Multi-Criteria Decision Analysis of Test Endpoints for Detecting the Effects of Endocrine Active Substances in Fish Full Life Cycle Tests


Fish full life cycle (FFLC) tests are increasingly required in the ecotoxicological assessment of endocrine active substances. However, FFLC tests have not been internationally standardized or validated, and it is currently unclear how such tests should best be designed to provide statistically sound and ecologically relevant results. This study describes how the technique of multi-criteria decision analysis (MCDA) was used to elicit the views of fish ecologists, aquatic ecotoxicologists and statisticians on optimal experimental designs for assessing the effects of endocrine active chemicals on fish. In MCDA qualitative criteria (that can be valued, but not quantified) and quantitative criteria can be used in a structured decision-making process. The aim of the present application of MCDA is to present a logical means of collating both data and expert opinions on the best way to focus FFLC tests on endocrine active substances. The analyses are presented to demonstrate how MCDA can be used in this context. Each of 3 workgroups focused on 1 of 3 species: fathead minnow (*Pimephales promelas*), Japanese medaka (*Oryzias latipes*), and zebrafish (*Danio rerio*). Test endpoints (e.g., fecundity, growth, gonadal histopathology) were scored for each species for various desirable features such as statistical power and ecological relevance, with the importance of these features determined by assigning weights to them, using a swing weighting procedure. The endpoint F1 fertilization success consistently emerged as a preferred option for all species. In addition, some endpoints scored highly in particular species, such as development of secondary sexual characteristics (fathead minnow) and sex ratio (zebrafish). Other endpoints such as hatching success ranked relatively highly and should be considered as useful endpoints to measure in tests with any of the fish species. MCDA also indicated relatively less preferred endpoints in fish life cycle tests. For example, intensive histopathology consistently ranked low, as did measurement of diagnostic biomarkers, such as vitellogenin, most likely due to the high costs of these methods or
INTRODUCTION

Fish full life cycle (FFLC) tests are being included in conceptual frameworks for the assessment of endocrine active substances (e.g., OECD 2002; USEPA 2002). However, it is unclear how such tests should best be optimized in terms of design and degree of replication to provide ecologically relevant and statistically sound results, while minimizing cost and animal use. This study presents the outcome of a workshop organized for the CEFIC (European Chemical Industry Council) Long-Range Research Initiative (http://www.cefic-lri.org/). The workshop comprised an invited group of aquatic ecotoxicologists, fish ecologists and statisticians (the authors of this study).

This study uses data and structured expert judgment to identify which fish life cycle test endpoints are the most preferred for assessing the ecological risk of chemicals in general and endocrine active chemicals in particular. The question of endpoint selection is one that arises regularly in the field of ecotoxicology and is most logically addressed by considering endpoint sensitivity, statistical power, ecological significance, and cost. A common problem in answering it is that many ecotoxicologists pursue sensitivity of response without adequately considering power and ecological significance. On the other hand, those ecologists who take an interest in ecotoxicology may not always understand the expense, difficulty, and sometimes the technical impossibility, of running a fish study to detect small changes in highly variable endpoints. In essence, appropriate endpoints in fish life cycle tests must be 1) of biological importance at the population level, defined herein as ecologically relevant and hence relevant for regulation, 2) sensitive to chemicals and the level of effect which is biologically significant should be understood, and 3) of known statistical power for detecting biologically important levels of response. Omission of any of the above when deciding on appropriate endpoints is likely to lead to suboptimal choices. Finally, all these criteria need to be considered under overarching considerations of cost and animal ethics. The best test that could possibly be conducted is not always reasonable from a resource perspective.

The use of FFLC experiments specifically to evaluate the potential environmental impact of endocrine active chemicals is quite recent. Many effects of endocrine active chemicals are sub-lethal, for example altered sexual differentiation (Andersen et al. 2003) and reduced fecundity (Nash et al. 2004). The complete FFLC test ensures that these effects are identified, allowing longer-term impacts on populations to be simulated. Most proposed test designs involve the addition of endocrine-relevant ‘mechanistic’ endpoints to standard FFLC test designs, such as sex ratio, gonadal histopathology and vitellogenin induction (OECD 2008). It is probable that concern for endocrine-mediated effects in fish (most likely derived from findings in the mammalian toxicology database or a positive result from an in vivo fish endocrine screening test) would be used to trigger a FFLC test with endocrine endpoints. This is likely to be the case under the revisions to the European Plant Protection Products Directive (91/414) and Registration, Evaluation, Authorisation and restriction of Chemicals (REACH; as substances of ‘equivalent concern’).

Similarly, in the US the Endocrine Disrupter Screening Program (http://www.epa.gov/endo/) would require a FFLC study as a Tier II test following a weight of evidence evaluation of any positive indicators of endocrine activity at Tier I (screening). Consequently, the demand to conduct FFLCs to support risk assessments for endocrine-active chemicals may increase in the future.

Currently, FFLC tests have not been internationally standardized and validated (intercalibrated), which is not surprising, considering the length, cost, and complexity of the studies, although the US Environmental Protection Agency (USEPA) and other jurisdictions have published guidance documents (Hansen et al. 1978; Benoit 1981; Anon 2002) based largely on tests conducted before the need for risk assessments for endocrine-active substances. Validation is particularly important for FFLC tests because they are difficult, time consuming, animal intensive, and expensive to conduct. In addition to this, some endpoints (e.g., fecundity) measured in these tests, while frequently sensitive, are inherently variable, producing noisy data with relatively low statistical power (Crane and Matthiessen 2007). Any simplifications or improvements in fitness-for-purpose of test design and aids to interpretation of data generated would be highly desirable.

FFLC tests generally begin with fertilized eggs (F0 generation), which are continuously exposed to the test substance until the eggs have developed into adults, a proportion of which (while still exposed) are allowed to breed and produce offspring (F1), which are followed until at least the swim-up (early fry stage) or an early life stage (e.g., 30 d after hatch). However, different designs are now also under consideration. In some cases, life cycle tests are continued until the point where the F1 generation is sexually differentiated (e.g., an extended 1-generation test). Two-generation tests have also been conducted, as have a few multigeneration tests which continue until the F2 generation reaches swim-up (reviewed in OECD 2008). These extended designs have been proposed to address the potential issues of maternal transfer of strongly bioaccumulative substances or endocrine-mediated transgenerational effects. There is limited evidence from a recent Detailed Review Paper (DRP) by the Organization for Economic Cooperation and Development (OECD) that prolonged tests of this type might be more sensitive to strongly bioaccumulative substances (OECD 2008). The evidence for transgenerational effects in these fish tests is also limited.

Although the FFLC test might be considered definitive in ecotoxicological terms (by covering all portions of the life
cycle), it should be recognized that the test does not precisely mimic reproduction, as would occur in the field (particularly for species with reproductive strategies which differ strongly from those used in tests), and, therefore, the results may be difficult to apply directly to field conditions. For example, species used for most FFLC tests perform best as pairs or small breeding groups, but many fish species (e.g., roach, *Rutilus rutilus*) breed in much larger groups, and are not constrained by the limited space available under laboratory conditions. There is also evidence that sperm release in some species (salmonids) is very sensitive to chemical interference, with the male olfactory epithelium able to detect female pheromonal prostaglandins (e.g., Moore and Lower 2001), but it is doubtful whether current FFLC tests take this process into account fully, or at all. Breeding behavior in the laboratory may, therefore, not be representative of some fish species in the wild, which could result in an important uncertainty given that some endocrine active chemicals alter sexual behavior (e.g., Martinović et al., 2007; Maunden et al. 2007). Laboratory fish also tend to be of limited genetic background when compared with their wild conspecifics (Coe et al. 2009), which could mask the individual variations in sensitivity that can occur. Fish in confined test chambers will experience an altered degree of stress in captivity compared with the field, as well as artificial social hierarchies, potentially modifying an individual’s sensitivity to toxicants. Conversely, wild fish populations may be subject to multiple “background” variables or stressors (aside from toxicity due to any particular chemical), which contribute to environmental stochasticity and the possibility of population decline (Brown et al. 2003). Environmental stressors (e.g., temperature fluctuations, competition, and predation) are absent in most laboratory life cycle fish exposures. Competition (for mates, food, or space) is not taken into account when fish are paired or placed in mating groups, and factors such as predation (of adults or young) are missing entirely. Finally, domestication effects lead to divergence of many life history traits between wild and cultured or laboratory populations (Thorpe 2004; Lorenzen 2005). However, these laboratory-to-field extrapolation uncertainties are accounted for, at least to some extent, in the risk assessment evaluation procedure by the use of assessment (safety) factors applied to the data derived from FFLC tests.

Published FFLC test data on endocrine active chemicals are sparse, and predominantly involve just 3 species: the Japanese medaka (*Oryzias latipes*), the fathead minnow (*Pimephales promelas*), and the zebrafish (*Danio rerio*). OECD (2008) lists 15, 3, and 7 life cycle tests with endocrine active chemicals for each species, respectively. In that document there are a further 16 life cycle test datasets for nonendocrine active chemicals with, predominantly, the fathead minnow, and limited additional data for sheephead minnow (*Cyprinodon variegatus*), flagfish (*Jordanella floridana*), brook trout (*Salvelinus fontinalis*), medaka, and zebrafish.

An additional uncertainty is the fact that few field-based experiments with fish populations exposed to endocrine active chemicals have been conducted, so the ability of FFLC tests to predict effects at the population level is largely unknown. Furthermore, the question whether natural fish populations have actually been damaged by endocrine active chemicals remains unresolved (Mills and Chichester 2005), despite abundant evidence that wild fish have inter alia been feminized by exposure to estrogens (e.g., World Health Organization 2002; Matthiessen 2006). A key reason for this apparent discrepancy may be the capacity of fish populations to compensate for reductions in reproductive output, survival or growth through density-dependent increases in other vital rates (Rose et al. 2001). The strength of compensatory processes differs between populations, but in general compensatory reserve is greater in juvenile survival than in adult growth or reproductive traits (Lorenzen 2008). This suggests that population abundance is more sensitive to ecotoxicological effects on adult life history traits than to effects on juvenile traits, with the exception of effects on sex ratio, which are caused during juvenile exposure. Consequently, test endpoints relating to adult growth, survival and reproductive traits (including those that are caused during juvenile sexual development) are particularly relevant to population level effects. In large populations in which density dependence acts mostly in a compensatory manner, effects on population abundance will generally be smaller than effects measured in the laboratory for particular life stages and vital rates. However, small populations may show depensatory density dependence (Allee effects; Allee 1931), such that ecotoxicological reductions in vital rates may be larger and possibly lead to catastrophic effects on population abundance.

One example for which limited field data exist concerns the synthetic estrogen ethynylestradiol (EE2), which is widespread in sewage effluents and surface waters at low ng/L concentrations. Kidd et al. (2007) conducted a whole-lake dosing experiment in Canada, in which EE2 was added to Lake 260 every year from May to October for 3 y, producing mean annual measured concentrations in epilimnetic waters of 6.1, 5.0, and 4.8 ng/L over the 3 years of dosing. Recruitment of fathead minnows substantially failed in the third year, possibly due to reduced fertility, and the population had almost disappeared by year 6 (i.e., 3 y after dosing ceased). It is instructive to compare these Canadian field data with laboratory-based life cycle tests that have been conducted with fathead minnows exposed to EE2 (Länge et al. 2001; Parrott and Blunt 2003), which have generated Lowest Observed Effect Concentrations (LOECs) for abnormal testicular development, fertilization success and sex ratio in the range <0.32-4 ng/L. Similar results have been obtained with the medaka (Balch et al. 2004), the zebrafish (Wenzel et al. 2001; Van den Belt et al. 2003), and the Chinese rare minnow *Gobioatyris rarus* (Zha et al. 2008). By constructing a life table based on the vital rates of survival and fecundity for fathead minnows from the study by Länge et al. (2001), Grist et al. (2003) used a Leslie matrix model to show that EE2 concentrations of 0.53-3 ng/L would reduce the intrinsic rate of population increase (r) by 20% and 100% compared with the control. In broad terms, therefore, this demonstrates reasonably good correspondence between the results of FFLC and field experiments, as confirmed by agreement with a Predicted No Effect Concentration (PNEC) of 0.35 ng/L derived from a species sensitivity distribution approach (Caldwell et al. 2008).

It is probable that FFLC tests will become part of a suite of tools to assess the potential ecological risk of endocrine-active chemicals. However, there are many options in terms of experimental design, species and endpoints. Unfortunately, little comparative information exists on which to base recommendations for appropriate FFLC test designs. This study describes how the technique of multi-criteria decision analysis (MCDA) was used to elicit the views of fish
METHODS

A 4-d workshop was convened in Palma, Mallorca, Spain, 23–26 September 2008, to which 25 fish ecologists and ecotoxicologists were invited, along with experts in statistics and environmental regulation. The first part of the workshop was spent considering what evidence is available on the relationship between fish reproduction and population-level effects in laboratory and field studies. The results of sensitivity and power analyses from Crane and Matthiessen (2007) were also presented at the workshop.

The workshop participants were asked to consider the evidence and then provide information for use in an MCDA to determine what endpoints in fish tests provide the optimum balance between sensitivity, power and ecological significance, at a study size and cost that is practical to implement. First they were asked to construct a “value tree,” which is a graphical representation of the different criteria that they wanted to use to appraise their different decision options. The value trees could be used to organize criteria into tiers (called subnodes) if required for clarity of analysis. The groups were then asked to score each of the options against each of the criteria, either directly on a preference scale of 0 to 100 or on an equivalent natural scale (e.g., cost) that was subsequently converted to a scale of 0 to 100. Finally they used the technique of swing weighting to assign weights to each of the criteria (Belton and Stewart 2002). This approach determines how a swing in weight from 0 to 100 for 1 criterion compares with a similar swing for another criterion. For example, swing weighting can be used to decide whether a difference in the range of cost per study endpoint from, say, $20 000 (least expensive, so scored 100) to $24 000 (most expensive, so scored 0) is more or less important than a difference in number of fish used per study endpoint of 1000 (lowest number, so scored 100) to 2000 (highest number, so scored 0). Individuals asked to consider this question might believe that cost per se is an important criterion when choosing between endpoints, but that the difference in cost between the least expensive and most expensive endpoints in this example is rather insignificant when compared with the difference between numbers of animals used to determine these endpoints.

The participants performed these tasks in 3 separate breakout groups—1 for each of the main small fish species used in FFLC tests (zebrafish, medaka, and fathead minnow, with each group comprising a selection of fish ecologists, ecotoxicologists, statisticians, and environmental regulators). These groups then reported back to plenary sessions. After considering the views of the other groups, the experts were asked to revisit their conclusions and produce a final view. This 2-stage approach is valuable, because it is common for there to be a learning process for participants during MCDA (Sparring and Tarbotton 2000).

The MCDA software program HiView Version 3.2 (Catalyze, Winchester, UK) was used to collate and analyze the data and produce the following outputs for each group: 1) A value tree, 2) A table of the final scores and weights, which identified the preferred endpoints and the criteria that contributed to these preferences, 3) Sensitivity analyses to determine what percentage change in weight would be required to change the preferred endpoint to an alternative endpoint. These changes were banded <5%, 5–15%, and >15% change in weight.

Fish test endpoints considered for this evaluation included the following: Time to hatch; Hatching success (number of embryos that complete hatching, expressed as a percentage of eggs deemed fertile); Fry survival; Growth; Condition factor [(weight × 100) / length3]; Sex ratio (macroscopic observation and histological confirmation); Sexually undifferentiated ratio (fish that have not sexually matured); Secondary sexual characteristics (e.g., fatpad and tubercles in fathead minnows, and anal fin shape and papillary processes in medaka, which are under endocrine control and can be enhanced, suppressed or induced); gonadosomatic index (gonad weight/body weight × 100); Gonad sex determination (sex determination based on gonads, rather than external appearance); Major histological abnormalities (e.g., intersex is the development of both ovarian and testicular tissues in the gonads); Intensive histology (a variety of alterations in gonad histology that may be associated with exposure to an endocrine active substance, and intensive techniques such as staging and potentially assessment of multiple tissues); Vitellogenin induction and/or reduction (a female egg yolk protein that can be induced in males following exposure to estrogenic substances or reduced in females in response to estrogen antagonists); Time to spawn (age at spawning or surrogate, i.e., time elapsed during study before spawning); Behavior (reproductive behavior, e.g., loss of territorial aggressiveness or spawning behavior); Pheromone production; Fecundity (number of viable eggs produced by females); Fertilization success (the number of fertile eggs, expressed as a percentage of the number of total eggs); and Population growth (expressed as the intrinsic rate of increase).

RESULTS

Fathead minnow

The fathead minnow group developed the final value tree shown in Figure 1. They identified 24 endpoints that could logically be measured in fathead minnow FFLC tests with endocrine-active chemicals: Time to hatch (F0, F1); Hatching success (F0, F1); Fry survival (F0, F1); Growth (F0, F1); Secondary sexual characteristics (F0, F1); Gonadosomatic index (F0, F1); Gonad sex determination (F0, F1); Major histological abnormalities (F0, F1); Intensive histology (F0, F1); Vitellogenin induction (F0, F1); Time to spawn (F0); Behavior (F0); Fecundity (F0); and Fertilization success (F1).
Scores for endpoints against the criteria in Figure 1, plus criteria weights derived by the group through the technique of swing weighting are shown in Table 1. The values in Table 1 (and also in Tables 2 and 3) should be interpreted as follows. The different potential FFLC test endpoints identified by the group are listed in the top row, and the criteria used to choose between these endpoints are listed in the first column. Values in bold are for higher node criteria (Tier 1) in the value tree that integrate any criteria below. Values in normal font are for Tier 2 criteria that are integrated by tier 1, and values in italics are for tier 3 criteria that are integrated by tier 2 (only the fathead minnow group decided to include tier 3 criteria). The scores from 0 to 100 given by the group to each endpoint against each criterion are shown in the columns under each endpoint. Finally, the swing weights given to each of the criteria are shown in column 2. Note that swing weights given to lower nodes propagate to higher nodes where they could again be checked for plausibility by the group.

The results in Table 1 show that the group weighted relative ecological significance of the different endpoints most highly, with differences in endpoint sensitivity to endocrine active substances weighted almost as highly. All other criteria (e.g., operating costs, statistical power, required level of laboratory expertise) for selecting endpoints were weighted as being considerably less important. F1 fertilization success emerged as the endpoint with the highest overall score, when criteria scores and weights were combined. This is because this endpoint scored relatively highly across several criteria, including those that were weighted most highly by the group. Other endpoints that scored highly overall were F0 and F1 secondary sexual characteristics, and F0 fecundity.

Sensitivity analysis showed that the most preferred endpoint of F1 fertilization success is sensitive to relatively small changes in weighting of less than 5%. Less than a 5% increase in weighting on test duration or field effect, or a 5–15% increase in weighting on endpoint measurement and number of fish used, would change the most preferred endpoint to F0 secondary sexual characteristics. A 5–15% decrease in weighting on laboratory expertise would change the most preferred endpoint to F0 time to hatch and F0 fecundity, respectively.

The lowest scoring endpoints were F0 and F1 intensive histopathology because of relatively low scores against all of the criteria.

In summary, for fathead minnow, the workshop group identified F1 fertilization success, F0 fecundity, F0 time to hatch, and F0/F1 secondary sexual characteristics as the most preferred endpoints in fathead minnow FFLC tests, primarily on the basis of their high ecological significance (see Discussion section for further assessment) and their sensitivity to endocrine active substances.

**Medaka**

The medaka group developed the final value tree shown in Figure 2. They identified 25 endpoints measured in medaka FFLC tests: Time to hatch (F0, F1); Hatching success (F0, F1); Fry survival (F0, F1); Growth (F0, F1); Sex ratio (F0, F1); Major histological abnormalities (F0, F1); Vitellogenin induction (F0, F1); Behavior (F0, F1); Population growth (F0, F1); Pheromone production (F0); Time to spawn (F0); Fecundity (F0); Condition factor (F0); Gonadosomatic index (F0, F1); and Fertilization success (F1).
Table 1. MCDA scores and weights for fathead minnow FFLC endpoints. Total scores and weights are for the highest nodes ( tier 1) in the value tree. Subnodes that contribute to the main nodes are in normal text ( tier 2) or italicized text ( tier 3).

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Validation costs</th>
<th>Operating costs</th>
<th>Test duration</th>
<th>Endpoint measurement</th>
<th>Public acceptability</th>
<th>Field effects</th>
<th>Test ethics</th>
<th>Fish number</th>
<th>Severity</th>
<th>Laboratory expertise</th>
<th>Ecological significance</th>
<th>EAS sensitivity</th>
<th>Statistical power</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10 100 100 100 80 100 80 60 40 80 80 0 75 100 100 100 100 100 80 100 80 60 40 80</td>
<td>20 100 100 88 52 52 52 40 34 28 43 50 32 18 18 30 30 30 24 24 24 12 6 0 15</td>
<td>14 100 99 83 40 40 40 40 40 40 29 29 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td>6 100 100 100 80 80 80 40 20 0 50 100 40 60 60 100 100 100 80 80 80 40 20 0 50</td>
<td>40 79 84 14 55 65 55 65 65 55 31 55 72 60 51 46 51 5 46 56 46 56 56 46 22</td>
<td>4 0 50 50 0 100 0 100 100 0 100 0 50 50 50 0 50 50 0 100 0 100 100 0 100</td>
<td>36 87 87 10 61 61 61 61 61 61 23 61 74 61 51 51 51 0 51 51 51 51 51 51 13</td>
<td>13 100 29 29 29 29 29 29 29 29 29 29 0 0 0 0 0 0 0 0 0 0 0 0</td>
<td>23 80 80 0 80 80 80 80 80 80 20 80 100 80 80 80 0 80 80 80 80 80 80 80 20</td>
<td>40 100 100 100 100 80 100 60 40 20 70 100 0 100 100 100 100 100 100 80 100 60 40 20 70</td>
<td>100 80 100 100 80 90 70 80 30 0 0 80 90 100 100 80 100 100 80 90 70 80 30 0 0</td>
<td>90 0 0 0 20 90 40 90 100 60 100 20 80 100 90 0 0 0 20 90 40 90 100 60 100</td>
<td>50 100 0 0 0 50 50 0 50 50 50 0 50 100 50 0 0 50 50 0 50 50 0 50 50 0 50 50</td>
<td>350 66 65 50 52 78 61 65 57 34 49 65 63 76 80 58 58 45 49 75 75 58 62 54 31 46</td>
</tr>
</tbody>
</table>
Table 2. MCDA scores and weights for medaka FFLC endpoints. Total scores and weights are for the highest nodes (tier 1) in the value tree. Subnodes that contribute to the main nodes are in normal text (tier 2) or italicized text (tier 3).

| Criteria | Criteria weights | F0 time to hatch | F0 hatching success | F0 fry survival | F0 growth | F0 sex ratio | F0 major histological abnormalities | F0 time to spawn | F0 fecundity | F0 condition factor | F0 gonadosomatic index | F0 fertilization success | F0 hatching success | F1 time to hatch | F1 hatching success | F1 fry survival | F1 growth | F1 sex ratio | F1 major histological abnormalities | F0/F1 Biomarker effects (vitellogenin) | F0/F1 Behavioral effects | F0/F1 Pheromone effects | F0/F1 Population level effects (e.g.,) |
|----------|------------------|-----------------|---------------------|-----------------|----------|-------------|-----------------------------|-----------------|----------------|------------------|----------------------|------------------|----------------|-----------------|-------------------|----------------|-------------|-------------------------------|-----------------------|---------------------|---------------------|-------------------------|
| Financial | 20               | 100             | 100                 | 90              | 85       | 65          | 50                          | 66              | 79             | 79               | 39                   | 39               | 39            | 21              | 16                | 15             | 0           | 78                           | 78                   | 74                  | 37                  |                         |
| Fish number | 4                | 100             | 100                 | 100             | 100      | 100         | 100                         | 100             | 100            | 100               | 0                    | 0                | 0             | 0               | 0                 | 0               | 0           | 100                          | 100                   | 100                 | 100                 |                         |
| Technician time | 6               | 100             | 100                 | 88              | 88       | 24          | 24                          | 24              | 44             | 44               | 68                   | 68               | 36            | 36              | 36                | 8               | 0           | 0                             | 68                   | 68                 | 68                  | 8                      |
| Lab costs | 10               | 100             | 100                 | 86              | 77       | 74          | 46                          | 78              | 62             | 76               | 77                   | 57               | 57            | 57              | 57                | 38             | 31          | 0                             | 76                   | 76                 | 68                  | 30                      |
| Non financial | 260              | 61              | 66                  | 61              | 58       | 59          | 55                          | 57              | 50             | 51               | 36                   | 78               | 61            | 65              | 59                | 46             | 58          | 51                           | 29                   | 28                 | 8                   | 61                      |
| Ecological relevance | 100             | 60              | 93                  | 87              | 73       | 73          | 47                          | 93              | 100            | 73               | 7                    | 100              | 60            | 93              | 87                | 73             | 60          | 33                           | 0                    | 53                 | 13                  | 100                     |
| Public explainability | 10              | 21              | 37                  | 37              | 79       | 100         | 89                          | 37              | 68             | 58               | 95                   | 68               | 5             | 37              | 26                | 37             | 96          | 89                           | 0                    | 26                 | 0                   | 23                      |
| Statistical power | 90              | 100             | 74                  | 59              | 49       | 49          | 59                          | 29              | 0              | 39               | 39                   | 74               | 100           | 74              | 59                | 23             | 49          | 59                           | 29                   | 18                 | 8                   | 13                      |
| EAS sensitivity | 50              | 0               | 0                   | 18              | 35       | 41          | 53                          | 35              | 35             | 35               | 82                   | 54               | 13            | 5               | 18                | 35             | 71          | 71                           | 100                  | 0                   | 0                   | 71                      |
| Risk of false positives | 10              | 67              | 67                  | 83              | 67       | 67          | 67                          | 67              | 17             | 17               | 17                   | 17               | 17            | 17              | 17                | 17             | 17          | 17                           | 0                    | 0                   | 0                   | 100                     |
| TOTAL | 280              | 64              | 68                  | 63              | 60       | 60          | 54                          | 57              | 51             | 53               | 39                   | 39             | 39            | 39              | 39                | 39             | 39          | 39                           | 39                   | 39                 | 39                  | 39                      |
Table 3. MCDA scores and weights for zebrafish FFLC endpoints. Total scores and weights are for the main nodes in the value tree (bold text). Total scores and weights are for the highest nodes (tier 1) in the value tree. Subnodes that contribute to the main nodes are in normal text (tier 2).

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Costs</th>
<th>Biomarkers</th>
<th>Histopathology</th>
<th>Time to spawn</th>
<th>Fecundity</th>
<th>Fertilization success</th>
<th>Time to hatch</th>
<th>Hatching success</th>
<th>Sex ratio</th>
<th>Undifferentiated ratio</th>
<th>Survival</th>
<th>Length</th>
<th>Weight</th>
<th>Behavior</th>
<th>Overall “fitness”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Costs</td>
<td>220</td>
<td>41</td>
<td>27</td>
<td>49</td>
<td>66</td>
<td>87</td>
<td>65</td>
<td>82</td>
<td>74</td>
<td>68</td>
<td>82</td>
<td>46</td>
<td>53</td>
<td>67</td>
<td>64</td>
</tr>
<tr>
<td>Study duration</td>
<td>10</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>25</td>
<td>25</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Operating costs</td>
<td>40</td>
<td>71</td>
<td>0</td>
<td>94</td>
<td>63</td>
<td>46</td>
<td>89</td>
<td>89</td>
<td>89</td>
<td>89</td>
<td>85</td>
<td>82</td>
<td>100</td>
<td>96</td>
<td>89</td>
</tr>
<tr>
<td>Laboratory capability/capacity</td>
<td>10</td>
<td>38</td>
<td>6</td>
<td>75</td>
<td>88</td>
<td>75</td>
<td>88</td>
<td>25</td>
<td>25</td>
<td>100</td>
<td>75</td>
<td>100</td>
<td>25</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>False positive results</td>
<td>50</td>
<td>85</td>
<td>69</td>
<td>92</td>
<td>100</td>
<td>100</td>
<td>92</td>
<td>100</td>
<td>92</td>
<td>8</td>
<td>100</td>
<td>0</td>
<td>23</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>False negative results</td>
<td>50</td>
<td>0</td>
<td>46</td>
<td>15</td>
<td>77</td>
<td>92</td>
<td>85</td>
<td>92</td>
<td>31</td>
<td>92</td>
<td>92</td>
<td>100</td>
<td>92</td>
<td>92</td>
<td>92</td>
</tr>
<tr>
<td>Public’s views</td>
<td>60</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>100</td>
<td>0</td>
<td>50</td>
<td>100</td>
<td>100</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Benefits</td>
<td>310</td>
<td>45</td>
<td>41</td>
<td>71</td>
<td>60</td>
<td>75</td>
<td>39</td>
<td>50</td>
<td>94</td>
<td>69</td>
<td>56</td>
<td>82</td>
<td>82</td>
<td>38</td>
<td>28</td>
</tr>
<tr>
<td>Specificity to EAS</td>
<td>20</td>
<td>100</td>
<td>100</td>
<td>10</td>
<td>10</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>10</td>
<td>10</td>
<td>80</td>
<td>0</td>
</tr>
<tr>
<td>Sensitivity to EAS</td>
<td>70</td>
<td>0</td>
<td>20</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>73</td>
<td>100</td>
<td>20</td>
<td>100</td>
<td>73</td>
<td>73</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ethical considerations</td>
<td>5</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>Ability to extrapolate across environments</td>
<td>95</td>
<td>100</td>
<td>80</td>
<td>100</td>
<td>40</td>
<td>80</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>80</td>
<td>0</td>
<td>100</td>
<td>50</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>Regulatory relevance</td>
<td>20</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Population relevance</td>
<td>100</td>
<td>0</td>
<td>13</td>
<td>75</td>
<td>100</td>
<td>100</td>
<td>75</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>50</td>
<td>75</td>
</tr>
<tr>
<td>TOTAL</td>
<td>530</td>
<td>43</td>
<td>35</td>
<td>62</td>
<td>63</td>
<td>80</td>
<td>50</td>
<td>64</td>
<td>85</td>
<td>69</td>
<td>67</td>
<td>67</td>
<td>70</td>
<td>50</td>
<td>43</td>
</tr>
</tbody>
</table>
Scores for endpoints against the criteria in Figure 2, plus criteria weights applied by the group are shown in Table 2. This shows that, like the fathead minnow group, the medaka group weighted ecological significance between the different endpoints most highly, with differences in statistical power being weighted almost as highly. All other criteria for selecting endpoints were weighted as being considerably less important.

F1 fertilization success emerged as the endpoint with the highest overall score, when criteria scores and weights were combined, which was also the case for the fathead minnow group. Other endpoints that scored relatively highly overall were F0 and F1 hatching success. Sensitivity analysis showed that the most preferred endpoint of F1 fertilization success is sensitive to changes in weighting of 5–15%. A 5–15% increase in weighting on fish numbers, technician time or laboratory costs would change the most preferred endpoint to F0 hatching success. A 5–15% decrease in weighting on endocrine active substance sensitivity would also change the most preferred endpoint to F0 hatching success. This is because the latter endpoint is generated at the beginning of the study and is attractive because of its low operational cost.

The lowest scoring endpoints were effects on vitellogenin induction, behavior and pheromones, although in some cases (e.g., pheromone production) it was argued by group members that this could be due to a lack of experience with and knowledge of these endpoints.

In summary, for medaka, the workshop group identified F1 fertilization success and F0 and F1 hatching success as the most preferred endpoints in medaka FFLC tests, primarily on the basis of their relatively high ecological significance and statistical power (see Discussion section for further assessment). The higher ranking of F1 fertilization success was also based on the understanding that the cost of a FFLC study is outweighed by the importance of completing the study and product approval process on schedule and avoiding any delay to market.

**Zebrafish**

The zebrafish group developed the final value tree shown in Figure 3. In contrast to the other 2 groups, this group combined the evaluation of endpoints in F0 and F1 generations, to prevent F1 endpoints being unfairly affected by certain criteria. This resulted in 14 combined generic endpoints: Biomarkers (vitellogenin and the male hormone, 11-ketotestosterone, plus specific biomarkers for other endocrine modes of action if these emerge in the future); Histopathology (full body, i.e., multiple organs); Time to spawn; Fecundity; Fertilization success; Time to hatch; Hatching success; Sex ratio; Sexually undifferentiated ratio; Survival; Length; Weight; Behavior (e.g., spawning behavior); and “Fitness.” (The group tried to consider endpoints which may be of ecological relevance but are currently not considered or tested yet. An example is the influence of a chemical on the response to stress, which could be tested as a laboratory challenge test, in which the organisms react to stimuli. This type of endpoint was referred to as “fitness.”)

Scores for endpoints against the criteria in Figure 3, plus criteria weights applied by the group are shown in Table 3. This shows that, like the other groups, the zebrafish group weighted the differences in ecological significance (which they called “population relevance”) between the different endpoints most highly, with differences in the ability to extrapolate results across different environments being
weighted almost as highly. Sensitivity to endocrine active substances also received a relatively high rating. All other criteria for selecting endpoints were weighted as being considerably less important.

Sex ratio emerged as the endpoint with the highest overall score, when criteria scores and weights were combined. This was mainly due to the recognized and irreversible effects of endocrine active chemicals on reproductive organs (e.g., Nash et al. 2004) and the potential negative effects that would result at the population level. However, it was noted that the genetic basis of sex determination in zebrafish is only partially understood (Jørgensen et al. 2008) and sex ratio may be variable even under controlled laboratory conditions according to recent ring test data for the Fish Sexual Development Test. Plasticity in sex ratio appears to depend on numerous environmental factors, including food availability and temperature during critical early development and also degree of inbreeding versus outbreeding (see Lawrence et al. 2008).

In common with medaka and fathead minnow, the other endpoint that scored relatively highly overall was F1 fertilization success. Sensitivity analysis showed that the most preferred endpoint of sex ratio is sensitive to changes in weighting of 5–15%. A 5–15% increase in weighting on study duration, laboratory capability or capacity, or false negative results would change the most preferred endpoint to F1 fertilization success. A 5–15% decrease in weighting on regulatory relevance would also change the most preferred endpoint to F1 fertilization success. The lowest scoring endpoints were effects on biomarkers, histopathology and overall fitness.

In summary, for zebrafish, the expert group identified sex ratio and F1 fertilization success as the most preferred endpoints in zebrafish FFLC tests, primarily on the basis of their relatively high ecological significance (see Discussion section for further assessment), the ability to extrapolate results across different environments, and sensitivity to endocrine active substances.

**DISCUSSION**

It is interesting to note that the 3 groups constructed different value trees. This does not appear to be due to any basic differences in the FFLC test under consideration, but instead reflects differences in the views of individuals within each group about how to frame the overall question. This illustrates the importance of taking the existence of such different views transparently into account, even when a decision is being informed by apparently objective experts.

However, despite developing initial value trees, criteria, scores and weightings independently, all 3 workshop groups arrived at rather similar overall conclusions. Fertilization success (F1) emerged as a high priority measurement endpoint from all groups. However, it became apparent subsequent to the workshop that participants were using different definitions of the term fertilization success. As a result, this endpoint scored highly because it was considered to be both ecologically relevant and of high statistical power, when in fact each of these is an attribute of 2 different definitions. Fertilization success, defined as the fertile proportion of the total number of eggs, does have high statistical power, because it is expressed as a percentage and is, therefore, normalized. However, this definition of fertilization success is not associated with high ecological relevance (e.g., a fertilization success of 90% could be 9 of 10 or 900 of 1000 eggs). The other definition used, which more accurately may be termed fertility, is the number of fertile eggs (e.g., the value of 9 or 900 eggs in the previous example, rather than the proportion). This endpoint does have high ecological relevance, but is of low statistical power, because it has been shown to be a highly variable endpoint. Fish full lifecycle studies should, therefore, be designed to gather data on both endpoints as efficiently and effectively as possible (as one is a function of the other). In terms of showcasing how MCDA can be used in this context, this provides a good example of how definitions need to be very clear at the outset so that all participants feed into the process with a common understanding.

---

**Figure 3.** Value tree for criteria when selecting zebrafish endpoints. EAS = endocrine active substance.
Some other preferred endpoints identified by the MCDA scored highly only in particular species, such as development of secondary sexual characteristics in fathead minnow and sex ratio in zebrafish. Indeed, sex ratio in zebrafish was considered by that group to be the most preferred endpoint, mainly due to the recognized effects of endocrine active chemicals on reproductive organs (e.g., Nash et al. 2004). Although not yet possible in zebrafish, the ability to determine sex genetically and compare this with the phenotypic sex greatly improves the statistical power and certainty in an effect. It could also reduce the number of excess fish which would otherwise be required to ensure equal numbers of males and females. Genetic sex probes already exist for the medaka (Matsuda et al. 2002; Nanda et al. 2002) and there is also the d-rR strain that possesses sex-linked pigmentation, which distinguishes genotypic sex (Aida 1921). Recent research in the United States and Denmark suggests that measurement of genetic sex in fathead minnow may soon be possible (A. Olmstead and G. Ankley, USEPA, personal communication). Other relatively highly preferred endpoints such as hatching success are also likely to be worth measuring in tests with all fish species.

The different groups were also in broad agreement about the least preferred endpoints in FFLC tests. Intensive histopathology consistently ranked low, as did measurement of biomarkers such as vitellogenin. These types of diagnostic endpoints can lend important insights into toxic mechanism of action, but are typically of less utility for predicting adverse effects in individuals or populations. Research has been undertaken recently to address this shortcoming and these endpoints may have potential in the future long-term use of FFLCs (Miller et al. 2007). Currently, in tiered testing programs for endocrine-active chemicals, diagnostic biomarker-type endpoints initially are used to flag chemical mechanisms of concern, while high-tier (e.g., FFLC) tests are intended to generate the type of population-relevant (reproduction) data needed for determination of ecological risk. This type of emphasis reflects an approach in which biomarkers are used as supporting evidence rather than directly in the risk assessment procedure (Hutchinson et al. 2006). It is, therefore, likely to be an inefficient use of available resources to require routine measurement of these endpoints in all FFLC studies, whose purpose is to ascertain population relevant impacts. That being said, in specific case studies or for specific regulatory requirements, the integration of population relevant endpoints and diagnostic biomarkers may be justified.

SUMMARY AND CONCLUSIONS

In summary, using the logical framework of MCDA, a limited number of preferred FFLC test endpoints for assessing endocrine active substances were identified by 3 workshop groups with expertise in statistics, fish ecology and aquatic toxicity testing with fathead minnow, medaka, and zebrafish. Effects on F1 fertilization rate emerged as a generally preferred endpoint for all species, and also highlighted the need for clear definitions and a common understanding of all terms used in MCDA. Some endpoints scored highly for particular species, such as development of secondary sexual characteristics in fathead minnow and sex ratio in zebrafish. Indeed, sex ratio in zebrafish was considered by the group evaluating that species to be the most preferred endpoint due to its environmental (population) relevance. Other relatively highly preferred endpoints such as hatching success would also be worth measuring in tests with any of the fish species.

There was also broad agreement about the least preferred endpoints in FFLC tests. Whereas histological confirmation of gonadal sex was considered important, intensive histopathology consistently ranked much lower, as did measurement of biomarkers such as vitellogenin. It is, therefore, likely to be an inefficient use of available resources to insist on routine measurement of these latter endpoints.

In conclusion, our analyses are preliminary and could certainly be refined further through discussion with a wider range of stakeholders, particularly to determine what swing weights should be allocated to criteria. For certain criteria there was considerable disagreement (e.g., public acceptability of testing) and further discussion matched with a public questionnaire would enable more uniform weighting to this criterion to be assigned. A decrease in the variability of swing weight scoring for this and other criteria would improve the MCDA results further. However, the examples presented here illustrate how MCDA can be used to provide a logical framework for scientists and regulators in which different options are identified, the criteria for choosing between them are discussed, scored, and weighted, and the sensitivity of the final results can be judged. In an economic environment in which government and industry scientists are continually required to do more with fewer resources, MCDA is a tool that can help identify essential and optimal ways forward. A clear and transparent audit trail is available when the MCDA process is followed, allowing anyone to return to and amend input data in the light of further scientific data, or different opinions about subjective criteria. This stands in contrast to the often opaque and untraceable outputs that can result from unstructured “expert judgment.”

Acknowledgment—This work was funded by the Cefic Long Range Research Initiative (Project EMSG 47). The conclusions and recommendations made in this study reflect the views of the authors as individual scientists and do not represent a position of the organizations to which the authors are affiliated. We thank 2 anonymous peer reviewers for helpful comments on the original manuscript.

REFERENCES


