecological risk assessment based on these test results indicates that a proposed use would pose an unreasonable adverse effect, the registration may be referred to special review for risk/benefits analysis or additional laboratory or field testing may be required to substantiate or rebut the presumption of risk.

Currently, the avian reproduction test is the only standardized test in Subdivision E that simulates a chronic dietary exposure, and thus represents the only method for detecting the long-term effects of chemicals that are either persistent in the environment or that are applied to the environment repeatedly or continuously. In addition to measuring chronic toxicity and reproductive impairment, the chronic dietary exposure presents a method for evaluating the potential for pesticide bioaccumulation in organism tissues and deposition into eggs. Additionally, the avian reproduction test is the only standardized laboratory test in Subdivision E that focuses on parameters other than mortality.

Although the avian reproduction test provides information on pesticide toxicity not found in any other laboratory test, there are several concerns with the current test and the associated data analysis. First, the test data generated are often highly variable so that biologically significant effects on reproduction may not be statistically detectable. Second, guidance for carrying out statistically sound analyses of test data is lacking in the current protocols. Third, chronic dietary exposure methods may not provide the most representative exposure scenario for detecting effects on reproduction from newer, less persistent pesticides. Effects on avian reproduction do not result only from chronic exposure, but can occur after short-term pesticide exposures. Fourth, because reproducing birds may come in contact with pesticides for the first time during any phase of the reproductive period, including incubation and brood-rearing periods, many potential effects on reproductive output may not be evaluated by the current test guidelines. Fifth, standardized field methods for evaluating pesticide effects on avian reproduction are limited and their relationship to the laboratory avian reproduction test are not well understood. Therefore, it is difficult to conduct field tests that adequately evaluate the risks predicted by laboratory tests. The purpose of this report is to provide an overview of current test methods, evaluate the bases for the first four concerns listed above, and suggest alternative methods for addressing these concerns. This report will not address field methods or the comparability of laboratory and field methods. However, this report will discuss some aspects of how the avian reproduction test relates to what is known about pesticide effects on the reproduction of wild birds.

To achieve the goal set for the future of OPP's ecological risk assessment process (USEPA 1986, p. 1) of broadening assessments to the population, community and ecosystem levels of concern, a broader evaluation of direct and indirect effects of pesticide use is required, including nonlethal effects on reproductive potential. The dynamics of a defined wildlife population are governed by the three principle population parameters--birth, death, and movements in and out of the defined population. Any effect of the use of a pesticide that changes the birth rate per female of reproductive age is considered a reproductive effect. Effects on the birth rate per female may occur as a change in the number of eggs per clutch, number of clutches per year, rates of egg fertility, embryo survival, or hatch success, or as mortality of reproducing adults. Additionally, mortality of hatchlings before recruitment into the population is often measured as a reproductive effect. The deaths of pre-recruitment young may result from direct toxicity from embryonic or post-hatching exposure to the pesticide or pesticide-related effects to the parent. Ideally, an ecological risk assessment evaluates if, and to what extent, the proposed use of a pesticide would adversely affect the number of young produced per female of reproductive age and their recruitment to the population.

The avian reproduction test does not and can not provide all this information for an ecological risk assessment. It is the first line screening test for identifying potential reproductive effects to birds. As such, it utilizes a suite of measurement endpoints to evaluate the potential effects of pesticides on several components of avian reproduction and to help identify cause and effect relationships. In practice, the avian reproduction test is often the first step in a multiple-step process of assessing the risk to avian reproduction from the proposed use of a pesticide. The focus of this report is to discuss methods for improving the avian reproduction test as a screening tool for ecological risk assessment.

SECTION 2

OVERVIEW OF CURRENT AVIAN REPRODUCTION TEST METHODS

CRITERIA FOR DETERMINING WHEN TEST IS REQUIRED

Current Criteria

The Subdivision E guidelines (USEPA 1982a) state that "data on avian reproduction are required by 40 CFR 158.145 to support the registration of an end-use product which meets one or more of the following criteria:

1) Its labeling contains directions for using the product under conditions where birds may be subject to repeated or continuous exposure to the pesticide or any of its major metabolites or degradation products, especially preceding or during the breeding season.

2) The pesticide or any of its major metabolites or degradation products are stable in the environment to the extent that potentially toxic amounts may persist in avian feed.

3) The pesticide or any of its major metabolites or degradation products is stored or accumulated in plant or animal tissues, as indicated by the partition coefficient of lipophilic pesticides (tests 165-3, -4, and -5 of Subdivision N)(USEPA 1982c), metabolic release and retention studies (test 83-4 of Subdivision F)(USEPA 1982b), or as indicated by structural similarity to known bioaccumulative chemicals.

4) Any other information, such as that derived from mammalian reproduction studies (test 83-4 of Subdivision F)(USEPA 1982b), that indicates that reproduction in terrestrial vertebrates may be adversely affected by the anticipated use of the pesticide product".

Potential Reproductive Effects Not Covered by Criteria

The above criteria were developed when the test was primarily used to determine effects of organochlorine pesticides and other persistent chemicals and reflect the concern for pesticides with chronic exposure patterns. The criteria would not necessarily trigger a test for pesticides that pose risk of adverse reproductive effects from short-term exposure. Several pesticides have been shown to reduce egg production within days after initiation of dietary exposure (Bennett and Bennett 1990, Bennett et al. 1991). Effects on eggshell quality (Bennett and Bennett 1990, Haegele and Tucker 1974) and incubation and brood rearing behavior (Bennett et al. 1991, Brewer et al. 1988, Busby et al. 1990) have also resulted from short-term pesticide exposures.

PROCEDURES FOR CONDUCTING AVIAN REPRODUCTION TESTS

Selection of Measurement Endpoints

The Subdivision E guidelines (USEPA 1982a) provide an extensive list of data requirements to be reported for complying with the test standards. Much of this information relates to the general test conditions (see section 70-3, General test standards), with specific information on test methods and the test substance and organism required. Additional information on the test conditions is specified in section 71-4, Avian

Reproduction Test. Information that defines the conditions for the entire test includes: ambient temperature and humidity, photoperiod and lighting intensity, description of the test diet (including composition and proximate analysis), source of food and water supply, pretest and test history of medical and chemical administration, dimensions and materials of test pens, temperature and humidity of egg incubators, and egg turning frequency. These measurements are critical for evaluating the appropriateness of test conditions and for comparing results between tests.

Other test standards specified in section 71-4 represent measurements made on each experimental unit (i.e., pen) in the test to be analyzed for pesticide treatment effects. Several of the measurements are not discussed in the section on statistical analysis and are often used as qualitative indicators. However, each of these measurements could, and sometimes are, quantified and analyzed statistically. These include signs of abnormal behaviors, signs of intoxication in hatchlings, observed morphological and physiological responses of adults, observations on the palatability or repellency of test diets, and postmortem necropsy findings. These measurements are often critical for explaining treatment differences in the quantitative measurements.

Measurements endpoints obtained for each treatment (dietary concentration group) specifically mentioned for statistical analysis include:

--number of adult mortalities by sex
--adult body weight by sex
--average adult food consumption per pen per day or season
--average eggshell thickness per pen per day or season
--number of cracked eggs per pen per day or season

--proportion of cracked eggs per eggs laid per pen per day or season

--number of eggs laid per pen per day or season

--number of fertile eggs laid per pen per day or season

--proportion of fertile eggs per eggs set per pen per day or season

--number of live 3-week embryos per pen per day or season

--proportion of live 3-week embryos per fertile eggs per pen per day or season

--number of hatchlings per pen per day or season

--proportion of hatchlings per live 3-week embryos per pen per day or season

--proportion of hatchlings per eggs laid per pen per season

--number of 14-day-old survivors per pen per day or season

--proportion of 14-day-old survivors per number of hatchlings per pen per day or season

--proportion of 14-day-old survivors per eggs laid per pen per day or season

--average weight of hatchlings per pen per season

--average weight of 14-day-old survivors per pen per season

Most of the above measurements are based on the entire test period. Body weights are recommended at the initiation and termination of the test and at biweekly intervals until egg production begins. Measuring body weight during egg production is discouraged because it could have adverse effects on egg production. Hughes and Black (1976) found that handling laying hens adversely affected eggshell quality by increasing the incidence of cracks and equatorial bulges. Also, handling laying hens can lead to broken eggs in the oviduct. The average eggshell thickness per pen is calculated by collecting and measuring all eggs laid on one day every two weeks. The Subdivision E guidelines recommend that, for consistency, eggs be collected during weeks 1, 3, 5, 7 and 9 of the egg laying period.

The Subdivision E guidelines recommend that food consumption be recorded at least biweekly throughout the test. With mallards, it may be necessary to record consumption and change feed more frequently because they tend to put water in their food, which may become moldy. Fungi in the genus Aspergillus thrive on moist grain and feeds and can cause respiratory infections and death (Friend 1987). Also, it is more difficult to separate food consumption from spillage with mallards if the spilled food becomes wet so that the weight is difficult to estimate. Unless measures are taken to quantify spillage, the food consumption measurements represent a combination of food consumed and spilled, which compromises the detection of treated-related changes in consumption if birds vary greatly in the amount of food spilled.

The Subdivision E guidelines do not discuss the relative importance of each of the listed endpoints. A variety of endpoints are measured so that the most sensitive endpoints for each tested pesticide can be determined. Each endpoint may be an important indicator of potential effects on avian reproduction. This provides valuable information for determining the specific mechanism of action of the pesticide. However, of the above list of measurement endpoints, the number of 14-day-old survivors per pen may be the most biologically meaningful and comprehensive at evaluating overall effects of a pesticide on avian reproduction. The number of 14-day-old survivors per pen is a function of the number of eggs laid, the proportion of fertile eggs of those set, the proportion of live embryos of those fertile, the proportion of hatchlings of eggs with live embryos, and the proportion of hatchlings that survive to 14 days of age. Significant effects on these endpoints also may be observed as a significant effect on the number of 14-day-old survivors per pen if the power of the test is sufficient and there is no compensatory effect in another endpoint.

Experimental Design

The goal of an avian reproduction test should be to detect pesticide-related effects on the reproduction of the test population at the lowest dietary concentrations that produce biologically significant effects. This can be accomplished by choosing the appropriate measurement endpoints and developing a sufficiently powerful experimental design that will control variability unrelated to treatment in the test system by reducing influence of confounding variables. Failure to address these aspects of the test can lead to a test that fails to identify effects or can detect only very large effects, thus failing to detect effects at lower concentrations that may be biologically significant.

However, circumstances beyond the scientist's current knowledge or control often make it difficult to ensure that all avian reproduction studies are sufficiently powerful. Some general guidelines will be discussed in this report.

An experimental design is defined to be the choice of number and type of treatments (in this case, diets amended with pesticide at various concentrations) and experimental units (in this case, pens of reproductive birds) and the process of assigning the treatments to experimental units. Note that because the amended diet is provided to a pen and not to single birds, the pen is the experimental unit. A treatment group consists of several replicate experimental units that are administered the same dietary treatment. In a broad sense, other techniques necessary for the development and maintenance of the study, such as animal husbandry and the choice of birds, can also be thought of as aspects of the design of the study.

The Subdivision E guidelines list the minimum number of treatments (with one serving as a control group), replicates per treatment, and numbers of males (M) and females (F) per pen for an acceptable protocol for two avian species, the northern bobwhite (Colinus virginianus), hereafter referred to as bobwhite, and mallards (Anas platyrhynchos). These are:

Species	Treatments	No. of birds	Replicates
Bobwhite	3	1M & 2F	12
	3	1M & 1F	>12
Mallard	3	2M & 5F	5
	3	1M & 1F	>12

Although pairs are acceptable, these guidelines recommend the use of the group-pen (more than one male and one female per pen) design, based on an unreferenced analysis showing greater sensitivity to detect pesticide effects using the group-pen design. However, many investigators prefer the use of pairs for a variety of reasons, including the fact that both species are monogamous. Given the amount of data collected since the Agency's original analysis, it would be worthwhile again to compare the sensitivity of these two approaches for housing test animals.

The Subdivision E guidelines also provide an example of an avian reproduction protocol that is considered an acceptable design for both bobwhites and mallards. While the protocol states minimum values for the number of treatments and replicate pens per treatment and guidance on the selection of appropriate dietary concentrations, it is the responsibility of the investigators to determine an adequate experimental design and justification for concentrations chosen for the particular pesticide tested.

Selecting dietary concentrations--

The guidance in Subdivision E for selecting dietary concentrations states that they "should be based on measured or calculated residues expected in the diet from the proposed use pattern(s). The concentrations should include an actual or expected field residue exposure level and multiple level such as five" (USEPA 1982a). One of the concerns with the current test guidelines is that if no significant differences are detected in a test, it is very difficult to determine if this reflects that the concentrations tested are truly below those causing reproductive effects or if the test was inadequately designed to detect effects that occur. In such a case, the NOAEL is the highest concentration and the LOAEL is not determined. If a test fails to detect a reproductive effect, little is learned about the types of effects potentially caused by the pesticide at higher concentrations or about how close the tested concentrations are to those causing effects. Conversely, if reproductive effects are observed in all treatment groups, it may not be possible to determine if the same or different effects would be observed at lower concentrations. If at a later date, a new use for the pesticide is proposed that would result in higher or lower environmental concentrations, the test results may be of little value in assessing the potential risk of the new use, and the reproduction test would have to be repeated at different concentrations.

An alternative approach to selecting dietary concentrations on the basis of estimated environmental concentrations (EEC) is to conduct the test using concentrations in the range that produces effects in the more sensitive endpoints, unless the dietary concentrations required would be much greater than the EEC (e.g., 100 x EEC). American Society for Testing and Materials (ASTM 1990) discusses this approach as one option for choosing dietary concentrations. The test would be conducted with at least one treatment group producing significant effects on reproductive parameters and one occurring below effect concentrations. The advantages to this approach are: 1) the test could more accurately define the NOAEL and LOAEL for making a regulatory decision; 2) the relative distance between effect concentrations and the EEC for a proposed use could be established; and 3) the types of reproductive effects caused by a pesticide would be identified. The disadvantages of this approach are: 1) a range finding test or tests would be required to establish appropriate dietary concentrations for a definitive test; and 2) the determination of tissue and egg residues may be less meaningful if the test was conducted at concentrations considerably different from the EEC.

Length of test--

The Subdivision E guidelines state that birds should receive the treated diet ad libitum for the duration of the study, which includes "at least 10 weeks prior to the onset of egg laying" and an egg laying period of approximately 10 weeks, although the total duration of the test will be dependent on the onset of egg production. Most of the measurement endpoints are based on the calculated value for the entire egg laying season for each pen.

One confounding issue in this approach that may be unrelated to treatment is that some birds do not consistently lay eggs or go completely out of egg production before the end of a 10-week laying period. As the proportion of birds that terminate laying increases through the test, the variability in parameters, such as the number of eggs laid, increases among pens within a treatment. For example, all control birds may start out as consistent egg producers, but late in the laying period, some control birds will go out of production whereas others will be consistent producers until the end of the test. As intratreatment variability increases during the course of the test, the ability to detect treatment effects will be reduced. Although the test was originally designed to produce a large number of eggs per pen so that the estimates for each endpoint were based on a large sample, the 10 week laying period pushes the biological limits of some birds for egg production.

ASTM (1990) guidelines suggest an alternative approach that addresses this situation. It suggests that "it may be unnecessary to collect more eggs than might be laid in the wild with two clutches. For the mallard and bobwhite, egg collections may be terminated when all control pens produce 25 eggs; or 6 weeks after 50% of the control hens have laid one egg". The advantage to this approach is that setting the egg laying period at a length that is within the biological limits of the test species should reduce one source of variability that is unrelated to the pesticide treatment. However, this approach would reduce the numbers of eggs upon which the proportional measures are calculated, which may increase the variability in those endpoints.

On a related issue, some test substances may delay the onset of egg production, even though they may not affect the total number of eggs laid during a laying season. Using the ASTM criteria for defining a shorter laying period for the test may increase the probability of detecting treatment differences due to delayed onset of production. However, greater care must be exercised in the interpretation of biological significance of such an effect. Is the delay in onset of laying biologically significant if reproduction is not otherwise affected?

Role of food consumption--

Some pesticides are known to reduce food consumption at the dietary concentrations that affect reproduction parameters. The reduction in food consumption alone can affect egg production. Because energy requirements for egg formation (34 kcal/bird/day) in bobwhite at 25°C are similar to their basal metabolic requirements (35 kcal/bird/day)(Case 1972), laying bobwhite approximately double their food consumption during egg production. It has been estimated that the daily maximum cost of mallard egg production is 52 to 70% of daily energy intake at constant body weight (King 1973). If a pesticide-treated diet results in reduced food consumption, are observed reductions in egg production the result of direct pesticide effects on reproductive mechanisms or a function only of reduced caloric intake? To address this question, several investigators have used a parallel "pair-fed" test that pairs a treated pen of birds with a pen that receives the same amount of untreated food each day as was consumed by the treated pen. Tests with methamidophos (Stromborg 1986a), methyl parathion (Bennett and Bennett 1990), and monocrotophos (Stromborg 1986b) did not produce evidence of a direct pesticide effect other than could be explained by reduced food consumption alone. Stromborg (1981) found that diazinon produced an effect on egg production beyond that produced by reduced food consumption alone. Rattner et al. (1982) found that parathion reduced bobwhite egg production by directly altering the secretion of reproductive hormones without a significant reduction in food consumption.

The mixed results from these tests with organophosphorus insecticides do not definitively answer the question of the role of food consumption in explaining reproductive effects, but it is clear that pesticide-induced reductions in food consumption can play an important role in affecting egg production. What has not been addressed is how the responses observed in laboratory tests for pesticides with significant reductions in consumption compare to bird responses in the field. It is not known to what extent wild birds may experience the same reductions in food consumption following an exposure to an agricultural application of pesticide or respond to the presence of the pesticide by seeking less contaminated food sources. Some pesticides that are avoided in laboratory tests can still be acutely poisonous to birds in the field (Grue et al. 1983). The laboratory test provides a worst case exposure scenario. Investigators should document the effects of each pesticide on food consumption to aid in the interpretation of reproductive effects data.

Randomization--

Once the number of dietary concentrations, number of birds per pen and number of replicate pens has been decided upon, it is necessary to assign the individual birds to pens and pens to treatment groups. The Subdivision E guidelines (section 70-3) state that "organisms in each test, should as nearly as practicable, be of uniform weight, size and age. Organisms should be randomly assigned to test groups." It further states (section 70-4) that the report should include a "randomization plan for treatments." However, further guidance for carrying out a randomization plan is not discussed.

The random allocation of birds to pens and pens (i.e., experimental units) to treatments is necessary to remove the possibility of systematic error in the estimation of treatment effects (Cox 1958) and will allow for the estimation of experimental error. Note that because the pens are the experimental units (and not the birds), pens must be randomly assigned to treatments. However, it is clear that it is also important to randomly assign the birds to the pens, and in the process, assure that the pens contain birds of uniform weight, size and age. It is also desirable to determine the compatibility of pairs, although this is often not apparent until after the beginning of the test. The incompatibility of birds may result in infertile eggs, injuries, or death of one of the pen mates and is a problem because it increases variability in measurement endpoints from causes that may be unrelated to treatment. One possible solution is to have an acclimation period of sufficient length to identify and remove incompatible pairs from the test population. This should not be done after dietary treatment has begun.

One method of randomizing birds is to divide the pool of test birds into classes of equivalent weight, size and age. Then randomly assign the birds to pens such that there are equivalent distributions of birds based on weight, size and age in each pen. This ensures that a particular treatment group is not dominated by heavy or older birds.

Note that "randomize" does not mean to choose what appears, to the researcher, to be a haphazard order. The careful use of a random order generator such as a table, computer or blind draw is essential. See Cox (1958) for thorough examples of good and bad randomization schemes and a discussion of what can go wrong if randomization is not carried out properly. Further discussion of randomization can be found in Appendix B.

When the entire pool of pens are randomly allocated to treatment groups in one generation of random order, the design is said to be a "completely randomized design (CRD)" (Peterson 1985). This is the type of design that Subdivision E assumes is being used in its discussion of design and in its prescription of a one way analysis of variance to analyze the data. There are however, other randomization plans that may be justified for some circumstances that give rise to a variety of other designs. Two of the most important are called "randomized block designs (RBD)" and "split plot designs (SPD)". It is important to recognize when these designs have been used because the design must be incorporated into the analysis.

Designs other than the CRD can be more efficient and precise at estimating and detecting treatment effects if they are used in the proper context. Their use is mandatory in some circumstances. RBDs are useful when groups of pens are similar within the group but dissimilar between groups; for example, when one group of pens is housed in a separate test room from another group. This design allows the differences between the groups to be accounted for in the analysis but separated from treatment effects. Split plot designs are used when a factorial arrangement of treatment is used; one factor is randomly applied to the experimental units and then the other factor is randomly applied within the first factor assignment.

When using alternative experimental designs, the rationale for why it is more appropriate than the CRD design outlined in Subdivision E should be stated clearly. Such rationales can be incorporated into protocols submitted to EPA prior to the study. Blocking factors and randomization patterns should be well documented. It is also important to recognize when designs other than CRDs have been used, because the one-way analysis will not apply to these cases. The SPD is especially important to note because it involves the generation of two experimental error terms and its analysis is non-trivial. Use of a one way analysis with this design will lead to incorrect inference. Definitions of RBDs and SPDs and descriptions of their respective analyses can be found in Peterson (1985).

Power--

In designing a reproductive study and in presenting the results, it is important to know that the study will be able to detect a meaningful difference between the treatments if a difference exists. This is the statistical concept of power. Technically, power is defined to be the probability of detecting a **statistically** significant treatment effect in the sample when such an effect exists in the population. Note that since "power" is a probability, its calculation requires that the distribution of the response is known or assumed. Alternatively, given the power of a study, one can determine the number of pens **per sample** necessary to detect treatment differences in the population.

In discussing the major issues involved with the avian reproduction test, the Subdivision E guidelines state that the Agency has studied the sensitivity of these avian reproductive tests and has concluded that with the group design "a statistically significant reproductive impairment of 20% or more" can be detected using the recommended number of pens given above. It goes on to state that "for pairs testing to achieve this sensitivity, more than 12 replicates are necessary; calculations have indicated that as many as 25 replicates may be necessary". Section 70-4 states that the "number of pens needed for a particular level of sensitivity" be calculated and the reader is referred to Walpole and Myers (1972) for a method. The ASTM (1990) standard practice for conducting avian reproduction tests is that a test should be designed to detect a difference of 25% with the probability of a Type I error less than 5% in 8 out of 10 such experiments (Power = 0.8). The Office of Pesticide Programs, whenever possible, follows this guidance when evaluating test results.

No references are provided to substantiate the claim by the Agency that 12 replicates are adequate to achieve the desired power. Differences in variability between studies, treatment groups and endpoints may make the "12 replicate" value an oversimplification. One method of assessing this would be to examine power curves and sample sizes from a variety of avian reproduction studies that have already been conducted. Such a study is currently being carried out at the National Wildlife Research Centre of Environment Canada (Pierre Mineau, Wildlife Toxicology Division, pers. comm.). The results of such a study would be extremely valuable in helping the Agency update its expectations of power and detectable differences. It should also be noted that the required number of replicates are theoretically adequate only when the one way ANOVA is the appropriate analysis to use.

Two areas of concern regarding the calculation of power and number of replicates are: 1) the power depends on the variability of the response under consideration and many responses are measured in an avian reproduction study; and 2) these calculations depend on the assumption that the data are obtained from a normal distribution of responses.

Since many reproductive endpoints are analyzed using ANOVA and each endpoint has a different level of variability and the power calculations depend on endpoint variability, the power that is attained for one endpoint will be different from the power for another endpoint. Thus for a single avian reproduction study some endpoints will be more or less useful, in a statistical sense, for detecting significant effects. When designing the study, the most variable endpoint should be considered and the sample size chosen to obtain the desired power, based on that endpoint. Then the proposed design will, at least, achieve the required power for all endpoints. Alternatively, power could be calculated based on a selected endpoint. In this case the choice of endpoint should be well documented and it should be recognized that the power statements will only apply to endpoints that are, at most, as variable as it.

A more worrisome problem has to do with the assumptions of normality and homogeneous variance of treatment groups implicit in the analysis of variance and the assumption, implied in the calculation of power, that the difference between treatments is a shift in the mean value of a normally distributed response. When the response is not normally distributed or when the effect of the treatment is more complicated than just a shift in the average of a normal population, power calculations are not accurate. For example, if eggshell thickness is normally distributed in the population of control birds, but the effect of treatment is not to change the mean but to result in a skewed distribution of eggshell thickness, then the computation of power will not be accurate. As discussed in the section on data analysis, it is important to check that there is reasonable evidence in the data to support the assumption of normality and homogeneity of variance for all treatment groups and control. See Appendix B for methods of checking these assumptions. Note also, that when a transformation of the response is used to achieve the assumptions of homogeneous variance and normality, then the power calculations apply to, and must be carried out on, the transformed scale.

In interpreting the results of the analysis of an avian reproduction study, attention should be paid to the power and the value of the difference from control that was detectable for each endpoint. Although there may be pesticide related effects in a variety of responses, the power may not have been sufficient for detecting it in all of them. Statistically significant effects in some endpoints and not in others may be due to differences in power. In addition, the biological significance needs to be considered. For example, if the power was high and a very small difference from control was detectable and statistically significant in the lowest dietary concentration, it should be considered whether or not that difference is biologically significant. Very small differences may not be biologically meaningful.

Selection of Test Animals

OPP recommends the bobwhite and mallard as the test species for use in avian reproduction studies. The Subdivision E guidelines (USEPA 1982a) state the criteria used for selecting species and the rationale for not including other species at this time. Briefly, the criteria used were: 1) species should have demonstrated sensitivity to pesticides as determined from an extensive data base; 2) species should be ecologically significant; 3) species should be aesthetically or economically important; 4) species should be readily available for test purposes; and 5) species should have characteristics that are appropriate for this type of test. There is little data available to demonstrate that laboratory tests with songbirds would be more useful than tests with mallards and bobwhite for predicting the reproductive effects of a pesticide to all exposed bird species. However, given their smaller body size and higher metabolic rates, songbirds may be more vulnerable to pesticide exposures during this energetically demanding period. Based on acute toxicity testing, many songbird species are more sensitive to pesticide exposures than bobwhite or mallards (Schafer and Brunton 1979). Much more work is needed to evaluate the relative sensitivity of birds to pesticide exposures during reproduction and to develop songbird reproductive test methods, if appropriate.

The Subdivision E recommendation for test animals used is "pen-reared birds, previously untreated, approaching their first breeding season, and phenotypically indistinguishable from wild birds" (USEPA 1982a). Birds may originate from research breeding colonies or from gamebird vendors. If birds are shipped, they should be examined for physical injuries, and it is advisable to have a 2- to 6-week health observation period prior to the start of the test. The state of health of shipped birds can also be assessed by necropsy of representative birds at a diagnostic laboratory at the time of arrival. Bennett and Fairbrother (unpubl. data) compared the health of mallards from three commercial game farms and found birds to be generally healthy and free of parasites and contagions, but all had low levels of organochlorines and heavy metals in their tissues.

The ASTM (1990) standard practice for avian reproduction tests makes several additional recommendations that are intended to reduce sources of variability that are unrelated to treatment effects. First, birds must come from one source and strain. Second, the age of birds should be within ±10% of the mean age of the test population. Third, birds must be rejected for test purposes if they are deformed, in poor physical condition, or different in plumage from wild birds.

Although test birds may appear to be "phenotypically indistinguishable from wild birds," it is extremely difficult to judge their genotypic comparability to wild birds. One problem that can arise is that phenotypically indistinguishable adults may produce young that clearly do not look like wild stock. Gile and Meyers (1986) found in a study with mallards that 3 out of 24 pairs produced white ducklings, all of which were males. These white ducklings were considerably larger than normal-color ducklings. The USEPA Environmental Research Laboratory-Corvallis has also observed gamefarm-reared bobwhite that produced nonnormal colored chicks in a reproduction test. This genetic variation may be a confounding factor that introduces greater uncertainty to the validity of test results, especially if the atypical young differ in body weight, growth, or survival rates.

Selection of Testing Environment and Husbandry

All birds should be housed in breeding pens of adequate size for the species and number of birds per pen. Minimum floor space specifications for quail are 0.023 m^2 per bird, and for 0.5 to 1.5 kg chickens are 0.093 m^2 (National Institutes of Health 1985). Similar NIH standards do not exist for mallards. However, all birds must be provided with sufficient head room to stand erect and wide enough to fully extend the wings. The Canadian Council on Animal Care (1984) specifies that mallards in a single pen have a minimum space requirement of 0.33 m^2 floor area with a 35 cm height and small groups should have 0.66 m^2 floor space per bird. ASTM (1990) provides a comparative table of pen sizes and numbers of birds per pen for several species from the published literature. For other recommendations on housing and care of laboratory animals, see the "Guide for Care and Use of Laboratory Animals" (National Institutes of Health 1985). Control of test room temperature and humidity are desirable and should be recorded throughout the test period. The recommended environmental conditions are 21°C and 55% relative humidity (USEPA 1982a). The NIH recommends a room ventilation rate of 10 to 15 room air exchanges per hour (National Institutes of Health 1985).

The photoperiod throughout the test is extremely important to the success of a test. It is recommended in Subdivision E that birds be maintained under 7 hr light: 17 hr dark each day during the first eight weeks of the test. The dark period should not be interrupted during this period. As little as 15 minutes of light exposure during the dark period can cause increased gonadal development (Kirkpatrick 1955). Bobwhite will become sexually active if given 60 minutes exposure to light during the dark period. Consequently, during the first eight weeks of the test, the dark period of each day must be maintained to keep birds in nonreproductive condition.

After eight weeks of the short day length, the photoperiod should be increased to 16 to 17 hours of light per day. This increase is often accomplished in one step, but sometimes bobwhite become aggressive after a sudden increase in day length, causing injuries to pen mates (Bennett et al. 1990b). An alternative method is to increase day length gradually over a two week period to reduce aggressiveness. The Subdivision E guidelines also suggest that the light period can be maintained either at a constant duration or increased by 15 minutes each week until the end of the test. Unless there is a documented benefit to continuously increasing the light period, it may be better to maintain a constant maximum light period, rather than adding a potentially confounding variable.

The Subdivision E guidelines state that a lighting intensity of 6 footcandles at the bird level is adequate. Lighting intensity seems to be less critical than day length to reproductive performance. Bobwhite reproductive performance was similar at 0.1, 1, 10, and 100 footcandles for a 17 hour light period (Kirkpatrick 1955). The Subdivision E guidelines recommend that shorter wavelength "cool white" fluorescent lights that do not emit the daylight spectrum should be avoided.

Food and water should be provided <u>ad libitum</u> throughout the test. Mallards consume large amounts of water and there are several ways of providing it. Water can be provided by regularly filling a static water bowl, using a continuously flowing water system (Bennett et al. 1991, Heath et al. 1969, Heinz 1974), or using a licking valve (Bennett et al. 1990a). Licking valves have the advantage of providing clean water on demand while reducing the flow of water, but birds need to be provided a bowl of water at least twice a week so they can clear accumulated matter around the nares.

Egg Collection, Storage and Incubation

It is recommended in the Subdivision E guidelines that eggs be collected daily, marked with the pen number from which they were collected, and stored at 16°C and 65% relative humidity. ASTM (1990) recommends storage at 12 to 16°C. Eggs should be set in an incubator at weekly intervals. Incubator temperature, humidity, and egg turning rate are not specified in Subdivision E. Mallard eggs have been incubated at 37.4 to 37.5°C with relative humidity ranging from 62 to 80% (Heinz 1976a, 1976b, Greenwood 1975, Holmes et al. 1978). Prince et al. (1969) found that the incubation time of mallards eggs decreased as incubation temperature increased from 35.6 to 39.4°C, with the highest percent hatch occurring at 37.5°C and relative humidity of 70 to 80%. Flegal and Sheppard (1976) recommend that quail eggs be incubated at 37.6°C and 60% relative humidity.

The Subdivision E guidelines recommend that eggs laid on one day every two weeks should be collected for measuring eggshell thickness. It is recommended that eggs be candled at day 0 of incubation to check for eggshell cracks, approximately half way through incubation to determine early embryo mortality, and approximately three quarters through incubation to measure embryo survival.

The rate of egg cracking can be sufficiently variable among tests that it warrants discussion. Eggs can be cracked intentionally or unintentionally by the birds or accidentally by test personnel or equipment at any point between laying and hatching. Cracked eggs typically are discarded from further analysis because of the adverse effect most cracks have on embryonic development. The Subdivision E guidelines state that the proportion of eggs laid that are cracked should be statistically analyzed to determine treatment effects. This represents one measure of eggshell quality. However, there are several factors, unrelated to the pesticide treatment, that may produce confounding influences on the sensitivity of egg cracking as a measure of eggshell quality. Non-treatment egg cracking can be primarily a function of the physical test system, with key factors being cage materials, slope of cage floor, and access of birds to eggs. If cracked eggs originate primarily from a small number of pens within each treatment group, this may be a reflection of the abnormal reproductive behaviors of some birds in reaction to the physical test system. Consequently, test systems with a low rate of background egg cracking may have a different sensitivity for identifying treatment-related effects on cracking than test systems with a high rate of background cracking, especially when the distribution of cracked eggs is concentrated in a fraction of the pens (See example in Appendix C). Consequently, the rate of cracked eggs may be a very poor indicator of treatment effects on eggshell quality.

Pesticide-related changes in eggshell quality are determined more directly and efficiently by measuring eggshell thickness or other measures of eggshell quality, rather than using the indirect measure of cracked eggs. The higher the rate of cracked eggs, the greater the confounding influence on other measures of reproductive success. The proportional measures of fertility, embryonic development, hatchability, and survival will be affected by reducing the number of eggs on which the proportion is based. The greatest confounding effect may be in reducing the number of 14-day-old survivors per pen for reasons unrelated to the test substance. The higher the rate of cracked eggs, the less meaningful this variable becomes as a measure of treatment effects. Incidental cracking of eggs is a nuisance variable that should be minimized to the extent possible. The rate of eggshell cracking may be more indicative of the quality of the physical test system than as a indicator of treatment-related eggshell quality. Rates of egg cracking should be used primarily to determine if cracking rates in control groups are within acceptable criteria established by OPP from historical data.

The Subdivision E guidelines recommend that at day 21 of incubation for bobwhite and day 23 for mallards the eggs are transferred to a separate nonturning hatcher or hatching incubator set at 39°C and 70% relative humidity. Flegal and Sheppard (1976) recommends that hatchers be set at 37.2°C and 85% relative humidity for bobwhite and pheasant. Stromborg (1986a, b) used a hatcher temperature of 37.6°C for bobwhite. Heinz (1974) used a hatcher set at 37.5°C for mallards. The incubator trays should be partitioned so that hatchlings can be identified as to parental pen.

Observations of Progeny

The Subdivision E guidelines recommend that hatchlings be removed from the hatcher at day 24 of incubation for bobwhite and day 27 for mallards and housed in brooders according to parental pens for 14 days on a control diet. The rate of hatchability and hatchling survival to 14 days of age should be recorded for each pen. Hatchlings are weighed at 14 days of age (or older if mortality is observed late in the observation period). Since the analysis of weights of 14-day-old survivors is conducted on the mean weight for each parental pen, either individual weights for all hatchlings in a season or the average of weighted weekly means are required to calculate the mean weight properly. However, individual weights would allow investigators to examine the distribution of weights for each pen and compare distributions among treatments.

The method of segregating hatchlings by parental groups as a means of determining survival rates and body weights for each parental pen requires several assumptions. First, it has to be assumed that the rate of survival is equal for hatchlings housed singly (i.e., only one egg hatched for a pen) compared to those housed in groups. Second, it must be assumed that there are no differences in the environmental conditions for housing chicks among all the parental pens. Third, it must be assumed that the system will not allow hatchlings from one parental pen to escape into a section holding hatchlings of another. If these assumption are not met, the determination of percent survival will be confounded by variables unrelated to the pesticide treatment.

An alternative approach is to individually mark hatchlings (e.g., with numbered wing tags or leg bands), record their parental pen number, and house hatchlings in common brooders. This method adds additional requirements to individually mark hatchlings and assumes that they will not lose their identification markers, but standardizes group size and environmental conditions to reduce potentially confounding variables that could affect hatchling survival.

Data Analysis and Interpretation

The Subdivision E guidelines (section 71.4) require that "for continuous variables, experimental groups should be compared to controls by analysis of variance. For most discrete variables, survival percentages should be computed and arc sine transformed prior to analysis. Alternatively, a chi square analysis of survival (contingency table) may be used". In addition, Section 70.4 states that, in addition to the data at each treatment level, LD50, LC50, EC50 and 95 percent confidence intervals be calculated when "sufficient doses and test organisms are used to establish a dose-response line" and that "no observed effects levels" and statistical methods used and a reference to them in published literature be provided for each avian reproduction test.

To assist in the evaluation of test reports, OPP has prepared a SAS (1985) program that is currently used to carry out a one way analysis of variance (ANOVA) on some response variables and a one way weighted ANOVA on arc sine square root transformed survival proportions, where the weights are the denominator of the survival proportions. The current statistical analysis process involves running this program on data submitted for review and assessing the statistical significance of the effects as provided by the program.

The chi-square contingency analysis that the Subdivision E guidelines also suggest for analyzing discrete response variables does not seem to be widely used. This may be due to the fact this analyses can lead to the conclusion that an association exists but there are no conclusions as to the pattern of that association. It is not an informative analysis. Generalized linear models, discussed in Section 3, can provide an analysis that will describe the patterns of dependency in this type of data.

Use of the Analysis of Variance (ANOVA)--

Because an analysis of variance, or any statistical analysis, is based on assumptions, the analysis is valid only to the extent that the assumptions are justified. It is important that the assumptions be checked because some assumptions may not be justified for some responses obtained from reproduction tests.

The ANOVA and associated F-tests assume: 1) the data are sampled from a normal distribution of response such that although the mean response of the treatment groups may be different; 2) the variability is assumed to be the same from group to group (homogeneity of variance); and 3) the effects of the treatment will be to change the mean response by the addition of a quantity due to the treatment while the variability will remain unchanged. This additive quantity is called the treatment effect.

The assumption of additive treatment effects, the third assumption above, can not always be checked using the data. The experience of the scientist and the biology of the problem will often bear on this decision.

The ANOVA is known to be robust to departures from the assumption of normality (Miller 1986). That is, the p-values obtained from the ANOVA F-test for treatment effect will provide the correct conclusion from the data even if the distribution of the response is somewhat nonnormal. However, although valid, these tests may not be the most powerful tests for the nonnormal distribution. Thus verifying an approximate normal distribution or using a transformation to help achieve this will ensure that powerful tests are being employed.

In order to check the first assumption above, it is important to recognize what types of responses may or may not be normally distributed. First, continuous measurements, which may take on any positive or negative value can be normally distributed. Some responses, such as weight which usually cannot be negative, can be normally distributed if the average value of the response is not close to 0. Also, some discrete variables (which can only take on integer values) can be thought of as having normal distributions if the average count is large, say 25 or more. This is the reason that counts of numbers of eggs laid per pen can be analyzed using ANOVA, if the number of eggs laid per pen is large.

Generally, ratios, proportions, small counts of individuals or responses that differ by orders of magnitudes, will not be normally distributed. In addition, these types of responses will have variances associated with them that depend on the mean value of the response and thus the ANOVA assumption of homogeneous variance is almost always violated for these types of responses. However, there can be continuous variables which do not have normal distributions and proportions, counts or ratios which do. It is important to evaluate the shape of the distribution and the extent to which variability is the same from treatment group to treatment group on a case by case basis.

The distributions of all endpoints to be analyzed should be examined prior to analysis using histograms, box plots or scatter plots (as described in Appendix B), in order to decide if the first ANOVA assumption is justified. The plots should be obtained for each treatment group including control. The distributions should be approximately symmetrical, tapering off evenly in both directions from a single mode. Examination of the distributions of the data can also provide justification of the second assumption. If the range over which the data are spread out is approximately the same for each group then the homogeneous variance assumption can be justified. However, the authors have seen cases of toxicological data where this is not the case.

If an ANOVA has already been carried out then residuals from the ANOVA can be plotted to check assumptions. Residuals (datum minus the treatment group average) from all treatment groups should be centered around a value of zero and the variability should be approximately the same in all groups. In SAS (SAS 1985) residuals can be obtained in PROC GLM using an OUTPUT statement. If the assumption of homogeneous variance is not justified, then statistical tests may not be carried out at the intended level and incorrect inference can occur. Wetherill (1981) and Miller (1986) discuss how the conclusions obtained from ANOVA can be affected when the variance is not equal between treatment groups.

It is often suggested that Bartlett's test (Snedecor and Cochran 1980) for homogeneity of variance be used to check the second assumption, but it is known that Bartlett's test is very sensitive to departures from normality (Miller 1986). Levene's test (Levene 1960) is a simple and robust alternative to Bartlett's test and it is described in Appendix B.

When the number of pens per treatment is small, it can be difficult to decide if the data are normally distributed or if the variability is the same between groups. In such cases it is suggested that good scientific judgement, as well as scatter plots and histograms, be employed. In addition, the advice of a trained statistician can be very helpful.

Transformations of the Data--

Although some types of responses may not appear to satisfy the assumptions of an ANOVA, there may be a transformation of the response which does. That is, on a different scale, the ANOVA assumptions are justified. Some transformations have been developed for specific types of problematic data. For example, the arc sine square root transformation of proportions is typically suggested for adjusting the heterogeneous variance of proportions. The square root transformation for counts of individuals is another example developed for the same reason.

The Subdivision E guidelines suggest that the arc sine square root transformation be applied to all proportions and then ANOVA be carried out on the transformed variables. Although these transformations may be helpful in making the distributions of the data more bell-shaped, transformations will not necessarily achieve normality, homogeneous variance nor additivity of effects in one fell swoop. McCullagh and Nelder (1989) and Atkinson (1985) discuss this point and acknowledge that frequently a transformation of the data is not the best solution to the problem. Therefore, once the data have been transformed, it is imperative to recheck the data in order to verify that the assumptions are justified on the new scale.

The arc sine square root transformation is suggested for the proportions obtained in an avian reproduction study, but if the transformation does not help to achieve normality and homogeneity of variance, one recourse is to investigate other simple transformations. Box and Cox (1964) developed a family of transformations from which a choice of the most appropriate transformation for achieving normality and homogeneous variance could be chosen based on the data. Some common transformations that might be used for counts or proportions, other than the arc sine square root transformation, include taking the square root or the logarithm of the response. One problem with using the logarithmic transformation for positive data is that data values of zero need to be recoded since it is clear that they do not belong at minus infinity (log(0)). One approach, similar to a method suggested by Tukey et al. (1985) is to choose Y^*, the recoded value of log(0), such that:

$$[\mathrm{Log}(Y_1) - Y^*]/[\mathrm{Log}(Y2) - \mathrm{Log}(Y1)] = [Y_1 - 0]/[Y_1 - Y_2]$$

where Y_1 is the smallest datum other than 0 and Y_2 is the next smallest datum.

In some cases, even on a transformed scale, the assumptions for ANOVA cannot be justified. If these data are analyzed using ANOVA, the conclusions that are reached may be incorrect and means and variances may not be well estimated. A practical approach to an analysis in which there is some question as to the validity of the assumption is to carry out the analysis but provide written documentation of the inconsistencies in assumptions or data. Make it clear that the conclusions reached from the statistical analysis are valid only to the extent that the assumptions are justified.

There may be cases when the failure of the ANOVA assumptions leads a scientist to look toward methods other than ANOVA such as non-parametric statistics (Daniel 1978) or generalized linear models (McCullagh and Nelder 1989) such as logit regression or log linear models or a trend test. These methods and other motivations for them are discussed in Section 3.

Alternative Designs--

As discussed previously, it is important that the statistical analysis used reflect the design of the experiment. It is possible that designs other than a Completely Randomized Design (CRD), as described in Section 2, are judged to be the more appropriate for the test in a given laboratory.

For example, Randomized Block Designs (RBD) could be used to reduce variability due to using multiple testing chambers. It is important that the use of alternative designs be clearly stated and justified so that test data can be evaluated by the Agency using the correct statistical analyses. It is quite possible for the data from a well designed RBD study, in which effects are pronounced, to show no statistically significant effect when analyzed as if it was a CRD.

As another example, data tabulated from a design that results in multiple error terms and requires a special analysis such as a split plot design, may not appear any different from data tabulated from a CRD. The difference lies within the design description. Analyzing this type of data with a one way ANOVA would lead to statistical tests that use the wrong error terms and hence results would be highly suspect.

Therefore, it is crucial that the statistical analysis be carried out based on the design that was used to generate the data. Test protocols should be broadened to include provisions for basic experimental designs other than the completely randomized design and scientists should be aware of alternative analyses required by these designs. There is no reason to believe that the CRD is the best or most efficient design for all situations.

Interpretation of Results--

Along with using statistical methods to identify significant differences between control and treatment groups, the biological reasons for such differences also should be considered. Finding a statistically significant effect does not necessarily represent a biological effect. For example, there may be unmeasured confounding variables, unrelated to the pesticide effect, that contribute to the apparent significant difference. Or in the case of a large number of replications, the statistical power may be great enough that treatment differences may be identified as statistically significant even though they may not be large enough to be of biological significance. On the other hand, the lack of a significant effect does not mean that there is no information in the data about pesticide-related effects. Random variation in biological studies can be quite large and this may make it difficult to detect statistically significant effects even when steps have been taken to design a powerful study. When trends or differences between treatment groups are clearly noticeable, even if they are not significant, their presence should be noted and this information added to the "weight of evidence" for the pesticide in question.

The biological significance of apparent treatment effects can be explored through a series of questions posed after the study has been completed. Is the relationship between patterns of effects and dietary concentrations explainable in the context of our understanding of dose/response relationships? Are effects observed in one measurement variable also observed in other related variables? For example, if there is a treatment-related decrease in the number of eggs produced, it is possible that there would also be a treatment-related difference in the number of 14-day-old juveniles produced, if the test has sufficient statistical power. If not, is there a reason why the outcomes for the variables are different? Are there mitigating factors or does the increase in variability with each successive step during development mask the effects of treatment? Are the differences observed in quantitative measurements supported by the qualitative observations in the test? For example, are observations of behavioral abnormalities in adults correlated with measurements such as food consumption or egg production?

It is also important to ask questions of data sets where no statistically significant differences were detected. Are there trends in response variables that are not statistically significant, but may represent biologically significant effects? If so, would effects be significant using a more powerful experimental design? Are there qualitative observations of pesticide effects in adults or juveniles?

The benefit to this interpretation is the identification of the questions that have been answered, questions that have remained unanswered or unclear, and questions that have been generated from test results that merit further consideration. If further testing is required to develop the ecological risk assessement, it is very important to identify the questions of concern arising from the avian reproduction test. Identifying the questions of greatest concern will help to determine the most effective method for addressing the questions, such as repeating the laboratory test, using a specialized laboratory test, or conducting a field test.

SECTION 3

ALTERNATIVE APPROACHES FOR EVALUATING REPRODUCTION EFFECTS

There are many approaches that could be taken in evaluating the effects of pesticides on avian reproduction. However, there are several important considerations in determining when alternative methods would be appropriate or acceptable for supporting a regulatory action. There is a need for consistency in testing methods for comparability of results. Consequently, to achieve consistency, it is desirable to have a "standard" protocol used for testing all chemicals. On the other hand, there is a need for flexibility in testing methods because some questions or data sets may not be appropriate for a standard protocol. In these cases there should be strong scientific justification for using an alternative approach that is suited to the question being addressed or the data set obtained. The use of alternative methods should be clearly explained and justified as to why they are superior to methods recommended in the Subdivision E guidelines.

Alternative approaches would be justified and useful in situations where previous experience (i.e., experience with related chemical or similar data distribution) indicates that a new or existing pesticide may pose reproductive effects that are not adequately tested and evaluated using the current guidelines. Because avian reproduction tests are used primarily as a toxicity screening tool, test approaches must be broad enough in scope to identify potential effects from a variety of pesticides, unless they are designed for a specific regulatory question or specific chemical. Especially in the case of existing pesticides where considerable information may be available concerning both the laboratory and field effects, alternative test protocols may be clearly indicated to address specific questions to support a pesticide registration. The following section discusses several alternative approaches to avian reproduction testing, with special attention given to the advantages and disadvantages of the alternatives and guidance on the situations in which the alternative may be appropriate.

ALTERNATIVE TEST PROTOCOLS

Determination of a Dose-response Relationship

When possible, a statistical analysis of quantitative responses from an avian reproduction study is desired. Under the current Subdivision E guidelines, this analysis results in estimating the mean response for each treatment group, a description of which treatment groups are statistically different from control, and the identification of a NOAEL and LOAEL dose level. Although the concepts of a NOAEL and LOAEL are easily understood, they are scientifically difficult to define and measure accurately. What is identified as a LOAEL under one experimental design may be the NOAEL under a similar design with a different choice of dietary concentrations or replicates.

More information about the pattern of response over the range of dietary concentrations and the potential effects of pesticides could be obtained by using a design and analysis for the determination of a dose-response relationship. The 'dose' in this case, would be the dietary concentration used in the study and the 'response' could be any of the quantitative responses listed in Section 2. The test design would include a range of several dietary concentrations, including control, that produce effects on reproductive parameters for calculating the dose-response relationship. This approach could be comparable to the dose-response approaches used for measuring relative toxicity (e.g., LD50, LC50).

The advantages to this approach are: 1) the relative distance between effect concentrations and the expected environmental concentration for a proposed use could be established; 2) the types of reproductive effects caused by a pesticide would be identified; 3) the pattern of toxicity over the dietary concentrations could be established; 4) a more accurate definition of the NOAEL and LOAEL; and 5) the ability to calculate additional statistics that help define the pattern of toxicity such as EC50's and the slope of the curve defining the dose-response.

As was discussed in Section 2, if a test failed to detect reproductive effects using current methods, nothing would have been learned about the types of effects that might have been observed under different exposure scenarios. Conversely, if reproductive effects are observed in all treatment groups, it may not be possible to determine if the same or different effects would be observed at lower concentrations. The alternative approach would circumvent these problems.

The disadvantages of this dose-response approach are: 1) a range finding test or tests would be required to establish appropriate dietary concentrations for a definitive test; and 2) the determination of tissue and egg residues may be less meaningful if the test was conducted at concentrations considerably different from the EEC. Another possible disadvantage is that, as the number of dietary concentrations increases and the number of replicate pens per treatment group is held constant, the size of the test will increase. However a trade-off may be made by reducing the number of pens per treatment. The number of replicate pens per treatment may be different from the currently prescribed number if an analysis other than the one way ANOVA is used (e.g., maximum likelihood methods).

The appropriate choice of the number of treatment groups and the number of replicates per treatment group depends on which statistics are to be obtained and the data analysis method to be used. For example, if it is desired to estimate the EC50 then the dose (dietary concentration) groups should be centered around the likely values of the EC50. If a logit regression analysis is to be used and the endpoint is the proportion of hatchlings out of eggs set per pen, then it is important to have a large number of eggs set over all pens in the treatment group than to have simply many pens. A statistician should be consulted at this point. Designing dose-response studies is not always the same as designing the type of study defined under Subdivision E. Analysis methods can include simple and multiple regression for normally-distributed response variables, logit or probit regression for proportions and log-linear models for data in the form of integer counts.

The dietary concentrations used in the study should be spread out over the range of response that is of interest, i.e., extrapolating the dose-response beyond the range of the data is statistically dangerous. Increasing the number of dietary concentrations used will result in more precise statistics.

However, some statistical analyses that do not rely on assumptions of normal distributions do not have the same types of requirements concerning replication as does the analysis of variance (see the example two paragraphs above). In addition, power calculations relevant to an ANOVA may not apply in these cases. It is important not to assume that the methods defined in Subdivision E carry over exactly to an alternative type of study. Seek the advice of a trained statistician when nonstandard methods are used.

Test for Effects from Short-term Exposure

The current avian reproduction test is designed to determine the effects of pesticides with chronic exposure patterns (i.e., continuous or repeated exposure) on reproductive mechanisms. Since the development of reproductive capacity in birds begins months before initiation of egg laying (Kirkpatrick 1959), the test originally was designed to start treated diets well in advance of laying because of the bioaccumulative properties of the organochlorine insecticides. However, many newer pesticides are much less persistent in the environment and their use patterns are such that the initial contact with these pesticides may come at any time during reproduction. The current test protocol using a chronic exposure period may not effectively identify potential effects of less persistent pesticides. An alternative method is to utilize a shorter exposure period initiated after the test population is in egg production. This approach is mentioned in the ASTM guidelines (1990), though not discussed. Bobwhite tests with organophosphorus (OP) insecticides have shown significant reproductive effects with treatment periods of 8 days (Bennett and Bennett 1990), 10 days (Rattner et al. 1982), and 3 weeks (Stromborg 1981, 1986a, 1986b, and Bennett et al. 1990b). A mallard test with an OP insecticide used a treatment period of 8 days and showed a pesticide-related response (Bennett et al. 1991). The length of tests using shorter exposure periods has not been standardized, but one approach would be to use environmental chemistry data to set the test length as a function of the environmental degradation rate.

Several of the organophosphorus and carbamate insecticides have been shown to significantly reduce egg production within days of initiation of dietary exposure (Bennett and Bennett 1990, Bennett et al. 1991). Changes in eggshell quality also have been observed within days after treatment (Bennett and Bennett 1990). Further, because many of the less persistent pesticides will have shorter residual times in the environment, they may not meet the current criteria for initiating an avian reproduction test, even though they present a potential hazard to reproduction from short-term exposures. Therefore, there is a need to establish new criteria to initiate a short-term test. Given the relationship between the amount of food consumption and the rate of egg production, one possible criterion is testing pesticides that produce significant reduction in food consumption (e.g., >50%) in the higher concentrations of an acceptable avian dietary toxicity test (i.e. LC50 test). Since the relationship between pesticide concentrations that affect food consumption of juveniles and those that affect reproduction in adults is poorly understood, an adult feeding test may be required to determine the effect levels in reproducing adults. It is also unclear if all chemicals that reduce the food consumption of juvenile birds will result in effects on reproduction.

There are several potential advantages to using a short-term exposure test. First, the test can be conducted with known layers of fertile eggs, thus reducing the variability in test data by not including pairs that produce no eggs or only infertile eggs. Second, the pretreatment values for measurement endpoints can be used as controls for each pen (i.e., covariates). For example, birds typically lay eggs of consistent size, shape, and eggshell strength (Thompson et al. 1983), making the use of pretreatment values from each bird as covariates a useful method for reducing between-hen variation. Finally, the test is timed to coincide with the maximum egg production. All three of these points will reduce the variability from factors that are unrelated to the pesticide treatment and increase the chances of detecting reproductive effects if they exist. This approach could also be used in conjunction with the dose-response alternative described above.

A disadvantage of this approach is that it would not detect a pesticide-related delay in the onset of laying because treatment would start during the laying period. Also, effects may be delayed or the severity of the effects may increase during the treatment period, so a short treatment period could not assess delayed effects. Effects that are first expressed near the end of the treatment period may be overlooked if the data analysis is conducted only on mean values for the entire treatment period. The short-term test should not be used with pesticides with slow-developing or delayed effects, since the current test guidelines would provide a better approach.

Parental Incubation

Because the current test guidelines (USEPA 1982a) recommend artificial incubation, parental influences on hatching and survival of juveniles are eliminated. The test does not provide information about pesticide effects on parental behavior after oviposition. However, many pesticides have been observed to affect parental behaviors that lead to nest or brood abandonments, abnormal parental care or other behaviors that adversely affect production of juveniles.

There are several field examples of pesticide-related behavioral effects during incubation. Forest spraying operations with fenitrothion resulted in disruption of incubation and nest abandonment by white-throated sparrows (*Zonotrichia albicollis*) (Busby et al. 1990); reproductive success in a sprayed forest area was one-third of that observed in a control area. Laughing gulls (Larus atricilla) dosed with parathion incubated significantly less time during the following three days than controls (White et al. 1983), although no effects on nest defense behavior or hatching success were observed (King et al. 1984). Fyfe et al. (1976) found that merlin (Falco columbarius) nests with high DDE residues in the eggs were deserted more and defended less than nests with lower residues, although Fox and Donald (1980) concluded that the reduced nest defense associated with high DDE egg concentrations was of minor importance in explaining nest failures. Fox et al. (1978) hypothesized that pollutant-induced endocrine dysfunction in Lake Ontario herring gulls (Larus argentatus) may have caused reduced nest attentiveness and defense.

Insecticides have also been observed to adversely affect interactions of parents with their young. Female starlings dosed with dicrotophos made significantly fewer sorties to feed their broods and were absent for longer periods of time than control females (Grue et al. 1982). Meyers and Gile (1986) found that mallards breeding on ponds supplemented with food treated with 80 ppm chlorpyrifos did not produce any ducklings surviving to 7 days old, partially because the hens were not attentive to the broods, while control ponds produced six to eight ducklings per hen. Brewer et al. (1988) observed brood abandonment by wood duck (Aix sponsa) and blue-winged teal (Anas discors) hens after agricultural applications of methyl parathion.

Several studies have been conducted with captive ducks using parental incubation to more closely simulate field situations (Bennett et al. 1991, Custer and Heinz 1980, Finley and Stendell 1978, Franson et al. 1983, Haseltine et al. 1980, Longcore and Samson 1973). Black ducks (Anas rubripes) fed 10 ppm DDE produced eggshells 22% thinner at the equator than controls, with a cracking rate of naturally incubated eggs that was fourfold greater than artificially incubated eggs (Longcore and Samson 1973). Breeding black ducks fed 3 ppm mercury appeared hyperactive, and fewer incubated their clutches (Finley and Stendell 1978). A range of responses were observed in 23 mallard hens exposed to dietary methyl parathion for 8 days either early or late in incubation (Bennett et al. 1991). No response to treatment was observed in 6 hens that successfully hatched broods compared to 7 hens that abandoned clutches, 6 hens that exhibited reduced nest attentiveness by leaving their nests for extended periods on one or more days, and 4 hens that died while incubating. An additional treatment group exposed to methyl parathion for 8 days during the egg laying period was observed to either begin incubating immediately after the end of the treatment period or resume egg laying after treatment by either completing their clutch or abandoning their first clutch and starting a new nest (Bennett et al. 1991). Even though this group was treated only during egg laying, these kinds of reproductive behaviors could only be observed in a test where birds were allowed to incubate their own eggs.

Parental incubation tests using pesticides and other environmental contaminants also have been conducted with several other species, such as ring doves (Stretopelia risoria) (Haegele and Hudson 1973, McArthur et al. 1983, Peakall et al. 1972, Peakall and Peakall 1973), Bengalese finch (Lonchura striata) (Jefferies 1971), screech owls (Otus asio) (McLane and Hughes 1980), barn owls (Tyto alba) (Mendenhall et al. 1983), and kestrels (Falco sparverius) (Porter and Wiemeyer 1969, Wiemeyer and Porter 1970). Ring doves fed dietary concentrations of 10 ppm polychlorinated biphenyl (Aroclor 1254) were less attentive of their nests than controls and produced fewer hatchlings than eggs that were artificially incubated from treated birds (Peakall and Peakall 1973).

The primary advantage to using parental incubation is that pesticide effects on parental behavior can be observed at the transition from laying to incubation and during incubation and brood rearing. Tests using artificial incubation can not identify pesticides that induce birds to: 1) incubate smaller than normal clutches; 2) continue laying excessively large clutches without incubating; 3) abandon nests or broods; or 4) reduce parental care for the nests or broods. These behaviors could have as great an effect on the production of young as any measurements made in the current avian reproduction test.

There are several disadvantages to using parental incubation. First, variability among birds within treatments may increase requiring substantially larger numbers of replicate pens per treatment to achieve the same power in the test as using artificial incubation. By keeping each hen as the incubator instead of using one mechanized incubator, incubation behavior of individual hens becomes a factor that can add significant variability to reproduction parameters in all treatment groups. Not all control hens that initiate clutches will begin incubation (Bennett et al. 1991, Haseltine et al. 1980). Not all control hens that begin incubation will hatch young (Bennett et al. 1991, Haseltine et al. 1980, Finley and Stendell 1978). Due to the variability in incubation behavior among individuals, hatching percentages can range from 0 to 100% in control groups. Bennett et al. (1991) found that 35 mallard pairs per treatment would have been required to detect the 57% decrease in ducklings in one treatment group as statistically different with a Type I error rate of less than 5%. Second, different pens may be required to induce hens to initiate a nest and incubate. Besides providing nest bowls or boxes, additional space and a greater degree of isolation may be required for hens to behave normally. This is especially true for bobwhite, which are more sensitive than mallards to human contact. Third, fewer eggs would be laid per pen so that proportional values (i.e., fertility, hatchability, and survival) would be calculated from smaller samples.

Parental incubation tests should be used when the concern for a pesticide is primarily with effects on adult behavior. It can be an effective method for identifying potential effects during incubation and brood rearing prior to attempting a field study of reproductive effects. However, investigators conducting tests using parental incubation should justify clearly the rationale for using this method and the experimental design chosen.

ALTERNATIVE ENDPOINTS

Are there other endpoints that should be considered when conducting an avian reproduction test? The value of alternative endpoints should be judged based on satisfying one or more of the following criteria: 1) the endpoint detects reproductive effects that are not detected by current methodology; 2) the endpoint is more sensitive at detecting effects than existing endpoints; and 3) the endpoint provides important insight into exposure patterns or mode of action of the pesticide. There is a variety of biochemical endpoints (e.g., serum chemistry, enzyme activity, endocrine function) that may be affected by particular pesticides or classes of pesticides at the concentrations that affect other reproductive endpoints, although they do not represent reproductive effects themselves. These endpoints can be very useful in understanding individual differences in sensitivity and identifying causes for other observed effects. However, it is beyond the scope of this report to address the broad range of possible endpoints. This section discusses a few alternative endpoints that can be useful in explaining reproductive effects and may address the criteria above.

Eggshell Strength

Eggshell thickness was the first measurement used to determine the effects of environmental contaminants on avian eggshell quality (Hickey and Anderson 1968, Ratcliffe 1967). The quality of an eggshell is reflected in its ability to maintain a protective environment for the developing embryo by resisting cracking or puncture and is a function of its thickness, ultrastructural characteristics, and the size and shape of the egg. While eggshell thickness is an extremely valuable measure of quality, it may not be the most sensitive indicator of adverse effects on eggshell quality posed by all chemicals. Eggs of mallards exposed to dietary DDE had reduced shell strength at lower dietary concentrations than were observed to affect shell thickness (Carlisle et al. 1986). Bobwhite exposed to sulfanilamide produced eggs with reduced breaking strength without reducing shell thickness (Bennett et al. 1988). Scanning electron microscopy of the eggshells indicated changes in their ultrastructure, characterized by poorly formed mammillae. In eggs collected from wild white-faced ibis (Plegadis chihi), eggshell strength decreased to a greater extent than shell thickness with increasing DDE residues in the yolks (Henny and Bennett 1990). In addition to DDE, many other classes of pesticides have been shown to reduce eggshell thickness under laboratory conditions (Haegele and Tucker 1974, Bennett and Bennett 1990), but few studies of contaminant effects have also measured eggshell strength. Until more is known about the effect of chemicals on eggshell strength, it would be a useful additional endpoint to measure for all pesticides.

Two methods have been used for measuring eggshell strength. The puncture test uses a punch to penetrate the shell and represents a measure of shear fracture force. The compression test uses two parallel flat surfaces and represents a measure of the tensile fracture force. Both methods provide results that are linearly related to eggshell thickness in uncontaminated eggs, although the puncture test was more highly correlated with thickness (Hunt et al. 1977). The puncture method has the advantage that repeated measures can be made on each egg for calculation of an average puncture force value, whereas the compression test can only be performed once per egg. However, the compression test is a direct measure of eggshell strength that simulates field conditions (Hunt et al. 1977). Eggshell cracking is related to tensile fracture properties, which should provide an index of eggshell resistance to field insults. Another source of variation to consider is the occurrence of body-checked eggs or eggs cracked in the shell gland that are partially repaired by additional calcification (Roland 1982). Body-checked eggs were thinner than control eggs and required 18% less force to crack eggs in a compression test (Roland 1982). Candling can be used to identify body-checked eggs.

The primary advantage to measuring eggshell strength is that it may detect effects on eggshell quality at lower dietary concentrations than measuring thickness alone or effects where thickness is unaffected. The eggshell strength test integrates the effects on several parameters such as thickness and ultrastructural integrity into one measure of overall eggshell quality. Consequently, eggshell strength may be more indicative of potential field effects than thickness.

One disadvantage to this endpoint is that compression strength can not be measured on cracked or broken eggs, whereas eggshell thickness can. If the rate of egg cracking is high in a test, the number of data gaps due to unmeasured cracked eggs may be unacceptably high. This is a further reason to take all precautions to minimize the rate of cracked eggs in a test so that reproductive parameters can be measured directly. Another disadvantage is that the compression test requires mechanized testing equipment; models of table top testing instruments exist for under $10,000. Less expensive instruments that do not mechanically control compression are available, but their accuracy and precision need to be rigorously documented before they are acceptable for measuring treatment effects on eggshell quality.

Plasma Calcium Concentrations in Females

Concentrations of calcium in plasma of laying birds can be a helpful indicator of how pesticides influence eggshell quality and egg production. Eggshells consist of 98 to 99% calcium carbonate and 1 to 2% of proteins and polysaccharide material (Wilbur and Simkiss 1968). In chickens, the formation of the eggshell takes 16 to 20 hours (Talbot and Tyler 1974). Eggshell formation takes approximately 20 hours for the ring dove with 60% of the calcium derived from dietary sources and the remainder from the marrow of bones (Peakall 1970). During eggshell formation, only a small fraction (1-2%) of the calcium needed for the shell is in the blood at any one time. Reductions in the concentrations of calcium in the blood may lead to abnormally thin eggshells or birds may cease egg laying rather than producing thin eggs.

Plasma calcium concentrations in female mallards are approximately twice as high during egg laying as other times of the year (Fairbrother et al. 1990). Serum calcium concentrations of laying bobwhite were 2.3 times higher than in males (Bennett et al. 1990b). Dietary methyl parathion produced a dose related decrease in serum calcium concentrations of laying females at levels that also affected egg production, where calcium concentrations in males were unaffected by treatment (Bennett et al. 1990b). This may be related to the pesticide-induced reduction in food consumption, which would reduce overall intake of calcium. Reduced food consumption alone has been observed to reduce eggshell thickness and strength (Haegele and Tucker 1974, Bennett and Bennett 1990).

The eggshell thinning observed from exposure to DDE is not related to reductions in blood concentration of calcium. Because the calcium concentration in shell gland cell was higher than in control birds, Lundholm (1984) concluded that the translocation of calcium between blood and the shell gland was not impaired. Eggshell thinning from DDE is a function of reduced secretion from the shell gland to the egg. Consequently, the value of measuring plasma calcium concentrations is to use it as a diagnostic tool for determining the reasons for observations of reduced eggshell quality or egg production in an avian reproduction test. It may be especially useful for addressing specific questions of the effects of a particular pesticide on eggshell quality prior to attempting to document the presence of these effects under field conditions.

Parental Organ Size and Weight

The Subdivision E guidelines state that the test report should contain information from post-mortem necropsies, but does not mention specifically the reporting of information on the size and weight of internal organs. Treatment-related changes in organ weights can be useful for determining causes of other reproductive effects. For example, the size and weight of reproductive organs (e.g., testes, hemipenis (mallard), oviduct and ovary) can provide much additional information to explain lack or cessation of egg production, infertility of eggs, or lack of mating behaviors. Changes in other internal organs may indicate the need for histological examinations to identify specific cellular changes. Identifying treatment-related changes in organs in conjunction with other reproductive effects can aid in the development of additional laboratory and/or field testing to assess pesticide risks to avian reproduction.

ALTERNATIVE DATA ANALYSIS METHODS

The data analysis methods required by Subdivision E are to provide a mathematical/statistical model for the data obtained in an avian reproduction study. This model (the one way linear model for normally distributed errors) has been shown capable of handling many kinds of data and is flexible to some deviations from the assumptions implicit in the analysis. But there may be times when another mathematical model is scientifically more defensible. This section presents some alternative methods that may provide better models in some cases. The decision to use an alternative analysis should be based on sound scientific reasoning and the reasons for choosing an alternative analysis should be clearly documented.

The one way ANOVA, recommended in the Subdivision E guidelines, tests for differences in means of treatment groups and makes no supposition of increasing effect with dose. However, in toxicological studies that is very often the case and interest may lie in testing such a hypothesis. Tukey et al. (1985) have proposed an alternative trend test for toxicological studies which can be carried out on data collected from an avian reproduction study as described by Subdivision E. This procedure identifies the highest dietary concentration at which no statistically significant trend is detected; a concept similar to a NOAEL. Tukey et al. (1985) claim that this test has high power when the response is highly correlated with the dietary concentration and reasonable power against a wide variety of response patterns.

Alternative data analyses may be required if the ANOVA assumptions of homogeneous variance and normality cannot be met even on a transformed scale or if it can be shown that the data fit a distribution other than the normal distribution and analysis is desired on the untransformed scale. Or, if a dose-response study is carried out it may be desired to fit the response as a linear function of both continuous covariates (such as dietary concentration) and classification variables (such as sex) but the data do not have a normal distribution or it is desired to use a statistical model that can account for patterns of variability that are different from the normal model. This latter alternative can apply when a dose-response type of study is carried out.

Some alternative methods that can be used in the above situations include nonparametric methods, generalized linear models for responses with nonnormal distributions and linear models for responses where only the mean/variance relationship (and not the entire response distribution) is known.

If some ANOVA assumptions do not seem to be justified by the data, then the choice of an alternative method of analysis may depend on which assumption is violated. If the variability in the treatment groups seems to be approximately similar but the distribution of the response is very skewed even when transformed then nonparametric methods (Daniel 1978) can be used (see Appendix B). But if heterogeneous variance between treatment groups is the problem, then some nonparametric methods that require homogeneous variance will not be appropriate either. Monte Carlo randomization tests are nonparametric tests which deserve mention. Such tests can provide exact statistical test but they may require the computer generation of all possible outcomes or the generation of many possible samples. Because of this, they are not useful as a routine analysis but can provide some insight when other methods cannot.

When the data consist of counts of individuals affected out of the total in the brooder, and where transformations have not helped to stabilize variance between groups, it may be possible to model the data as coming from a binomial distribution and use logistic or probit regression to decide if there are differences between treatment groups. If the response is not a proportion but an integer count, then the data may be able to be modeled as coming from a Poisson distribution and a log linear model can be used (McCullagh and Nelder 1989). The methodology is similar to that used for probit regression.

Toxicology researchers in human health applications have used the above types of analyses for studies using litters of mice which are analogous to pens or clutches of birds. There is extensive literature on the application of these methods to toxicological studies, especially where correlations between individuals in the same litter (analogous to a pen or clutch) exist. Williams (1975) described a maximum likelihood method for analyzing binary response data from experiments involving such litter effects. Haseman and Kupper (1979) reviewed various models that are used to account for litter effects in toxicological studies with dichotomous responses such as death or the presence of malformation. Haseman and Soares (1976) discuss the impact of litter effects on modeling and estimation and compare various analyses methods. Kupper et al. (1986) use the method given by Williams (1975) to model intralitter correlations and to study the biases and variances of the maximum likelihood estimators.

Intralitter correlations can lead to variability that is greater than that predicted by a simple parametric model such as the binomial distribution. This extra variability has been attributed to differences between groups of subjects that are not related to treatment effects. This same phenomenon occurs in avian reproduction studies and examples have been pointed out in Section 2. For example, the tendency for a hen to have higher numbers of cracked or infertile eggs due to her own characteristics can lead to extra variability. Finney (1971) labeled this extra variation heterogeneity and provided a method for estimating it in the context of probit regression. Williams (1975) suggested the use of a beta-binomial model to account for the extra variation.

Wedderburn (1974) provided a relatively simple and general method for estimating and accounting for extra variation in the evaluation of treatment effects in the context of a linear model such as regression or ANOVA. If the distribution of the response is not known, but the relationship between the mean response and the variance of the response can be assumed to be similar to that of a known distribution such as binomial or Poisson, except for a constant of proportionality, then Wedderburn has shown that statistically consistent and unbiased estimates of treatment effects can be obtained with quasilikelihood methods. Although these methods are similar to the maximum likelihood methods of logistic and probit regression, they do not assume the complete form of the distribution, only the relationship between the mean and variance. In fact, these methods apply not only to data in the form of counts or proportions, but to any data for which a linear model is postulated for some function of the mean. McCullagh and Nelder (1989) provide a good review and explanation of this topic. Such models can be fit using existing software such as SAS (1985). The software package GLIM (Baker and Nelder 1978) was specifically developed for fitting generalized linear models either by maximum likelihood or quasilikelihood and the package is widely available as well. In a more recent application, Chen and Kodell (1989) use the beta-binomial distribution to account for litter effects and a Weibull dose-response model for teratogenic effects to obtain low-dose risk estimates for reproductive and developmental toxic effects rather than a NOEL.

Although these alternative analyses are nonstandard and require extra effort and understanding by the data analyst and by the scientist, their use should not be discouraged for these reasons. When required, they may be the most appropriate methods to use. That is, they do not depend on assumptions unjustified by the data. They may also be used by registrants when more information is sought than is required by the registration process. It is inefficient to require the registrant to carry out a less informative experiment or analysis when the information required by OPP already exists in the alternative analysis. The expertise of a trained statistician can be invaluable in interpreting the results of such analyses.

SECTION 4

CONCLUSIONS

The avian reproduction test is unlike other standardized laboratory toxicity tests for birds where the primary endpoint of the test is death. Effects on avian reproduction can be expressed in a variety of ways. The current avian reproduction test measures a suite of parameters, but there are many more that are not or can not be measured by this test. Alternative methods can be employed to address parameters not currently addressed, but the obvious conclusion from the above discussion is that screening for reproduction effects is not a one-size-fits-all proposition. No one test or experimental design can adequately address all aspects of avian reproduction or be adequate for all chemicals. No one test can have equal sensitivity to detect treatment effects across all parameters.

The current avian reproduction test provides a strong basis for screening pesticides for potential effects if conducted in a manner that achieves sufficient statistical power to detect effects that exist. There are many steps that must be taken by investigators to control for sources of variability that are unrelated to the pesticide treatment. This report has discussed several means of reducing variability unrelated to the pesticide treatment. If the evaluation of test data leads to a conclusion that the use of a pesticide at proposed application rates would produce unacceptable adverse effects on avian reproduction, several options exist for rebutting the presumption of risk using laboratory or field methods. Many of the alternative approaches discussed above may be helpful in clarifying concerns raised in the avian reproduction test. Taking these concerns to the field will provide a more realistic scenario for verifying or rebutting presumed risks, although there is a paucity of standardized field methods for evaluating avian reproduction. Specialized laboratory tests may be very helpful in identifying specific questions to address in the field environment.

If in the evaluation of test data the Agency does not find that pesticide exposure at the dietary concentrations tested produced adverse effects on avian reproduction, many questions may remain unanswered. Was the experimental design and data analysis adequate to detect existing reproductive effects? Were the measurement endpoints appropriate for identifying the kinds of reproductive effects potentially produced by the pesticide in the field? Were the test species adequate for assessing potential reproductive effects to other taxonomic groups of birds? There is often little information available to answer these questions. Investigators can and should provide an analysis demonstrating that a test had sufficient power to detect pesticide-related effects if they existed at the dietary concentrations used. However, if a test that does not detect treatment effects, there is no information about the potential for effects at higher concentrations or the kinds of effects potentially caused by the pesticide. Consequently, we have recommended that tests should be conducted using dietary concentrations that produce effects in the more sensitive endpoints. This should improve the risk assessments by providing a measured effect level to compare with the EEC, rather than making a decision amongst the uncertainty surrounding a test that failed to detect effects.

There are also situations where an avian reproduction test may not be triggered for a particular pesticide because its use pattern will not result in repeated or continuous exposures to birds. Pesticide effects on avian reproduction are more than a chronic phenomenon. They can occur rapidly with some chemicals, with potentially significant consequences. Not only are new approaches needed to properly evaluate the effects of short-term exposures to pesticides, but new criteria need to be established to identify when pesticides may present a hazard to avian reproduction under any exposure scenario.

SECTION 5

RECOMMENDATIONS

This report has reviewed the current Subdivision E guidelines (USEPA 1982a) for conducting avian reproduction tests and discussed the advantages and disadvantages of several options and alternative methods. Many of the options are designed to strengthen the test described in the Subdivision E guidelines by reducing variability in the data from sources unrelated to the pesticide treatment. Many of the alternative methods are designed to add dimensions to the test that do not currently exist, but could be ecologically relevant for evaluating reproductive effects. The intent of the report has been to provide technical input for future revisions of the Subdivision E guidelines and for discussions on the future of the avian reproduction test.

The avian reproduction test is experimentally much more complex than other standardized avian tests. Consequently, there is a need for the Agency to clearly define the objectives of the test and to explain how test results are to be evaluated and used in the risk assessment process. There is also a need for investigators to be very clear and explicit in their descriptions of test methods, analytical methods, and data interpretation, so that the test reports can be properly reviewed by the Agency.

Several concerns were listed in the Introduction about the methods used for evaluating the potential effects of pesticides on avian reproduction that OPP should examine when reviewing guidelines for conducting avian reproduction tests. Many of these concerns can be addressed through modifications to existing guidelines and through alternative test methods, although some of these alternatives have not been standardized. Attention should be given to identifying and reducing sources of variation that are unrelated to the pesticide treatment. Examples of these sources of variation, which are generally related to test procedures and the design of physical test systems, are given throughout the text. This could greatly improve the utility of tests conducted under the existing test guidelines.

The test guidelines could be more specific in defining what constitutes an acceptable test by establishing acceptability guidelines for parameters such as rate of egg cracking, eggshell thickness, and percent fertility, hatching and survival. A compilation of historical control data would be invaluable for determining what have been achievable and acceptable values for these parameters and indicating what may be currently achievable. This is not to say that these acceptability guidelines for specific endpoints should be used to judge the acceptability of the entire test.

To maximize the amount of information gained from an avian reproduction test, the test should be conducted in the range of dietary concentrations that produce significant effects on avian reproduction, unless the concentrations required would be excessively high (e.g., 100 X EEC). This would better define the kinds of effects potentially produced by a chemical, better define the relationship between the EEC for a particular use and the range of dietary concentrations causing effects, and reduce the need for repeating tests when new uses are proposed.

It is also very important that the power of the avian reproduction tests be carefully considered. It is the responsibility of the investigator to use information on the test chemical and estimates of response variability observed in previous tests with a particular test system and species to design tests to achieve sufficient statistical power when possible. This is one of the most important issues to confront in future revisions of the guidelines. There is a need for a detailed analysis of existing reproduction test data to provide guidance on the adequate numbers of birds per pen, numbers of replicate pens per concentration, and numbers of concentration given various degrees of response variability. This analysis would be very important in determining the most efficient experimental design necessary to achieve the regulatory objectives defined by the Agency.

Pesticide effects on avian reproduction are not simply a function of chronic exposure. This report has discussed several examples of reproductive effects resulting from relatively brief exposures to pesticides. New criteria need to be established for pesticides that do not satisfy existing criteria for initiating an avian reproduction test, but may affect reproduction from short-term pesticide exposures. Test methods for measuring reproductive effects of pesticides with short-term exposures also need to be standardized.

In summary, future revisions of Subdivision E guidelines can focus on maximizing the information provided by an avian reproduction test, identifying and reducing sources of variability unrelated to the pesticide treatment, reevaluating questions relating to power and sample size, establishing baseline guides for endpoints based on past data and establishing criteria for effects resulting from short-term exposure.

REFERENCES

American Society for Testing and Materials. 1990. Standard practice for conducting reproductive studies with avian species. American Society for Testing and Materials, Vol. 11.04. E 1062-86. Philadelphia, PA, pp. 660-670.

Anderson,D. W. and J. J. Hickey. 1972. Eggshell changes in certain North American birds. In K. H. Voous, (ed.), Proc. XV Int. Ornithol. Cong. E. J. Brill. Leiden, The Netherlands. pp. 514-540.

Atkinson, A. C. 1985. Plots Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis, Oxford Science Publications, Oxford University Press, NY.

Baker, R. J. and J. A. Nelder. 1978. The GLIM System, Release 3, Numerical Algorithms Group, Oxford.

Bennett, J. K. and R. S. Bennett. 1990. Effects of dietary methyl parathion on northern bobwhite egg production and eggshell quality. Environ. Toxicol. Chem. 9:1481-1485.

Bennett, J. K., S. E. Dominquez and W. L. Griffis. 1990a. Effects of dicofol on mallard eggshell quality. Arch. Environ. Contam. Toxicol. 19:907-912.

Bennett, J. K., R. K. Ringer, R. S. Bennett, B. A. Williams and P. E. Humphrey. 1988. Comparison of breaking strength and shell thickness as evaluators of eggshell quality. Environ. Toxicol. Chem. 7:351-357.

Bennett, R. S., R. Bentley, T. Shiroyama and J. K. Bennett. 1990b. Effects of the duration and timing of dietary methyl parathion exposure on bobwhite reproduction. Environ. Toxicol. Chem. 9:1473-1480.

Bennett, R. S., B. A. Williams, D. W. Schmedding and J. K. Bennett. 1991. Effects of dietary exposure to methyl parathion on egg laying and incubation in mallards. Environ. Toxicol. Chem. 10:501-507.

Box, G. E. P. and D. R. Cox. 1964. An Analysis of Transformations. J. Royal Stat. Soc., Series B, 26:211.

Box, G. E. P., W. G. Hunter and J. S. Hunter. 1978. Statistics for Experimenters, An Introduction to Design, Data Analysis and Model Building, John Wiley and Sons, NY.

Brewer, L. W., C. J. Driver, R. J. Kendall, C. Zenier and T. E. Lacher, Jr. 1988. Effects of methyl parathion in ducks and duck broods. Environ. Toxicol. Chem. 7:375-379.

Brown, M. B. and A. B. Forsythe. 1974. Robust tests for the equality of variances. J. Amer. Stat. Assoc. 69:364-367.

Busby, D. G., L. M. White and P. A. Pearce. 1990. Effects of aerial spraying of fenitrothion on breeding white-throated sparrows. J. Appl. Ecol. 27:743-755.

Canadian Council on Animal Care. 1984. Guide to the Care and Use of Experimental Animals. Canadian Council on Animal Care. Vol. 2, pp. 39-46.

Carlisle, J. C., D. W. Lamb and P. A. Toll. 1986. Breaking strength: An alternative indicator of toxic effects on avian eggshell quality. Environ. Toxicol. Chem. 5:887-889.

Case, R. M. 1972. Energetic requirements for egg-laying bobwhites. Proceedings of the First National Bobwhite Quail Symposium. Stillwater, Oklahoma, April 23-26. pp. 205-212.

Chen, J. J. and R. L. Kodell. 1989. Quantitative risk assessment for teratological effects. J. Amer. Stat. Assoc. 84:966-971.

Conover, W. J., M. E. Johnson and M. M. Johnson. 1981. A comparative sutdy of tests for homogeneity of variances, with application to the outer continental shelf bidding data. Technometrics 23:351-361.

Cox, D.R. 1958. Planning Experiments, John Wiley and Sons, NY.

Custer, T. W. and G. H. Heinz. 1980. Reproductive success and nest attentiveness of mallard ducks fed Aroclor 1254. Environ. Pollution 21:313-318.

Dahlgren, R. B. and R. L. Linder. 1971. Effects of polychlorinated biphenyls on pheasant reproduction, behavior and survival. J. Wildl. Manage. 35:315.

Daniel, W. W. 1978. Applied Nonparametric Statistics. Houghton Mifflin Company.

Ellman, G. L., K. D. Courtney, V. Andres, Jr. and R. M. Featherstone. 1961. A new and rapid colorimetric determination of acetylcholinesterase activity. Biochem. Pharmacol. 7:88-95.

Fairbrother, A., M. A. Craig, K. Walker and D. O'Loughlin. 1990.

Changes in mallard (Anas platyrhynchos) serum chemistry due to age, sex, and reproductive condition. J. Wildl. Dis. 26:67-77.

Finney, D. J. 1971. Probit Analysis, Third Edition, Cambridge University Press, London.

Finley, M. T. and R. C. Stendell. 1978. Survival and reproductive success of black ducks fed methyl mercury. Environ. Pollution 16:51-64.

Flegal, C. J. and C. C. Sheppard. 1976. Managing Gamebirds. Extension Bulletin E-692, Cooperative Extension Service, Michigan State University, East Lansing, Michigan. 15 pp.

Fox, G. A. and T. Donald. 1980. Organochlorine pollutants, nest-defense behavior and reporductive success in merlins. Condor 82:81-84.

Fox, G. A., A. P. Gilman, D. B. Peakall and F. W. Anderka. 1978. Behavioral abnormalities of nesting Lake Ontario herring gulls. J. Wildl. Manage. 42:477-483.

Franson, J. C., J. W. Spann, G. H. Heinz, C. Bunck and T. Lamont. 1983. Effects of dietary ABATE on reproductive success, duckling survival, behavior, and clinical pathology in game-farm mallards. Arch. Environ. Contam. Toxicol. 12:529-534.

Friend, M. (ed.) 1987. Field guide to wildlife diseases. U. S. Fish and Wildl. Serv. Resour. Publ. 167. 225 pp.

Fyfe, R. W., R. W. Risebrough and W. Walker II. 1976. Pollutant effects on the reproduction of prairie falcons and merlins of the Canadian prairies. Can. Field-Nat. 90:346-355.

Gile, J. D. and S. M. Meyers. 1986. Effect of adult mallard age on avian reproductive tests. Arch. Environ. Contam. Toxicol. 15: 751-756.

Glaser, R. E. 1982 Encyclopedia of Statistical Sciences, Volume 4. S. Kotz and N. L. Johnson, Editors-in Chief. John Wiley and Sons, NY.

Greenwood, R. J. 1975. Reproduction and development of four mallard lines. Prairie Natural. 7: 9-16.

Grue, C. E., W. J. Fleming, D. G. Busby and E. F. Hill. 1983. Assessing hazards of organophosphate pesticides to wildlife. Trans. N. Amer. Wildl. Nat. Res. Conf. 48:200-220.

Grue, C. E., G. V. N. Powell and M. J. McChesney. 1982. Care of nestlings by wild female starlings exposed to an organophosphate pesticide. J. Appl. Ecol. 19:327-335.

Haegele, M. A. and R. H. Hudson. 1973. DDE effects on reproduction of ring doves. Environ. Pollution 4:53-57.

Haegele, M. A. and R. K. Tucker. 1974. Effects of 15 common environmental pollutants on eggshell thickness in mallards and coturnix. Bull. Environ. Contam. Toxicol. 11:98-102.

Haseman, J. K. and L. L. Kupper. 1979. Analysis of dichotomous response data for certain toxicological experiments. Biometrics 35:281-293.

Haseman, J. K. and E. R. Soares. 1976. The distribution of fetal death in control mice and its implication on statisical tests for dominant lethal effects. Mutation Res. 41:277-288.

Haseltine, S. D., M. T. Finley and E. Cromartie. 1980. Reproduction and residue accumulation in black ducks fed toxaphene. Arch. Environ. Contam. Toxicol. 9:461-471.

Heath, R. G., J. W. Spann and J.R. Kreitzer. 1969. Marked DDE impairment of mallard reproduction in controlled studies. Nature 224:47-48.

Heinz, G. H. 1974. Effects of low dietary levels of methyl mercury on mallard reproduction. Bull. Env. Contam. Toxicol. 11: 386-392.

Heinz, G. H. 1976a. Methylmercury: second year feeding effects on mallard reproduction and duckling behavior. J. Wildl. Manage. 40:82-90.

Heinz, G. H. 1976b. Behavior of mallard ducklings from parents fed 3 ppm DDE. Bull. Env. Contam. Toxicol. 16:640-645.

Henny, C. J. and J. K. Bennett. 1990. Comparison of breaking strength and shell thickness as evaluators of white-faced ibis eggshell quality. Environ. Toxicol. Chem. 9:797-805.

Hickey, J. J. and D. W. Anderson. 1968. Chlorinated hydrocarbons and eggshell changes in raptorial and fish-eating birds. Science 162:271-273.

Holmes, W. N., K. P. Cavanaugh and J. Cronshaw. 1978. The effects of ingested petroleum on oviposition and some aspects of reproduction in experimental colonies of mallard ducks (Anas platyrhynchos). J. Reprod. Fertil. 54:335-348.

Hughes, B. O. and A. J. Black. 1976. The influence of handling on egg production, eggshell quality and avoidance behavior of hens. Br. Poult. Sci. 17:135-144.

Hunt, J. R., P. W. Voisey and B. K. Thompson. 1977. Physical properties of eggshells: A comparison of the puncture and compression tests for estimating shell strength. Can. J. Anim. Sci. 57:329-338.

Jefferies, D. J. 1971. Some sublethal effects of p,p'-DDT and its metabolite p,p'-DDE on breeding passerine birds. Meded. Fakult. Landbouwwetenschappen, Gent. 36:34-42.

King, J. R. 1973. Energetics of reproduction in birds. In D. S. Farner (ed), Breeding Biology of Birds. National Academy of Science, Washington, D.C. pp. 78-107.

King, K. A., D. H. White and C. A. Mitchell. 1984. Nest defense behavior and reproductive success of laughing gulls sublethally dosed with parathion. Bull. Environ. Contam. Toxicol. 33:499-504.

Kirkpatrick, C. M. 1955. Factors in photoperiodism of bobwhite quail. Physiol. Zool. 28:255-264.

Kirkpatrick, C. M. 1959. Interrupted dark period: tests for refractoriness in bobwhite quail hens. In R. B. Withrow (ed), Photoperiodism. Amer. Assoc. Advan. Sci. Publ. No. 55, Washington, D.C. pp. 751-758.

Kupper, L. L., C. Portier, M. D. Hogan, and E. Yamamoto. 1986. The impact of litter effects in dose-response modeling in teratology. Biometrics, 42:85:98.

Levene, H. 1960. Robust tests for equality of Variances. In I. Okin, S. G. Shurye, W. Hoeffding, W. G. Madow, and H. B. Mann (eds.), Contributions to Probability and Statistics, Stanford University Press, Stanford.

Longcore, J. R. and J. R. Samson. 1973. Eggshell breakage by incubating black ducks fed DDE. J. Wildl. Manage. 37:390-394.

Lundholm, C. E. 1984. Effect of DDE on the Ca metabolism of the duck eggshell gland and its subcellular fractions; relations to the functional stage. Comp. Biochem. Physiol. 78C:5-12.

McArthur, M. L., G. A. Fox, D. B. Peakall and B. J. R. Philogene. 1983. Ecological significance of behavioral and hormonal abnormalities in breeding ring doves fed an organochlorine mixture. Arch. Environ. Contam. Toxicol. 12:343-353.

McCullagh, P. and J. A. Nelder. 1989. Generalized Linear Models, Second Edition. Chapman and Hall, NY.

McLane, M. A. R. and D. L. Hughes. 1980. Reproductive success of screech owls fed Aroclor 1248. Arch. Environ. Contam. Toxicol. 9:661-665.

Mendenhall, V. M., E. E. Klass and M. A. R. McLane. 1983. Breeding success of barn owls (Tyto alba) fed low levels of DDE and dieldrin. Arch. Environ. Contam. Toxicol. 12:235-240.

Meyers, S. M. and J. D. Gile. 1986. Mallard reproductive testing in a pond environment: A preliminary study. Arch. Environ. Contam. Toxicol. 15:757-761.

Miller, R. G. Jr. 1968. Jackknifing variances. Ann. Mathemat. Stat. 39:567-582.

Miller, R. G. 1986. Beyond Anova, Basics of Applied Statistics. John Wiley and Sons, NY.

National Institutes of Health. 1985. Guide for the Care and Use of Laboratory Animals. National Institutes of Health, NIH Publ. No. 85-23. 83 pp.

Peakall, D. B. 1970. Pesticides and the reproduction of birds. Sci. Amer. 222:72-78.

Peakall, D. B. 1985. Behavioral responses of birds to pesticides and other contaminants. Res. Reviews 96:45-77.

Peakall, D. B. and J. L. Lincer. 1972. Methyl mercury: Its effect on eggshell thickness. Bull. Environ. Contam. Toxicol. 8: 89-90.

Peakall, D. B. and M. L. Peakall. 1973. Effect of a polychlorinated biphenyl on the reproduction of artificially and naturally incubated dove eggs. J. Appl. Ecol. 10:863-868.

Peakall, D. B., J. L. Lincer and S. E. Bloom. 1972. Embryonic mortality and chromosomal alterations caused by Aroclor 1254 in ring doves. Environ. Health Perspect. 1:103-104.

Peterson, R.G. 1985. Design and Analysis of Experiments, Marcel Dekker, Inc., NY.

Porter, R. D. and S. N. Wiemeyer. 1969. Dieldrin and DDT: Effects on sprarrow hawk eggshells and reproduction. Science 165:199-200.

Prince, H. H., P. B. Siegel and G. W. Cornwell. 1969. Incubation environment and the development of mallard embryos. J. Wildl. Manage. 33:589-595.

Ratcliffe, D. A. 1967. Decrease in eggshell weight in certain birds of prey. Nature 215:208-210.

Ratcliffe, D. A. 1970. Changes attributable to pesticides in egg breakage frequency and eggshell thickness in some British birds. J. Appl. Ecol. 7:67-107.

Rattner, B. A., L. Sileo and C. G. Scanes. 1982. Oviposition and the plasma concentrations of LH, progesterone and corticosterone in bobwhite quail (Colinus virginianus) fed parathion. J. Reprod. Fert. 66:147-155.

Roland, D. A., Sr. 1982. Relationship of body-checked eggs to photoperiod and breaking strength. Poult. Sci. 61:2338-2343.

SAS Institute, Inc. 1985. SAS/STAT Guide for Personal Computers, 6th ed., Cary, NC.

Shafer, E. W., Jr. and R. B. Brunton. 1979. Indicator bird species for toxicity determinations: Is the technique usable in test method development? In J. R. Beck, ed., Vertebrate Pest Control and Management Materials. ASTM STP 680. American Society for Testing and Materials, Philadelphia, PA, pp. 157-168.

Shorack, G. R. 1969. Testing and estimating ratios fo scale parameters. J. Amer. Stat. Assoc. 64:999-1013.

Snedecor, G.W. and W.G. Cochran. 1980. Statistical Methods, Seventh Edition, Iowa State University Press.

Stromborg, K. L. 1981. Reproductive tests of diazinon on bowhite quail. In D. W. Lamb and E. E. Kenaga, ed., Avian and Mammalian Wildlife Toxicology: Second Conference. ASTM STP757. American Society for Testing and Materials, Philadelphia, PA, pp. 19-30.

Stromborg, K. L. 1986a. Reproduction of bobwhites fed different dietary concentrations of an organophosphate insecticide, methamidophos. Arch. Environ. Contam. Toxicol. 15:143-147.

Stromborg, K. L. 1986b. Reproductive toxicity of monocrotophos to bobwhite quail. Poult. Sci. 65:51-57.

Talbot, C. J. and C. Tyler. 1974. A study of the progressive deposition of shell in the shell gland of the domestic chicken. Br. Poult. Sci. 15:217-224.

Thompson, B. K., A. A. Grunder, R. M. G. Hamilton and K. G. Hollands. 1983. Repeatability of egg shell quality measurements within individual hens. Poult. Sci. 62:2309-2314.

Tukey, J. W., J.L. Ciminera and J. F. Heyse. 1985. Testing the statistical certainty of a response to increasing doses of a drug. Biometrics 41, 295-301.

U. S. Environmental Protection Agency. 1982a. Pesticide assessment guidelines. Subdivision E. Hazard evaluation: Wildlife and aquatic organisms. EPA-540/9-82-024 (NTIS PB83-

153908). Washington, DC.

U. S. Environmental Protection Agency. 1982b. Pesticide assessment guidelines. Subdivision F. Hazard evaluation: Human and domestic animals. EPA-540/0-82-025 (NTIS PB83-153916). Washington, DC.

U. S. Environmental Protection Agency. 1982c. Pesticide assessment guidelines. Subdivision N. Chemistry: Environmental fate. EPA-540/9-82-021 (NTIS PB83-153-973). Washington, DC.

U. S. Environmental Protection Agency. 1986. Hazard Evaluation Division Standard Evaluation Procedure: Ecological Risk Assessment. EPA-540/9-85-001. Washington, DC.

Walpole, R.E. and R.H. Myers. 1972. Probability and Statistics for Engineers and Scientists. The MacMillan Company, NY. pp. 387-392.

Wedderburn, R. W. M. 1974. Quasi-likelihood Function, generalized linear models and the Gauss-Newton method. Biometrika 61:439-447.

Wetherill, G.B. 1981. Intermediate Statistical Methods. Chapman and Hall, London.

White, D. H., C. A. Mitchell and E. F. Hill. 1983. Parathion alters incubation behavior of laughing gulls. Bull. Environ. Contam. Toxicol. 31:93-97.

Wiemeyer, S. N. and R. D. Porter. 1970. DDE thins eggshells of American kestrels. Nature 227:737-738.

Wilbur, K. M. and K. Simkiss. 1968. Calcified shells. In M. Flockin, (ed.), Comprehensive Biochemistry, Elsevier, Amsterdam, Vol. 26, pp. 229-295.

Williams, D. A. 1975. The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. Biometrics 31:949-952.

APPENDIX A

SUBDIVISION E GUIDELINES FOR CONDUCTING AN AVIAN REPRODUCTION TEST

This appendix includes the test guidelines for conducting an avian reproduction test (section 71-4) from the Subdivision E guidelines (USEPA 1982a) on Hazard evaluation: Wildlife and aquatic organisms. Other relevant information, not included in the appendix, concerning the conduct of all tests in the guidance document can be found in the introductory sections on General information (section 70-1), Definitions (section 70-2), General test standards (section 70-3), and Reporting and evaluation of data (section 70-4).

§ 71-4 Avian reproduction test.

(a) When required. (1) Data on avian reproductive effects are required by 40 CFR § 158.145 to support the registration of an end-use product which meets one or more of the following criteria:

(i) Its labeling contains directions for using the product under conditions where birds may be subject to repeated or continuous exposure to the pesticide or any of its major metabolites or degradation products, especially preceding or during the breeding season.

(ii) The pesticide or any of its major metabolites or degradation products are stable in the environment to the extent that potentially toxic amounts may persist in avian feed.

(iii) The pesticide or any of its major metabolites or degradation products is stored or accumulated in plant or animal tissues, as indicated by the partition coefficient of lipophilic pesticides (§§ 165-3, -4, and -5 of Subdivision N) metabolic release and retention studies (§ 85-1 of Subdivision F), or as indicated by structural similarity to known bioaccumulative chemicals.

(iv) Any other information, such as that derived from mammalian reproduction studies (§ 83-4 of Subdivision F), that indicates the reproduction in terrestrial vertebrates may be adversely affected by the anticipated use of the pesticide product.

(2) Applicants for registration of avicides should consult with the Agency prior to conducting this test.

(3) See 40 CFR § 158.50, "Formulators' exemption," to determine whether these data must be submitted. Section II-A of this Subdivision provides an additional discussion on this subject.

(b) Test standards. Data sufficient to satisfy the requirements in 40 CFR § 158.145 should be derived from tests which comply with the general test standards in § 70-3 and all of the following test standards:

(1) Test substance. Data shall be derived from testing conducted with the technical grade of each active ingredient in the product.

(2) Species. Testing should be performed on the bobwhite quail and mallard.

(3) Dose levels. At least two treatment level groups and a vehicle control group should be used.

(4) Number of test animals. When other test data reveal bioaccumulative potential, the number of test animals in the test group should be increased sufficiently to partly offset animal deaths or data-gathering problems associated with morbidity or with tissue residue determinations.

(5) Age. Birds approaching their first breeding season should be used.

(6) Duration of administration. Birds should be exposed to treated diets beginning not less than 10 weeks before egg laying is expected, and extending throughout the laying season.

(c) Reporting and evaluation of data. In addition to the information provided in § 70-4, the test report should contain:

(1) Test results. The following information, reported for all test groups:

(i) All observed abnormal behavior;

(ii) All observed morphological and physiological responses;

(iii) Post mortem autopsy.

(2) Test conditions. The following information, reported for each treated and untreated test group:

(i) Species;

(ii) Strain;

(iii) Age;

(iv) Body weight;

(v) Number of birds per test (include sex ratio);

(vi) Individual identification of birds;

(vii) Diet;

(viii) Storage;

(ix) Feed consumption (grams per day);

(x) Observation on palatability or repellancy;

(xi) Housing conditions of test birds, including:

(A) Space allocations for mating and nesting;

(B) Protection from weather and injuries; and

(C) Lighting program, including hours per day and wattage or foot candles at bird level;

(xii) Diagram of test layout;

(xiii) Temperature;

(xiv) Water supply;

(xv) Pretest and test history or medical and chemical administration; and

(xvi) Length of treatment period and observation period.

(3) Egg and hatching data. The following information, reported for each treated and untreated test group:

(i) Egg shell thickness;

(ii) Number and percent of cracked eggs;

(iii) Eggs laid (number eggs per bird per day and per season);

(iv) Hatching egg storage data:

(A) Temperature;

(B) Humidity;

(C) Incubation data;

(D) Eggs set; and

(E) Egg turning frequency;

(V) Fertility (viable embryos);

(vi) Live 3-week embryos;

(vii) Embryos that mature, embryos that pip shell, and embryos that liberate themselves, and a determination of hatchability;

(viii) Dead embryos;

(ix) Fourteen-day-old survivors;

(x) Crippled survivors;

(xi) Post-hatching mortability;

(xii) Weights of fourteen-day-old survivors; and

(xiii) Any signs of intoxication in post-hatching survivors.

(4) Feed analysis data. Levels of concentration of pesticide in the feed used in each test, and the rationale for choice of such levels.

(d) Acceptable protocol. Except where noted, the following example of avian reproduction protocol is acceptable for the testing of both bobwhites and mallards. This study is a modification of a study that appears on pages 23 to 50 in an unpublished draft report to EPA from the American Institute of Biological Sciences (AIBS), titled analysis of Specialized Pesticide Programs, Volume VI, Wildlife Toxicology Study. The report is dated October, 1974, and was funded under EPA Contract No. 68-01-2457.

> Test animals. Pen-reared birds, previously untreated, approaching their first breeding season, and phenotypically indistinguishable from wild birds, should be used as test animals. If shipped, all birds should be examined following shipment for possible physical injury that may have been encountered in transit. If deemed necessary, several birds may be randomly selected for pretreatment necropsy at a diagnostic laboratory to assess the state of health upon arrival. It is desirable to have a 2- to 6-week health observation period prior to selection of birds for treatment.

A history of rearing practice for the birds to be tested should be obtained if possible. This history should include lighting practices during rearing, disease record, drug and any other medication administered, and exact age.

Test groups - Bobwhite. A minimum of 3 test groups of bobwhite should be used. One group should serve as a control and 2 groups as treated birds. By random distribution, 1 male and 2 females per pen, replicated by a minimum of 12 pens, should be used per group. If individual pairs (1 male and 1 female) are to be used per pen, more pens (greater than 12) per test group should be used to proide similar sensitivity to the group testing design. To determine the number of pens needed for a particular level of sensitivity, see Walpole and Myers (1972). Control and treated birds should be kept under the same experimental conditions.

Test groups - Mallards. A minimum of 3 test groups of mallards should be used. One group should serve as a control and 2 groups as treated birdss. By random distribution, 2 males and 5 females per pen, replicated by 5 or more pens, should be used per group. If individual pairs (1 male and 1 female) are to be used per pen, considerably more pens (greater than 12 per test group should be used to provide similar sensitivity to the group testing design. To determine the number of pens needed for a particular level of sensitivity, see Walpole and Myers (1972). Control and treated birds should be kept under the same experimental conditions.

Diet preparation. Concentrations for the test substance should be based on measured or calculated residues expected in the diet from the proposed use pattern(s). The concentrations should include an actual or expected field residue exposure level and a multiple level such as five. The highest nonlethal level may be estimated from data developed from the avian dietary LC50 (§ 71-2).

The test material should be added to table grade corn oil or other appropriate vehicle and premixed with an aliquot of basal diet, utilizing a mortar and pestle or mechanical blender. It is recommended that the aliquot of basal diet used for the premix be screened to remove large particles of diet before blending in the corn oil and test material. The final diet should be a uniformly mixed composition consisting of 98 to 99 parts by weight of basal diet and 1 or 2 parts by weight of corn oil. The basal diet should be a commercial game bird breeder ration (or its equivalent) that is treated with an equivalent amount of vehicle. The premix should be stored under conditions which maintain stability. Test diets should be analyzed for pesticide concentrations at intervals during the tests. If other long-term animal tests have demonstrated a propensity for the test chemical to persist or bioaccumulate, the degree of bioaccumulation in birds should be determined by measurement of tissue residues in the birds from an extra pen group put through the reproduction test. Two or three tissues should be selected for residue analysis at the end of the exposure period, based on tissues known from other studies to hold highest residues.

Testing phase - test environment. The birds should be housed in breeding pens of adequate size conforming to good husbandry practices. The mallard pens should be screen-bottomed or kept clean of spilled food and excrement. It is desirable to offer mallards water in which to bathe.

Since light is extremely important, both during rearing and during the egg laying period, all birds should be maintained for the first 8 weeks under a regime of 7 hours of light per day for maximum egg production.

The photoperiod should then be increased to 16-17 hours of light per day and either maintained at this level or increased by 15 minutes per week for the following 12 weeks. (The 12-week period may vary depending upon the time required for the onset of egg production.) An illumination intensity of 6 footcandles at the bird level during the lighting phase of the reproductive study is adequate. Avoid the use of shorter wavelength "cool white" fluorescent lights which do not emit the daylight spectrum.

Temperature and relative humidity control throughout the reproductive test is desirable and should be recorded. Recommended levels are 21°C and 55 percent relative humidity. Ventilation is necessary.

Feeding and husbandry. All birds should receive the appropriate diet *ad libitum* for the duration of the study. Water is to be provided *ad libitum*. The test chemical should be administered for at least 10 weeks prior to the onset of egg laying.

Body weights should be recorded at test initiation prior to onset of laying, and at termination. During egg laying, body weight recording is discouraged because of the adverse effects that handling may have on egg production.

Food consumption should be recorded at least at biweekly intervals throughout the study.

Mortality should be recorded by date and morbidity (noted together with clinical signs) throughout the test phase. Gross pathology data should be obtained for birds that die during the course of the test phase and for some survivors.

<u>Egg collection, storage, and incubation</u>. All eggs should be collected daily, marked according to pen from which collected, and stored at 16°C and 65 percent relative humidity. Eggs should be set at weekly intervals for incubation in a commercial incubator. All eggs should be candled on day 0 for eggshell cracks; on approximately day 11 for bobwhites and day 14 for mallards to measure fertility and early death of embryos; and on day 18 for bobwhite and day 21 for mallards to measure embryo survival. For hatching, transfer of the eggs to a separate commercial incubator or hatcher should be made on day 21 for bobwhites and Day 23 for mallards.

Recommended temperatures and relative humidity during hatching phase are 39°C and 70 percent, respectively.

<u>Bobwhite chick observations</u>. On Day 24 of incubation, the hatched bobwhite chicks should be removed, hatchability recorded, chicks housed according to the appropriate parental grouping, and maintained on control diet for 14 days. The time period should be extended if mortality occurs appreciably late. The diet should be a commercial bobwhite starter diet or its equivalent.

<u>Duckling observation</u>. On Day 27 of incubation, the hatched mallard ducklings should be removed, hatchability recorded, ducklings housed according to the appropriate parental grouping, and maintained on control diet for 14 days. The time period should be extened if mortality occurs appreciably late. The diet should be commercial mallard starter diet or its equivalent.

Eggshell thickness. One day every two weeks newly laid eggs should be collected and measured for eggshell thickness. For consistency, the eggs used for thickness determinations should be collected during weeks 1, 3, 5, 7 and 9 of the egg-laying period. An accepted procedure is to crack open the eggs at the widest portion (girth or waist), wash out all egg contents, air-dry the shells for at least 48 hours, and then measure the thickness of the dried shell plus the membranes at 3 or 4 points around the girth using a micrometer calibrated to 0.01 mm units.

Analysis. Reproductive data consists of continuous variables (e.g., shell thickness, and body weight data) and discrete variables (e.g., number of eggs laid or 14-day-old survivors). For continuous variables, experimental groups should be compared to controls by analysis of variance. For most discrete variables, survival percentages should be computed (e.g., 14-day-old survivors of eggs laid) and arcsine transformed prior to analysis of variance. Alternately, a chi square analysis of survival (contingency tables) may be used for discrete variables. Analyses should include: body weight, food consumption, eggs laid, eggshell thickness, eggs cracked, viable eggs, fertility, live 3-week embryos, hatchability, number of normal chicks or ducklings, 14-day-old survivors (per number of eggs hatched, per hen, and per number of eggs laid). Sample units are generally the pens within each group.

Withdrawal. If the test substance is toxic (reduced reproduction evident), then a withdrawal study period should be added to the test phase. The withdrawal period need not exceed 3 weeks. Continued observations should be made on egg production, fertility, hatchability, and hatching survival.

Definitions:

1. Eggs laid. The total egg production during a breeding season (which is approximately 10 weeks).

2. Eggs cracked. Eggs determined to have cracked shells when inspected with a candling lamp; fine cracks cannot be detected without utilizing a candling lamp and if undetected will bias data by adversely affecting embryo development.

3. Eggs set. All eggs placed under insubation, i.e., total eggs laid minus cracked eggs and those selected for eggshell thickness analysis.

4. Viable embryos (fertility). Eggs in which fertilization has occurred and embryonic development has begun. This is determined by candling the eggs 6 to 14 days after incubation has begun. It is difficult to distinguish between the absence of fertilization and early embryonic death. This distinction can be made by breaking out eggs that appear infertile and examining further. This is especially important when a test compound induces early embryo mortality.

5. Live 3-week embryo. Embryo that is developing normally after 3 weeks of incubation. This is determined by candling the egg.

6. Hatchability. The percentage of embryos that mature, pip the shell, and liberate themselves from their eggs as computed from the number of fertile eggs. For quail this generally occurs on day 23 or 24 of incubation, and for mallard on day 25, 26, or 27.

7. 14-day-old-survivors. Birds that survive for 2 weeks following hatch.

8. Eggshell thickness. The thickness of the shell and the membrane of the egg at the girth after the egg has been opened and washed out, then the shell with membrane dried for at least 48 hours at room temperature.

(e) References. The following references can provide useful background information in developing acceptable protocols; some outline useful statistical procedures for handling data.

(1) Cochran, W.G. 1943. Analysis of variance for percentages based on unequal numbers. Am. Stat. Assoc. 38:287-301.

(2) Davidson, K.L., and J.L. Sell. 1974. DDT thins shells of eggs from mallard ducks maintained on ad libitum or controlled-feeding regimes. Arch. Environ. Contam. Toxicol. 2(3):222-232.

(3) Duncan, D.B. 1955. Multiple range and multiple F tests. Biometrics 11:1-42.

(4) Heath, R.G., J.W. Spann, and J.F. Kreitzer. 1969. Marked DDE impairment of mallard reproduction in controlled studies. Nature 224(5215):47-48.

(5) Heath, R.G., J.W. Spann, J.R. Kreitzer, and C. Vance. 1970. Effects of polychlorinated biphenyls on birds. Presented at the XV Internat. Ornith. Congress, The Hague, 30 Aug - 5 Sept., 1970. Pp. 475-485 in Proceedings of XV Internat. Ornith. Congress. K.H. Voous, ed. E.J. Brill, (pub.) Leiden.

(6) Heinz, G. 1974. Effects of dietary levels of methyl mercury on mallard reproduction. Bull. Env. Cont. Toxicol. 11:386-392.

(7) Longcore, J.R., F.B. Sampson, and T.W. Whittendale, Jr. 1971. DDE thins eggshells and lowers reproductive success of captive black ducks. Bull. Env. Cont. Toxicol. 6:485-490.

(8) Prince, H.H., P.B. Seigel, and G.W. Cornwell. 1969. Incubation environment and the development of mallard embryos. J. Wildlife Manage. 33:589-595.

(9) Stromborg, K.L., 1981. Reproductive test of diazinon on bobwhite quail, Avian and Mammalian Wildlife Toxicology: Second conference, ASTM STP 757, D.W. Lamb and E.E. Kanaga, Eds., American Society for Testing and Materials, pp. 19-30.

(10) Walpole, R.E. and R.H. Myers. 1972. Probability and statistics for engineers and scientists. The MacMillan Company, New York. Pp. 387-392.

APPENDIX B

SELECTED METHODS FOR EXPERIMENTAL DESIGN AND ANALYSIS

RANDOMIZATION

The proper use of randomization in a designed experiment removes the potential for systematic biases in the estimation of treatment effects. Randomization usually means assigning individuals to groups such that every individual is equally likely to fall into any group. Note however, that other randomization schemes that assign individuals to groups with unequal probabilities are also possible. The alternative to randomization is the assignment of individuals to groups based on an ordering that appears to be haphazard to the scientist. Cox (example 5.6, 1958) discussed this method and how its use can invalidate the results of designed experiments. Correct randomization procedures involve the use of an algorithm that has been scientifically shown to choose individuals with known, usually equal, probability.

One method of randomly assigning individuals to groups is to number the individuals from 1 to n (where n is the total number of individuals) and then, using a random number table, pick off a sequence of the numbers from 1 to n in the order given by the table. For example, given 5 individuals numbered 1 through 5, obtain from the random number table the random order 3, 2, 5, 1 and 4. Then individual 3 goes in the first pen, individual 2 goes in the second pen, individual 5 goes in the third pen and so on. This allocation plan can be used to assign birds in age classes to pens or pens to treatment groups (diets). Another method of obtaining random order would be to use a computer program with a good random number generator to generate orderings.

INVESTIGATING DISTRIBUTIONS OF RESPONSE VARIABLES

The results of a single avian reproduction test are assumed to be representative of results that would be obtained if all birds ever exposed to the pesticide were examined and the true distribution of the responses could be examined. However, since it is not possible to examine all birds, the distributions of the responses from only the birds in the test are used. A statistical analysis will attempt to describe the distributions of the response and the extent to which the results of the test might vary if different birds or pens had been used.

The determination of significant treatment differences is made by examining the distributions of responses obtained from different pens for different treatment groups and asking if the distributions appear to be "different" from each other in the face of the variability in response that is observed. For example, an analysis of variance will decide whether two or more distributions are centered at the same value assuming that the distributions have the same variance.

Usually, distributions are summarized with an average (or a median) to describe the most likely response and a standard deviation (or range or interquartile range) to describe the amount of variability in the response. The average and standard deviation are the theoretical best summary statistics for the normal or bell-shaped distribution and since normal distributions are quite common in biological studies, the use of these statistics has become almost routine. However, it is important to remember that some responses may have distributions that are not bell-shaped and in these cases other types of summary statistics, such as the median or interquartile range, can provide better summaries of the distribution of the response.

In any study, it is important to visually examine the distribution of the responses for each treatment group (Box, Hunter and Hunter, p 193, 1978). In addition, summary statistics can be calculated, but only by looking at the distributions of the observed responses can it be known what the means, medians or standard deviations are attempting to describe. Histograms, frequency charts, stem and leaf plots, box plots and scatter plots of the data are suggested methods for looking at distributions of data. These plots can be generated by SAS's PROC UNIVARIATE. Box plots, generated by PROC UNIVARIATE and the plot option, with means superimposed on them, are especially useful for evaluating whether or not unusual points (outliers) are present. In addition, these plots can be used to decide whether or not variability is similar between groups and whether it is likely that the data come from an approximately normal distribution or whether they come from a distribution that is quite un-normal.

The analysis of variance is relatively robust to deviations from normality and so small amounts of skewness can be tolerated (Miller, 1986). However, ANOVA is not robust to deviations from homogeneity of variance and it is important to be aware of this type of deviation in the data. Visual examination of the data allows the scientist to be aware of specific strengths and weaknesses in the data.

A hypothetical dataset, given in Table 1, will be used to illustrate how exploratory data analysis techniques can be used to validate the use of particular statistical tools. This dataset is based on a hypothesized avian reproduction study involving 3 dietary concentrations (Ctl, Lo and Hi), 1 hen per pen, 15 pens per conentration group and it supposes that the study is carried out for 10 weeks (70 days).

Figure B-1 is a frequency chart for each dietary concentration of the number of fertile eggs with the average, median and standard deviation superimposed on it. It is clear from the diagram that even if the responses from each dietary concentration came from normal distributions, they appear to have very different variances. Levene's test for homogeneity of variance (described below) was conducted using the absolute value of the residuals, rejected the hypothesis of equal variance with a p-value of 0.0063. Hence, using ANOVA on this data may not be appropriate since the assumption of equal variance does not seem to be justified. In addition, there is some evidence from a visual inspection of the plot that the distributions may not be normal. The test for normality, from PROC UNIVARIATE, gave p-values of 0.0001, 0.0419 and 0.1486 for the Ctl, Lo and Hi groups respectively. However, the issue of non-constant variance is of more importance since ANOVA is relatively robust to this departure from the assumptions.

Figure B-2 is a similar chart for the eggshell thickness data. The variability within each dietary concentration group appears, from the plot, to be about the same. Levene's test resulted in a p-value of 0.8705 indicating that there is no evidence to reject the hypothesis of equal variances. The p-values for the tests of normality were 0.5832, 0.4754 and 0.4044 for the Cnt, Lo and Hi groups respectively. Visual examination of Figure 2 shows that the responses appear to come from normal-like distributions. Thus there is no evidence to believe that an ANOVA should not be used.

This first step in any data analysis is relatively easy to do and yet it can provide more insight than many other sophisticated analyses. The generation of histograms, or similar plots, to check assumptions for future statistical analysis should be incorporated into the statistical protocol.

LEVENE'S TEST

Since Bartlett's test for homogeneity of variance is very sensitive to deviations from normality (Miller, 1986), Levene's test (Levene, 1960) is a robust alternative. It is easily programed in SAS as well. Levene's test is a statistical test of the hypothesis that the variance is constant in all of the treatment groups.

To carry out the test in an avian reproduction study where the treatment groups are different dietary concentrations of pesticide, run the appropriate ANOVA on the data and obtain the residuals from this analysis. Then square the residuals and treat these squared residuals as if they are independently, identically normally distributed (under the null hypothesis of constant variance) and apply the usual ANOVA F test to them. That is, carry out an ANOVA (using a model similar to the one used on the original data) on the squared residuals and test the hypothesis that the means of the residuals are all equal.

As discussed in Miller (1986), these squared residuals do not satisfy the assumptions imposed upon them by the test. However, as demonstrated by Levene (1960), Miller (1968), Shorack (1969) and Brown and Forsythe (1974) the test preforms well and is robust to departures from normality. A simple alternative, less sensitive to heavy tailed distributions of the original data, is to use the absolute value of the residual, rather than the squared residual.

A simulation study of the performance of more than 50 proposed homogeneity of variance tests as well as a bibliography of these tests is given by Conover et al. (1981).

A sample SAS program that carries out Levene's test using the absolute value of the residuals, for a one way model, is given below.

```
DATA ONE;INFILE 'TEST.DAT';
INPUT EGGS CHICKS TRT;
RUN;

PROC GLM; CLASS TRT;
MODEL CHICKS=TRT;
OUTPUT OUT=LEVENE R=RESID;
RUN;

DATA TWO;SET LEVENE;
ABSRES=ABS(RESID);
RUN;

PROC GLM DATA=LEVENE;CLASS TRT;
MODEL ABSRES=TRT;
RUN;
```

NON-PARAMETRIC STATISTICS

When the ANOVA assumptions of normality and constant variance cannot be justified by the data nor by a transformation of the data, then one alternative is to use non-parametric statistics. Daniel (1978) is a good reference for such techniques since assumptions inherent in these analyses and examples using real data are provided. SAS, in PROC NPAR1WAY, provides non-parametric analyses for one-way classifications of data.

It is important to know when nonparametric techniques are appropriate. First of all, because ANOVA is relatively robust to departures from normality, it is best to use ANOVA when the variance seems to be constant between treatment groups and there is only slight evidence that the distribution of the response is not symmetric (Miller, 1986). But if it is evident that the response distribution is exceedingly non-symmetric then non-parametric statistics may be used. Secondly, some non-parametric tests assume that the distribution of the responses are identical, except for a difference in mean or median. Thus, responses for which the variability is different for different treatment groups should not be analyzed by these methods. The Kruskal-Wallis one way analysis (given by PROC NPAR1WAY) is one example. However, the k-sample median test in PROC NPAR1WAY only assumes that the probability of being larger than the overall median is the same for all treatment groups (Daniel, 1978). For some cases of heterogeneous variance, this may not be a problem.

Thus, although non-parametric methods make less assumptions about the distribution of the response, they still require that the response behave in certain prescribed ways. As for ANOVA, it is best to be certain that the data satisfy whatever assumptions are inherent in the analysis.

TABLE B-1. SELECTED RESULTS FROM A HYPOTHETICAL AVIAN REPRODUCTION STUDY

Dietary	# Eggs	Eggshell
CONC.	FERTILE	THICKness
Ctl	53	0.24486
Ctl	54	0.24048
Ctl	54	0.23870
Ctl	54	0.22711
Ctl	55	0.24954
Ctl	55	0.25704
Ctl	56	0.21630
Ctl	56	0.23669
Ctl	56	0.23876
Ctl	56	0.23382
Ctl	56	0.23434
Ctl	57	0.23987
Ctl	57	0.24373
Ctl	57	0.25232
Ctl	58	0.25177
Lo	47	0.22984
Lo	48	0.25984
Lo	49	0.23661
Lo	49	0.23736
Lo	50	0.23780
Lo	50	0.25032
Lo	51	0.24532
Lo	52	0.24840
Lo	52	0.24282
Lo	53	0.24862
Lo	54	0.24724
Lo	55	0.23867
Lo	56	0.23219
Lo	57	0.21183
Lo	57	0.25195
Hi	37	0.19182
Hi	38	0.20519
Hi	40	0.19119
Hi	41	0.21250
Hi	41	0.19215
Hi	41	0.20661
Hi	41	0.19116
Hi	42	0.20605
Hi	42	0.21265
Hi	44	0.19921
Hi	45	0.20144

(continued)

Table B-1. (continued)

Dietary CONC.	# Eggs FERTILE	Eggshell THICKness
Hi	46	0.1978
Hi	46	0.19135
Hi	47	0.19420
Hi	48	0.17650

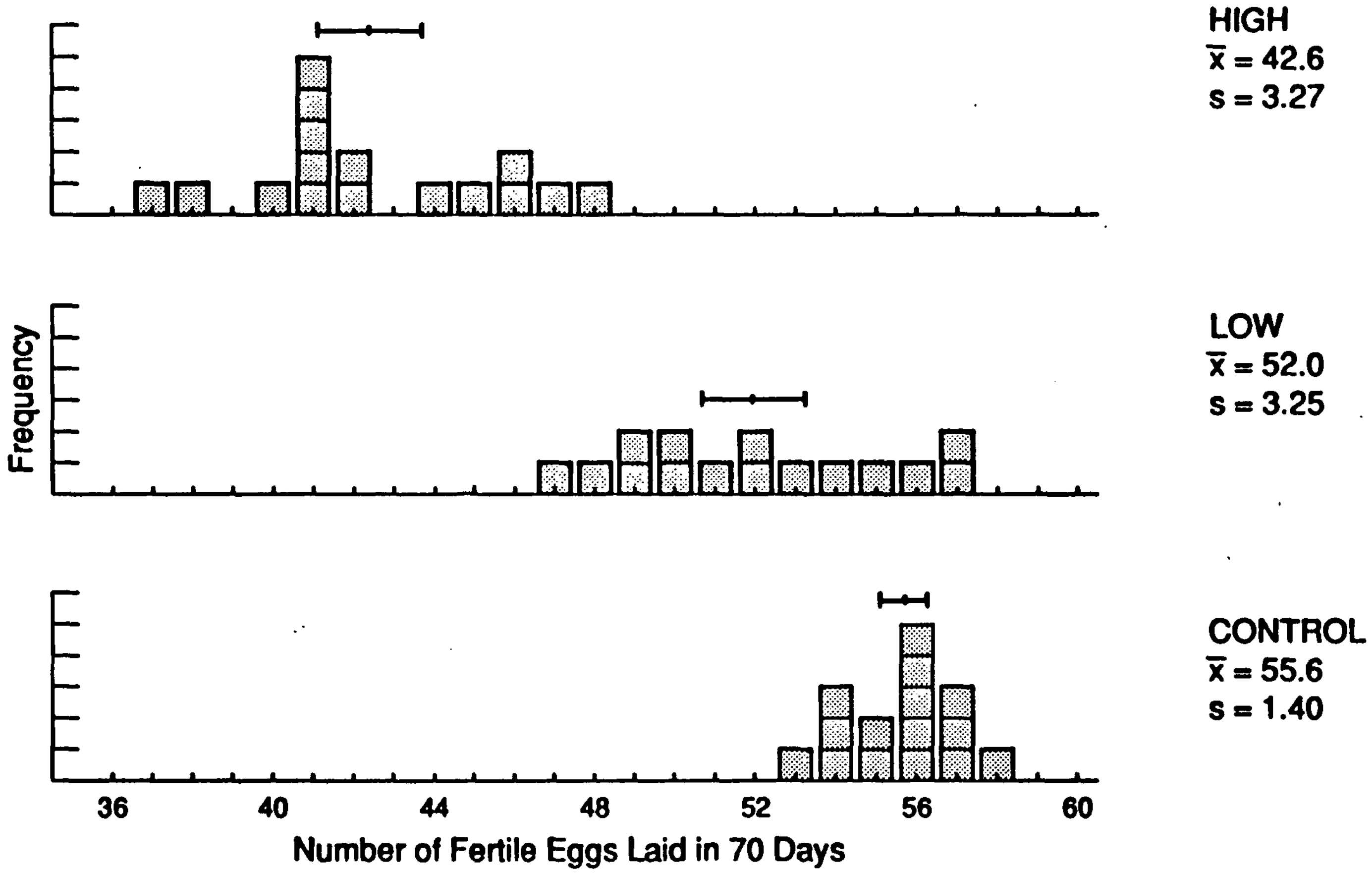

Fig. B-1 Frequency chart, summary statistics and 95% confidence intervals for the number of fertile eggs laid, by treatment group.

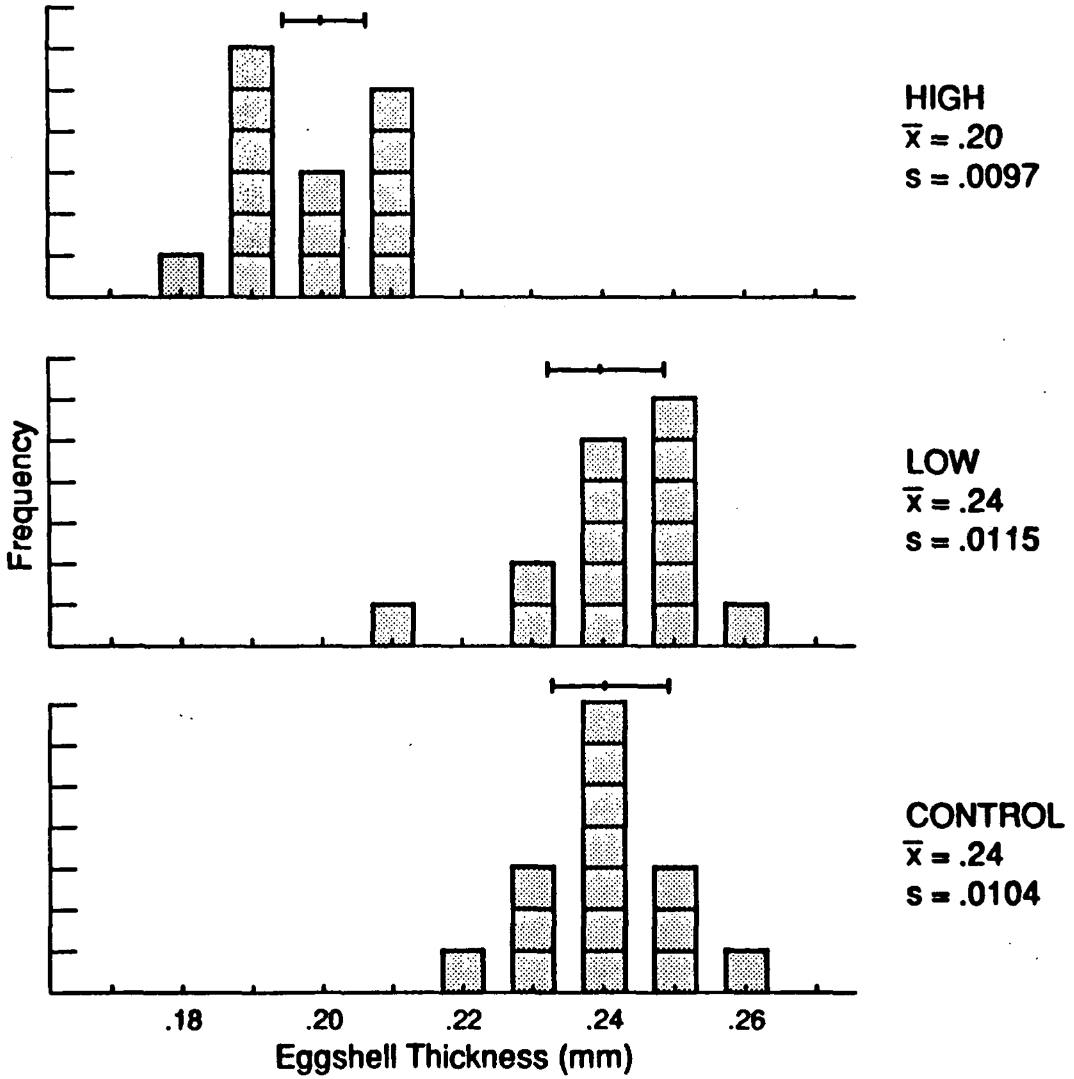

Fig. B-2 Frequency chart, summary statistics and 95% confidence intervals for eggshell thickness, by treatment group.

APPENDIX C

POTENTIAL EFFECTS OF CRACKED EGGS ON THE DETERMINATION OF PESTICIDE-RELATED EFFECTS

The number of cracked eggs in an avian reproduction study is often obtained and analyzed as an indicator or pesticide related effects or as a measure of the quality of the study. There have been many discussions concerning its usefulness and applicability as a valid endpoint. It should be recognized that eggshell cracking may be due to a variety of causes, including technical aspects such as pen construction, animal husbandry techniques such as handling of the birds, pesticide related effects or to characteristics of individual birds. It is this latter cause which has been observed to lead to problems in the analysis of other endpoints such as the number of 14 day old chicks.

It has been observed by one author that some hens are more likely, perhaps due to behavioral characteristics, to have clutches of eggs that contain cracked eggs than other hens. Thus in a study one may find hens with clutches that contain a large proportion of cracked eggs and other hens whose clutches contain no cracked eggs. Even if this pattern of cracked eggs is not related to the effects of the pesticide it can lead to incorrect conclusions concerning pesticide effects.

A hypothetical example, given in Table 2, was devised to illustrate the problem that cracked eggs may have on the detection of pesticide related effects on the number of 14 day old chicks. The example was designed to represent a hypothetical avian reproduction study involving bobwhite quail over a period of 10 weeks (70 days). In this example, it is theorized that cracking is <u>not</u> due to the pesticide treatment. The incidence of cracked eggs is the same in all of the treatment groups.

However, the presence and pattern of cracked eggs will make detecting the true effect of the pesticide more difficult. In order to provide a relatively simple example, endpoints other than the number of eggs laid, the number of viable embryos and the number of live 14 day old chicks are not included.

The hypothetical study from which these data arose had one hen per pen, 15 pens per dietary concentration group and 3 different dietary concentrations; a control (Ctl), a low concentration (Lo), and a high concentration (Hi). For each concentration group, the number of eggs laid was generated as a binomial random variable with a binomial sample size of 70. For the control group, the probability of laying an egg on any given day was set equal to 0.80. For the low group, the probability was also 0.80, but for the high concentration group, the probability was 0.60. Thus on average, the control and low concentration groups would lay about 56 eggs over the course of the study and the high concentration group would lay about 42 eggs.

It was decided to have a study-wide cracking rate of about 10%. However the pattern of cracking is crucial. This pattern reflects a typical pattern observed by one of the authors. For each dietary concentration group, 20% of the pens in each group have a pen-wide cracking rate of about 50%. The remaining pens have no cracked eggs.

The number of viable embryos in the test is calculated as the number of eggs laid minus the cracked eggs. The number of 14 day old chicks was generated as a binomial random variable with the binomial sample size equal to the number of viable embryos. For the control group, the probability that an embryo develops into a live chick is 0.65. For the low concentration group, the probability of an embryo developing into a live chick is 0.55, and for the high concentration group the probability is 0.40.

Thus, the dataset is generated to reflect a pesticide induced effect in the number of eggs laid in only the high concentration group, and an effect in the number of 14 day old chicks in both the high and low dietary concentration groups. The cracking rate is not affected by the pesticide, that is, the same cracking rate is seen in all groups.

When the number of eggs laid (LAID), viable embryos (EMBRYOS) and 14 day old chicks (CHICKS) are analyzed using a one-way ANOVA, a significant difference is detected between the high concentration group and control but not between the low concentration group and control That is, the analysis detected the true pesticide effect in the number of eggs laid, but failed to detect the difference between the control and low concentration groups for the number of 14 day old chicks.

A hypothetical scenario in which no cracking occurs is also simulated. The variable CHICK2 is generated as a binomial random variable with the binomial sample size equal to the number of eggs laid (and not the number of viable embryos). When CHICK2 is analyzed using the one-way ANOVA, a significant difference is detected between the low concentration and the control group and the high concentration and the control group.

That is, in the absence of the cracked eggs, the true effect of the pesticide is detected. However, when cracked eggs are present, the effects is masked. It should be noted that this phenomenon is due to the distribution of cracking, i.e., 50% in a few pens and 0% in other pens. If cracking occurs at the same rate in all pens this effect would not occur. In addition, the cracking in this example was hypothesized to be independent of treatment. There may be cases where cracking is due to the pesticide treatment or to a combination of husbandry/technical and pesticide effects. In these cases it is not easy to predict the effect on the analysis.

TABLE C-1. SELECTED RESULTS FROM A HYPOTHETICAL AVIAN REPRODUCTION STUDY

CONC.	# eggs	# eggs	# viable	# 14 day	# 14 day
group	LAID	CRACKED	EMBRYOS	CHICKS	CHICKS2
Ctl	54	0	54	39	34
Ctl	53	0	53	35	37
Ctl	55	27	28	19	40
Ctl	58	29	29	20	36
Ctl	49	0	49	33	29
Ctl	54	0	54	32	37
Ctl	53	0	53	31	25
Ctl	55	0	55	40	35
Ctl	58	0	58	37	40
Ctl	53	26	26	16	30
Ctl	57	0	57	35	33
Ctl	57	0	57	37	34
Ctl	57	0	57	36	42
Ctl	52	0	52	29	34
Ctl	48	0	48	32	31
Lo	51	26	25	12	23
Lo	58	29	29	18	30
Lo	54	0	54	28	30
Lo	54	0	54	34	29
Lo	54	0	54	26	34
Lo	48	0	48	33	23
Lo	55	0	55	30	32
Lo	58	0	58	37	33
Lo	53	0	53	27	30
Lo	62	0	62	34	27
Lo	56	0	56	23	35
Lo	56	0	56	28	30
Lo	60	30	30	17	30
Lo	56	0	56	30	22
Lo	60	0	60	31	32
Hi	42	0	42	17	21
Hi	41	0	41	18	15
Hi	41	0	41	13	10
Hi	47	23	24	9	18
Hi	36	0	36	14	16
Hi	42	21	21	9	14
Hi	42	0	42	23	21
Hi	44	0	44	22	18
Hi	43	0	43	17	15
Hi	40	0	40	14	13
Hi	41	0	41	15	26

(continued)

Table C-1. (continued)

CONC. group	# eggs LAID	# eggs CRACKED	# viable EMBRYOS	# 14 day CHICKS	# 14 day CHICKS2
Hi	45	22	23	4	19
Hi	40	0	40	15	20
Hi	42	0	42	18	16
Hi	47	0	47	15	17