More for Less: Analysis of the Performance of Avian Acute Oral Guideline OECD 223 from Empirical Data

Peter J Edwards,* ‡ Annegaike Leopold,‡§ Joann B Beavers, ‡|| Timothy A Springer, ‡|| Peter Chapman, ‡|| Samuel K Maynard, ‡ and Patrick Hubbard‡

‡Syngenta, Jealott’s Hill International Research Centre, Bracknell, Berkshire, United Kingdom
||Wildlife International, EAG, Easton, Maryland, USA
§Present address: Calidris Environment BV, Warnsveld, The Netherlands
||Present address: Private address

ABSTRACT
Since the publication of the Organisation for Economic Co-operation and Development (OECD) avian acute oral guideline, OECD 223, empirical data have become available to compare the performance of OECD 223 with statistical simulations used to validate this guideline and with empirical data for US Environmental Protection Agency Office of Chemical Safety and Pollution Prevention (USEPA OCSPP) guideline OCSPP 850.2100. Empirical studies comprised 244 for Northern bobwhite, of which 73 were dose–response tests and 171 were limit tests. Of the dose–response tests, 26 were conducted to OECD 223 (using 3–4 stages) and 33 to OCSPP 850.2100 (using the single 50-bird design). Data were collected from 5 avian testing laboratories from studies performed between 2006 and 2013. The success with which the LD50 and slope could be determined was 100% and 96% for OECD 223 (mean 26 birds per test) and 100% and 51% for OCSPP 850.2100 (mean 50 birds per test). This was consistent with the statistical simulations. Control mortality across all species and designs amounted to 0.26% (n = 2655) with only single mortalities occurring in any 1 study and <1% for any 1 species. The simulations used to validate the OECD 223 design showed that control mortality up to 1% will have no observable impact on the performance. The distribution of time to death for Northern bobwhite, zebra finch, and canary were obtained from 90, 29, and 17 studies, and mortalities appeared within 3 d for 71%, 95%, and 91% of birds tested, respectively.

INTRODUCTION
Animal welfare and compliance with ethical standards such as European Policy on Animal Welfare (EUPAW) and International Council on Animal Protection in OECD programs (ICAP) are essential considerations in the development and harmonization of Organisation for Economic Co-operation and Development (OECD) test guidelines. Legislation for the protection of animals used for scientific purposes (Directive 2010/63/EU) requires adherence to the principles of the 3Rs (Refinement, Reduction, and Replacement) by Russell and Burch (1959). This is a guiding principle for those organizations and staff that are involved in the conduct of toxicity tests. Following a workshop on avian toxicity testing organized by the Society of Environmental Toxicology and Chemistry (SETAC) and the OECD (OECD 1996), an OECD expert group was established to develop an acute oral toxicity guideline for birds (OECD 2010a) that would utilize fewer birds while providing adequate precision in toxicity estimates appropriate for risk assessment. This precision was achieved by computer simulations for single- and multiple-stage designs and by describing a simple framework of tests (limit, LD50 only, and LD50 slope).

The design of OECD 223 (OECD 2010a) is a multiple-stage design based on D-optimal designs (Neyer 1994). If there is prior knowledge that toxicity is likely to be low, a 5-bird single-dose limit test may be conducted. If there is no mortality in the limit test after 14 d, then the LD50 is above the limit dose and the study is complete. If there is mortality or the expectation of mortality, a single individual is treated at 4 different doses in Stage 1. The results from Stage 1 provide a working estimate of the LD50 that is used to set 10 doses in Stage 2, with 1 individual tested per dose. If there is no requirement for the estimation of slope and CLs, the data from Stages 1 and 2 are combined to
estimate the LD50 only. If a slope and CLs are required, the
LD50 estimated from Stages 1 and 2 becomes a working
estimate for setting doses in a third stage. There are 2
alternatives for Stage 3, depending on the number of reversals
in the dose–response sequence. A “reversal” is defined as a
change in response with increasing dose, that is, survival at a
higher dose than one in which mortality was recorded. For 2 or
more reversals, indicating a shallower slope, Stage 3a
comprises 5 birds at each of 2 doses placed at the working
estimate of the LD15 and LD85. If there is 0 or 1 reversal,
indicating a relatively steep slope, 2 birds are exposed at Stage
3b at each of 5 doses. If necessary, where the slope is steep,
further stages may be added to ensure the slope and CLs are
estimated. For convenience, birds are dosed 3 d apart for a
working estimate of the dose response, unless there is evidence
of delayed mortality identified from the development of clinical
signs of toxicity. Birds are observed for at least 14 d. The working
estimate of the LD50, doses for each stage and the final estimate
of the LD50, slope, and CLs are calculated by a
SEquential DESign Calculator (SEDEC) (Springer 2009).

Full descriptions of both test guidelines OECD 223 (OECD
2010a) and USEPA OCSPP 850.2100 (USEPA 2012) are
available online as part of OECD and USEPA OCSPP guideline
programs. A validation report was published in 2010 as part of
the OECD Series on Testing and Assessment (No 131) and is
referred to regularly in the present paper (OECD 2010b). This
validation report describes the results of statistical simulations
conducted to evaluate design performance of OECD 223,
including the effect of control or background and delayed
mortality on estimates of the LD50 and slope and the results of
an independent validation of these simulations. In addition,
the validation report describes SEDEC, the results of a reading
comprehension test and ring testing, together with feedback
from participating laboratories. The empirical data, presented
here, have not been reported before. The Methods section
describes how empirical data were collected and comparisons
were made with the statistical simulations.

The OECD expert group also reviewed background mortality
in laboratory-bred species and concluded that the inclusion of
controls was unnecessary when mortality is the primary
endpoint and background mortality is negligible. However,
there is a requirement in OCSPP 850.2100 to measure food
consumption and bodyweights, so in the interest of harmoniza-
tion, 5 control birds were included in OECD 223.

During development of OECD 223, the USEPA raised
concerns about the use of an LD50-only test and the influence
of background and delayed mortality on estimates of the LD50
and slope and the resulting classification of the endpoints for
risk assessment. These concerns were addressed through
statistical simulations reported in a validation report (OECD
2010b) prior to publication of OECD 223. Although USEPA
supported the publication of OECD 223 through the OECD
Working Group of the National Coordinators of the Guidelines
Program, there were some residual issues described in
Guidance for Classifying Studies Conducted Using the
OECD Test Guideline 223 (Brady 2011) that have been
addressed through recent revision of OECD 223. These issues
were essentially the use of OECD 223 with chemicals that
delayed mortality, the application of a 10% background
mortality threshold when only 5 control birds are used, the
acceptable background mortality in species other than
Northern bobwhite, and the potential bias toward higher or
lower LD50 values for OECD 223 compared to OCSPP
850.2100. In the interests of harmonization, revisions to OECD
223 included these: 1) The study is considered invalid if there
are any control mortalities when 5 control birds are used, and
2) background mortality of the chosen species should be
demonstrated to be ≤1% in the testing laboratory.

Since publication of the OECD 223, a number of studies
have been completed by avian testing laboratories, with both
Northern bobwhite Colinus virginianus and passerine species
(canary Serinus canaria domestica and zebra finch Taeniopygia
guttata). Zebra finch and canary are representative captive-
bred passerines generally used as a requirement for recent
legislation by USEPA (2011). The purpose of the present paper
is to assess the performance of OECD 223 using empirical data
generated between 2006 and 2013 and to compare these to
the statistical simulations published in OECD Validation
Report No 131 and to the performance of USEPA Guideline
OCSPP 850.2100. Performance measures were 1) the number
of animals required; 2) the percent success with which the
LD50, slope, and CLs or CLs can be calculated; 3) the influence
of delayed mortality on the estimate of the LD50 and slope;
and 4) the control mortality. Data on species other than
Northern bobwhite are presented to address the points raised
in the USEPA memorandum (Brady 2011).

METHODS

Collection of empirical data from testing laboratories

Avian testing laboratories were contacted in 2014 and
asked to complete preformatted Microsoft Excel spreadsheets
for their last 50 acute oral toxicity studies conducted
according to guidelines OECD 223 and OCSPP 850.2100,
with Northern bobwhite. If fewer than 50 studies were
available, laboratories were asked to provide all acute oral
studies performed in their laboratory since the beginning of
2006 when the first OECD 223 studies were conducted to
what was then a draft guideline. Five laboratories collabo-
rated, providing treatment and control data, for limit and
dose–response tests with Northern bobwhite for both
guidelines. Data were collected in an identical way for zebra
finch and canary, but from a single laboratory. The identity of
test substances was protected by collection of key toxicity
data (LD50 and CLs) into classes. More specifically, laborato-
ries were asked to provide the following information if
available in the report (no additional analysis was done):

1) Numbers of birds used in control and treatment groups for
each study
2) Estimation of an LD50, slope, and 95% CI
3) Estimated slope (rounded to a whole number) and 95%
   CLs as a ratio of the upper to lower bounds for the LD50
4) Counts of studies falling within given ranges of toxicity
a) For the LD50, values were collected in classes (0.01–7.3, 7.4–12, 13–20, 21–33, 34–56, 57–93, 94–155, 156–259, 260–432, 433–720, 721–1200, and 1201–2000 mg/kg bw)

b) For upper to lower ratios of 95% CIs about the LD50 were collected in classes (1.0–1.25, 1.26–1.59, 1.60–2.00, 2.01–2.51, 2.52–3.16, 3.17–3.98, 3.99–5.01, 5.02–6.31, 6.32–7.94, 7.95–10.0, 10.1–12.6, >12.6)

5) Control mortality
6) Daily distribution of mortalities in all treatment groups and studies with mortality.

Treatment group data were collected from 244 studies with Northern bobwhite, 60 with zebra finch, and 29 with canaries comprising both dose–response and limit tests (Table 1). Untreated control data were available for 209 Northern bobwhite, 56 zebra finch, and 29 canary control groups (Table 2). The numbers of studies with treatment groups and the number of control groups for Northern bobwhite and zebra finch differ because concurrent controls were used by laboratories to reduce numbers of birds. In some early studies with OECD 223, no controls were included because this was consistent with the draft OECD 223 guideline at that time.

Analysis of empirical data

Mean numbers of birds used in actual studies were compared with guideline recommendations to determine the extent to which problems encountered led to an increased use of birds. Tests conducted that were unable to measure an LD50, that is, LD50 was greater than the maximum dose, were defined as “limit tests.” Empirical data were analyzed to compare performance with the statistical simulations reported in the validation report (OECD 2010b). Not all empirical data complied with the guideline conditions. This limited compliant OECD 223 designs to 3 or more stages (LD50 dose–response test) and the single-stage OCSPP 850.2100 to 50 treated birds. Regurgitation in passerine studies further reduced the sample size. Due to limitations in sample size, a comparison was only attempted for Northern bobwhite (Table 1). The percentage of cases in which an LD50, slope, and 95% CI could be calculated in compliant empirical OECD 223 and OCSPP 850.2100 studies was compared directly with measurements in the statistical simulations. Distributions of LD50, slopes, and 95% CIs for empirical data from both guidelines were described in classes. These empirical data sets were assumed representative of modern pesticides tested and without bias. No account was taken of the pesticide class. Cumulative daily distribution of mortality was used directly to measure the proportion of deaths occurring within the first 3 d of a 14-d observation period. Mortality after 3 d may be considered “delayed” in the context of OECD 223. Background mortality was described as a percentage of birds that died in the control groups. Because there was no reason why guideline type would influence the proportion of deaths within 3 d or control mortality, data were combined for each species.

RESULTS

Performance of OECD 223 and OCSPP 850.2100 in the numbers of animals used

OECD 223 and OCSPP 850.2100 recommend 5 and 10 treated birds be used in a limit test, and 24 to 35 birds (3-4

Table 1. Numbers of studies making up the empirical data base used for comparative evaluation of performance between OECD 223 and OCSPP 850.2100

| Species    | Design (and conditions) | OECD 223a | OCSPP 850.2100b |
|------------|-------------------------|-----------|-----------------|
| Bobwhite   | Limit                   | 64        | 107             |
|            | Full                    | 31        | 42              |
|            | Full (compliant with guideline) | 26  | 33              |
| Zebra finch| Limit                   | 4         | 26              |
|            | Full                    | 3         | 27              |
|            | Full (compliant with guideline) | 3  | 23              |
|            | Full (compliant with guideline & no regurgitation) | 2  | 7               |
| Canary     | Limit                   | 3         | 9               |
|            | Full                    | 2         | 15              |
|            | Full (compliant with guideline) | 1  | 12              |
|            | Full (compliant with guideline & no regurgitation) | 0  | 6               |

OCSPP = US Environmental Protection Agency Office of Chemical Safety and Pollution Prevention; OECD = Organisation for Economic Co-operation and Development.

a Numbers of OECD 223 (OECD 2010a) studies evaluated.

b Numbers of OCSPP 850.2100 (USEPA 2012) studies evaluated.

Table 2. Numbers of control groups and individuals in the empirical database used to evaluate background mortality

| Species    | Design | Groups | Individuals |
|------------|--------|--------|-------------|
| Bobwhite   | OECD 223a | 65 | 328         |
|            | OCSPP 850.2100b | 144 | 1435        |
|            | Total  | 209    | 1763        |
| Zebra finch| OECD 223a | 8  | 85          |
|            | OCSPP 850.2100b | 48 | 492         |
|            | Total  | 56     | 577         |
| Canary     | OECD 223a | 5  | 45          |
|            | OCSPP 850.2100b | 24 | 270         |
|            | Total  | 29     | 315         |

OCSPP = US Environmental Protection Agency Office of Chemical Safety and Pollution Prevention; OECD = Organisation for Economic Co-operation and Development.

a OECD (2010a).
b USEPA (2012).
stages) and 50 birds in dose–response tests, respectively. Mean numbers of birds used in the empirical data set are presented in Table 3. Limit tests are defined as studies in which the LD50 was determined to be greater than the maximum dose tested, normally 2000 mg/kg. The definition of a limit test may have influenced the performance of OCSPP 850.2100, and it may have been that 49% of studies were failed dose–response tests. In OECD 223 limit tests with passerines, 10 rather than 5 birds were used because of uncertainty in the acceptability of using 5 birds to USEPA. Studies in which regurgitation was observed have been excluded from Table 3 because it prevented completion of the study due to uncertainty of the dose. Data lost through regurgitation in Table 3 represented 39% of 28 studies with canary and 12% of 60 studies with zebra finch, unlike Northern bobwhite for which regurgitation was absent from all studies.

Performance of OECD 223 and OCSPP 850.2100 in estimating the LD50, 95% CLs, and slope from empirical data

Performance was measured in all compliant OECD 223 (n = 26) and OCSPP 850.2100 (n = 33) designs and presented in Table 4. Success in estimating all parameters was high, with the exception of slope in OCSPP 850.2100. These empirical data are consistent with the statistical simulations reported in the validation report (OECD 2010b). The empirical distributions of the LD50 are presented in Figure 1, and LD50 estimates are truncated at the limit dose of 2000 mg/kg. Although the sample size is probably too small to detect bias in LD50 estimates, the distributions look visually similar.

The distribution of empirical slope estimates for different compounds may indicate that the OCSPP 850.2100 design slightly favors the shallow slopes (<5) with fewer steep slopes (>10) compared to empirical slopes for OECD 223 (Figure 2). Although this may reflect differences in the data set, it is considered further because of the difference in the success in estimating slopes (Table 4). The higher success rate of estimating slopes and the apparent distribution of slope estimates for OECD 223 are consistent with the results of the statistical simulations in the validation report (OECD 2010b) (see Figure 3). In these simulations, OECD 223 was compared to the best single-stage 50-bird design with 10 concentrations of 5 birds (OECD 2010b). The single-stage design performs

---

**Table 3. Average numbers of birds used in each study**

| Species         | OECD 223 limit (5) | OCSPP 850.2100 limit (10) | OECD 223 full (24–34) | OCSPP 850.2100 full (50) |
|-----------------|--------------------|---------------------------|-----------------------|-------------------------|
| Bobwhite        | 6.1                | 23.6                      | 26                    | 49.8                    |
| Zebra Finch     | 10                 | 20.2                      | 26                    | 51.5                    |
| Canary          | 11.3               | 22                        | NA (regurgitation)    | 50                      |

OCSPP = US Environmental Protection Agency Office of Chemical Safety and Pollution Prevention; OECD = Organisation for Economic Co-operation and Development.

*AAverage from all OECD 223 (OECD 2010a) studies evaluated.

*BAverage from all OCSPP 850.2100 (USEPA 2012) studies evaluated.

---

**Table 4. Percent success in measuring LD50, 95% confidence limits and slope for bobwhite using empirical data compliant with OCSPP850.2100 and OECD223 (2–3 stage dose response design)**

| Guideline      | Number of studies | LD50 (% success) | 95% confidence limits (% success) | Slope (% success) |
|----------------|-------------------|------------------|-----------------------------------|------------------|
| OECD223        | 26                | 100              | 85                                | 96               |
| OCSPP850.2100  | 33                | 100              | 94                                | 51               |

**Figure 1. Distribution of LD50 estimates from empirical data.**

**Figure 2. Distribution of slopes from empirical data.**

**Figure 3. Success in estimating slopes.**
poorly in estimating steep slopes and may explain why empirical data for OCSPP 850.2100 estimate fewer slopes (Table 4) and that those that are estimated are biased to shallow slopes. Furthermore, the success of estimating steep slopes (10) is lower in single-stage designs when dose ratios are large (50).

The 95% CIs for the LD50, expressed as a ratio of the upper to lower bounds from classes, appears to be skewed toward lower ratios for OCSPP 850.2100 (Figure 4). For OECD 223, the 95% CI was always estimated together with a slope. In the case of OCSPP 850.2100, 95% CI were estimated using binomial distribution theory or Spearman–Karber methods (Finney 1971) on 16 occasions when the slope could not be estimated. The absence of partial mortalities (all dead or live at adjacent doses) prevents fitting a probit model and estimation of a slope. Under these conditions, the empirical estimates of the 95% CI about the LD50 might be biased, although this possibility was not further explored.

Taking only empirical studies compliant with the guidelines, the high success rate in measuring the LD50 for OECD 223 was achieved at Stage 3 (100%) (Table 4). Stage 4 was required on only 3 occasions (11%) to estimate the slope. This resulted in mean animal use of 26 birds in OECD 223 dose–response test (Table 3). These empirical results compared well to the statistical simulations in which estimates of the LD50 were virtually 100% in Stage 3 (24 birds) and slopes were 74% in Stage 3 and 26% in Stage 4 (34 birds). Of those requiring Stage 4, 44% had true slopes of 10, thus showing the utility of Stage 4 in dealing with steep slopes. The sample size for zebra finch and canary (Table 1), especially after losses through regurgitation, do not allow for an adequate comparison of performance between the guidelines.

Background mortality

Empirical control data collected for Northern bobwhite, zebra finch, and canary are presented in Table 5. Because experimental conditions are similar, there are no reasons to expect the guideline to influence the background mortality, and data from both guidelines have been combined. Of 294 studies (all species and designs), there were 7 studies with single control mortalities. This represented 7 individual deaths from 2655 birds (0.26%). Simulations in the validation report (OECD 2010b) show that increasing natural mortality from 0% to 1% had very little impact on the success rate in estimating the LD50 and slope for OECD 223. As the percentage mortality increased to 5% and 10%, success rate in estimating the LD50 changed little, but estimates shifted downward. These levels of control mortality were not exceeded in the empirical data (Table 5).

Influence of delayed mortality on the estimate of the LD50 and slope

An analysis of the temporal distribution of mortalities with Northern bobwhite from 144 studies using OCSPP 850.2100 and 65 using OECD 223 from 5 testing laboratories (Figure 5a) demonstrated that 71% of mortalities occurred within 3 d (801/1123). These temporal distributions were similar to empirical data, collected from a single laboratory before 2006, used in the simulations for the validation report (OECD 2010b) and again here in Figure 5a. The temporal distribution of mortalities for zebra finch (Figure 5b) was taken from 56 studies, 48 with OCSPP 850.2100 and 8 with OECD 223. For canary (Figure 5c), they were combined for both guidelines (29 studies) because the data set for OECD 223 alone was too small in this species. The proportion of mortalities occurring within 3 d for zebra finch and canary, taken from combined data for both guidelines, was 95% (212/222) and 91% (115/127), respectively. Mortality appeared to occur more quickly after dosing in smaller passerines than was the case for the larger Northern bobwhite. Statistical simulations in the validation report showed that OECD 223 would estimate the LD50 in more than 99% of simulated experiments when mortality was taken from a distribution based on the empirical delayed mortality provided in Figure 5a. Distributions of estimates of the LD50 appear similar to those with no delayed mortality (OECD 2010b). The simulations of the OECD 223 design with empirical delayed mortality were less successful in estimating the slope (52%) but still better than the best single-stage design without delayed mortality (43%) (OECD

![Figure 4. Distribution of 95% CI ratios.](Image)

### Table 5. Control mortality

| Guideline | Total individuals | Individual deaths | % mortality |
|-----------|------------------|-------------------|-------------|
| Bobwhite (5 laboratories) | 328 | 1 | 0.30 |
| OECD223 | 1435 | 1 | 0.07 |
| OCSPP850.2100 | 1763 | 2 | 0.11 |
| Combined | 85 | 1 | 1.18 |
| Zebra finch (1 laboratory) | 492 | 3 | 0.61 |
| OECD223 | 577 | 4 | 0.69 |
| Combined | 45 | 0 | 0 |
| Canary (1 laboratory) | 270 | 1 | 0.37 |
| OECD223 | 315 | 1 | 0.32 |
2010b). Analysis of the empirical data for OECD 223 (Table 6) shows that mortalities occurring after 3 d had little impact on the success rate of measuring the LD50, 95% CI, or slope and represented better performance than in the simulations. This contrasted with OCSPP 850.2100, where estimation of the slope was poor in both simulations and the empirical data sets.

DISCUSSION

The present paper presents the first opportunity to evaluate and compare the performance of OECD 223 and OCSPP 850.2100 designs from empirical data with each other and to compare this performance with the statistical simulations described in the validation report. Increasingly, statistical simulations are used to evaluate and improve testing procedures (Rispin et al. 2002), especially where there needs to be a balance between measuring performance and reducing the numbers of animals used in the guideline development process (ring testing). The combination of virtually unlimited statistical simulations and highly restricted in-vivo testing of vertebrates represents an excellent combination for an initial evaluation of performance. More extensive in-vivo testing can only be justified during the course of regulatory testing after guidelines have been published. Since OECD 223 was published in 2010, there are now sufficient data available to confirm guideline performance or the need for revision. The USEPA issued guidance for classifying studies conducted using OECD 223 (Brady 2011). This guidance included further evaluation of a potential bias in the estimation of the LD50 compared to OCSPP 850.2100, applicability of the results of the validation for other species, numbers of controls used, and the implications of delayed mortality.

Quality of the endpoint estimate

Empirical data, like the ring testing in the validation report, did not include further direct comparisons for both designs with the same compounds. The combination of statistical simulations in the validation report and empirical data provides a good basis for concluding that both guidelines have a high probability of providing LD50 estimates (Table 4). With respect to further evaluation of a potential bias in the estimation of the LD50 compared to OCSPP 850.2100, sample size is probably too small to detect bias in LD50 estimates from empirical distributions (Figure 1).

The success rate in estimating slope was also similar between simulations and empirical data, as indicated in Figure 3. It should also be noted that the performance of OCSPP 850.2100 is poor for slopes >5. The single-stage design (such as OCSPP 850.2100) is dependent on the initial guess of the LD50, the number of doses, and the high-
low-dose ratio. Multiple-stage designs (OECD 223) correct for bad initial guesses and optimize subsequent doses by taking account of reversals and partial mortalities. When the required LD50, 95% CLs, or slopes are determined, no further stages (testing) are required. There is no limit to the number of stages, and this guarantees the required endpoints will be determined. The simulations and empirical data demonstrate that 3 to 4 stages, representing 24 to 34 animals, are adequate.

In addition to LD50 and slope, the width of the 95% CLs for the LD50 is also an important performance measure. As stated earlier, no direct comparisons between both designs are available for the same compounds, other than for the 2 ring test compounds described in the validation report. The distribution for the ratio of upper to lower bounds of 95% CLs in Figure 4 appears skewed, with a higher proportion of low ratios for OCSPP 850.2100 compared to OECD 223. Possible reasons for this include differences in the representation of classes of chemicals tested using the different guidelines, resulting in different distributions of individual bird responses, differences in CLs arising from use of different estimation methods (related to presence or absence of slope estimates), and use of smaller numbers of test animals in OECD 223 studies.

Control mortality

Empirical data demonstrate that, for captive-bred birds, Northern bobwhite, and smaller passerine species, typical background mortality rates are low: about 0.1% for Northern bobwhite and 1% for passerines. Simulations in the validation report show these levels have very little impact on the frequency or distributions of estimates of the LD50 or slope. The only impact of higher background mortality in the treated population (10%) is a slightly lower estimate of the LD50 (OECD 2010b). These data support the original position of the OECD expert group that untreated controls are unnecessary for captive species when the primary endpoint is the LD50. Control birds are only required to measure treatment-related effects on bodyweight and food consumption. For harmonization with the requirement of OCSPP, 5 control birds were included, with provision that an additional 5 would be included if 1 bird dies. Brady (2011) stated that the OECD 223 study is invalid if >10% of the controls die or additional control birds are added during the course of a study. While this was disputed, it was agreed that, in the interests of animal welfare, 5 birds would be included and if there were any deaths the study would be repeated. The incidence of rejecting a study with 1 in 5 control deaths (20%), when the background mortality rate is 0.1% and 1%, is 1 in every 200 and 20 studies, respectively. At this rate, fewer animals will be tested by using 5 birds and accepting that the occasional study will have to be repeated.

Delayed mortality

Study duration of multiple-stage designs like OECD 223 could be very long if the interval between stages was set at 14 d. Defining the doses for the next stage after 3 d largely overcomes this inconvenience. However, the use of mortality on day 3 to make decisions about the next set of doses may not always be appropriate if there is prior knowledge suggesting that delayed mortality may be common for the test substance or if clinical signs on day 3 indicate further mortality may occur. Thus, delayed mortality in OECD 223 may prevent the optimum choice of doses for the next stage. Under these circumstances, the calculation of the working estimate may be delayed until recovery of the remaining test birds is evident, which may be 14 d in some cases. The final estimates of the LD50 and slope from an OECD 223 study are no more influenced by delayed mortality than such estimates from an OCSPP 850.2100 study because, in both, the effects are observed after at least 14 d.

Brady (2011) stated that OECD 223 should not be used with chemicals that cause delayed effects. Evidence presented shows there is no more basis for this statement in relation to OECD 223 than to OCSPP 850.2100. Statistical simulations show that OECD 223 with empirical delayed mortality still performs better than single-stage designs with no delayed mortality. Furthermore, the empirical data presented in the present paper provide no evidence that mortalities occurring later than 3 d have any influence on the success in estimating the LD50, 95% CI, and slope (Table 6).

The cumulative distributions from empirical data in Figure 5 show that on average 71%, 95%, and 91% of mortalities appear during the first 3 d for Northern bobwhite, zebra finch, and canary, respectively, and that for Northern bobwhite, distributions were similar to those used in the simulations. We conclude that for a high proportion of compounds, 3 d is ideal for estimating doses for the next stage for all 3 species, and for the rare test materials in which delayed mortality does occur, the effects on the quality of LD50 estimates are minimal.

Managing multiple-stage designs

In addition to the longer duration of multiple-stage designs like OECD 223, the researcher is confronted with some complexity in the selection of doses to be used at each stage. An Excel workbook called SEDEC (Springer 2009) was developed to make this easy. An initial estimate of the LD50 is entered into SEDEC, and it determines Stage 1 doses. The program then guides the user through selection of doses in subsequent stages. After each stage, SEDEC calculates a working estimate of the LD50 and provides optimal doses for the next stage. At the end of the study, it calculates a final LD50, dose response, slope, and CIs. SEDEC also provides facilities for printing reports of the study results and providing a secure audit trail. This is freely available on the OECD website (Springer 2009).

Managing the numbers of birds required before the study starts can be more difficult when the tests and numbers of stages required are uncertain. This uncertainty makes it more difficult for laboratories with a low turnover of studies to meet the full potential of OECD 223.

Animal welfare

Application of animal welfare principles such as those of the National Centre for Replacement, Refinement and
Reduction of Animals in Research (NC3Rs) was an essential part of the development of OECD 223. Internationally, regulatory authorities require an estimate of the acute avian LD50 for risk and hazard assessment of chemicals, and complete replacement methods are not available at this stage. With this limitation, OECD 223 applies reduction and refinement through the efficient multiple-stage designs, limiting the use of controls and matching the precision required to the needs of the risk assessment by describing a limit test, LD50 only, and LD50 slope test. The use of limit tests is harmonized into international guidelines for chemicals of low toxicity, where a limit dose supports the conclusion of low risk. Further reduction in animal use may be achieved using the 2-stage LD50-only test (14 birds) in circumstances where a slope is not required or utilized to characterize the risk. Alternatively the LD50-only test may be used to characterize the species sensitivity distribution or geometric mean with more than a single species.

Empirical data collected (Table 3) show that the numbers of birds used in dose–response designs closely matched the guideline requirements. However, this is not the case for limit tests performed in studies identified as compliant with OCSPP 850.2100 in which more than twice as many birds are used as are required by the guideline. This unnecessary animal use may be rectified by more informed estimation of initial toxicity and adherence to the guidelines description for a limit test. For OECD 223 passerine tests, 10 birds were used in limit tests (Table 3) as a precaution over uncertainty in acceptability by USEPA. With the low background mortality observed in captive-bred birds, only 5 control birds should be required in future.

In the interest of animal welfare and efficient testing policy, it is essential that guidelines be globally harmonized to prevent the need for repeating studies. In addition to an efficient design like OECD 223, this harmonization includes the choice of the same species or acceptance of different species. To achieve this, national regulatory authorities have to allow more flexibility and improve their engagement in the principles of animal welfare.

The use of a limit test or delivering an LD50 and slope estimate 100% of the time with far fewer than 50 birds is a significant improvement with OECD 223, even if controls have to be included. In a recent evaluation of the need to test pesticide formulations when data were available for technical active substances (Maynard et al. 2014), it was calculated that a 61% reduction in use of birds could be achieved using OECD 223 to estimate the LD50 and slope, compared to OCSPP 850.2100. We hope the publication by Maynard et al. (2014) will encourage regulatory authorities to end their requirements for formulations studies when they can be calculated from technical active substances. In addition we hope the present paper will encourage greater use of OECD 223 for essential studies with technical material to deliver animal welfare benefits globally with better characterization of the dose response (getting more from less).

Sequential multistage designs like OECD 223 may be considered replacements for single-stage designs in other acute test guidelines where individual animals are dosed. CONCLUSIONS

The frequencies of estimating the LD50 and slope from empirical data were 100% and 96% for OECD 223 and 100% and 51% for OCSPP 850.2100, respectively. This was consistent with statistical simulations presented in the validation report (OECD 2010b). These performance statistics apply to minimum designs compliant with the guidelines for dose–response tests: 24 to 34 birds (3–4 stage design) for OECD 223 and 50 birds (single-stage design) for OCSPP 850.2100.

Mean control mortality for Northern bobwhite, canary, and zebra finch was 0.11%, 0.32%, and 0.69%, respectively. Background mortality rates below 1% were shown to have very little influence on estimates of the LD50 and slope, and are well below the 10% that causes a nonnegligible shift of the estimates of the LD50 downward (as demonstrated by the simulations conducted during guideline validation).

Mortalities for Northern bobwhite, canary, and zebra finch were observed within 3 d in 71%, 91%, and 95% of birds tested, respectively. Consequently extending the observation period beyond 3 d in OECD 223 to define concentrations for the next stage, in response to the observation of clinical signs, will not be necessary unless a high proportion of deaths occur after 3 d in the early test stages.

Simulations performed during the validation of OECD 223 showed that delayed mortality did not affect the probability that the LD50 could be estimated. Although the probability of estimating the slope was lowered, OECD 223 with delayed mortality still performs better than OCSPP 850.2100 without delayed mortality.

Acknowledgment—The authors would like to thank all scientists involved in the OECD Avian Expert group, with particular mention to Christine Hartless, Elizabeth Margosches, Wout Slob, Dave Farrar, Marie Chantal-Huet, and Carol Yarrow, in addition to Wildlife International, Smithers Viscient, Huntingdon Life Sciences, Syntech Research, and Bayer CropScience for providing the data that enabled this evaluation.

Data Accessibility—Data referenced in this paper are publicly available and have been supplied in Supplemental Data files published with this work.

SUPPLEMENTAL DATA

Supporting data for bobwhite compiled by all 5 contributing laboratories contains information on control mortality, distribution of time to death, frequency of estimating the LD50, 95% CI, CI ratio, slope, and bins providing distribution classes for LD50 and 95% CI.

Supporting data for canary compiled by Wildlife International. Contains information on control mortality, distribution of time to death, frequency of estimating the LD50, 95% CI, CI ratio, slope, and bins providing the distribution classes for LD50 and 95% CI.

Supporting data for zebra finch compiled by Wildlife International. Contains information on control mortality, distribution of time to death, frequency of estimating the LD50, 95% CI, CI ratio, slope, and bins providing the distribution classes for LD50 and 95% CI.
REFERENCES

Brady DJ. 2011. USEPA EFED guidance for classifying studies conducted using the OECD Test Guideline 223 (TG223) (acute avian oral sequential dose study). [cited 2016 August 26]. www.epa.gov/pesticide-science-and-assessing-pesticide-risks/guidance-classifying-studies-conducted-using-oecdoed

[EU] Official Journal of the European Union. 2010. Directive 2010/63/EU of the European Parliament and of the Council of 22nd September 2010 on the protection of animals used for scientific purposes. [cited 2016 August 26]. http://data.europa.eu/eli/dir/2010/63/oj

Finney DJ. 1971. Statistical methods in biological assay. 2nd ed. London (UK): Griffin. 668 p.

Maynard SK, Edwards P, Wheeler JR. 2014. Saving two birds with one stone: Using active substance avian acute toxicity data to predict formulated plant protection product toxicity. Environ Toxicol Chem 33:1578–1583.

Neyer BT. 1994. A D-optimality-based sensitivity test. Technometrics 36: 61–70.

[OECD] Organisation for Economic Co-operation and Development. 1996. Report of the SETAC/OECD workshop on avian toxicity testing - Series on Testing and Assessment No 5. Inter-organisation programme for the sound management of chemicals. Paris (FR). GD(96)166. [cited 2016 August 26]. www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?docid=66&doclanguage=en

[OECD] Organisation for Economic Co-operation and Development. 2010a. Test No. 223: Avian acute oral toxicity test. Paris (FR). [cited 2016 August 26]. https://doi.org/10.1787/9789264090897-en

[OECD] Organisation for Economic Co-operation and Development. 2010b. Report of the test method validation of avian acute oral toxicity test (OECD test guideline 223). Series on Testing and Assessment, No. 131. Paris (FR). [cited 2016 August 26]. www.oecd.org/env/ehs/testing/seriesontestingand assessmentطرقتجربةالسكتاتانكس

Rispin A, Farrar D, Margosches E, Gupta K, Carr G, Green M, D McCall. 2002. Alternative methods for the median lethal dose (LD50) test: The up and down procedure for acute oral toxicity. J ILAR 43(4):233–243.

Russell WMS, Burch RL. 1959. The principles of humane experimental technique. London (UK): Methuen.

Springer TA. 2009. SEquential DEsign Calculator - A tool for use with OECD TG 223: Avian acute oral toxicity test. User’s guide and programme. Version 1.3 October 5. Paris (FR). [cited 2016 August 26]. www.oecd.org/env/ehs/testing/softwaretobeusedwithtg223.htm

[USEPA] US Environmental Protection Agency. 2011. Data requirements for pesticides. 40 CFR Part 158.630.124. 7-1-11 ed. Fed Regist [cited 2016 August 26]. www.gpo.gov/fdsys/pkg/CFR-2011-title40-vol24/pdf/CFR-2011-title40-vol24-part158.pdf

[USEPA] US Environmental Protection Agency. 2012. Ecological effects test guidelines – OCSSP 850.2100: Avian acute oral toxicity test. Washington (DC). EPA712-C-025.