Original Article

Challenges and Approaches to Conducting and Interpreting the Amphibian Metamorphosis Assay and the Fish Short-Term Reproduction Assay

Katherine Kemler Coady,* Christine Marie Lehman, Rebecca J Currie, and Troy Alan Marino

The Dow Chemical Company, Midland, Michigan

The amphibian metamorphosis assay (AMA) and the fish short-term reproduction assay (FSTRA) are screening assays designed to detect potential endocrine activity of a test substance. These assays are included in a battery of assays in Tier 1 of U.S. Environmental Protection Agency's Endocrine Disruptor Screening Program. Based on our laboratory's experience with these two assays, we have noted several challenges in the conduct and interpretation of the AMA and FSTRA, including, but not limited to, diseased/parasitized test organisms, failure to meet some guideline performance criteria, and issues selecting and maintaining test concentrations. Various approaches are described for addressing the challenges associated with both the conduct and interpretation of these assays. Historical control data for both the AMA and FSTRA are presented to further understand background occurrences of histopathological phenomena and variability associated with the measured endpoints in these assays. In the historical control database for the AMA, wet weight on day 7 was the most variable endpoint (coefficient of variation = 26%), while developmental stage on day 21 was least variable (coefficient of variation = 0.47%). In the FSTRA, vitellogenin concentrations were the most variable endpoint (coefficient of variation = 47–84%), while fertility was the least variable endpoint (coefficient of variation = 1.5%) among historical controls. *Birth Defects Res (Part B)* 101:80–89, 2014. © 2013 Wiley Periodicals, Inc.

Key words: ***endocrine; Xenopus laevis; fathead minnow; thyroid; estrogen; androgen; EDSP***

INTRODUCTION

The 1996 Food Quality Protection Act directed the U.S. Environmental Protection Agency (USEPA) to institute a screening program to determine whether certain substances may have hormonal effects. In addition, a 1996 amendment to the Safe Drinking Water Act authorized the USEPA to screen substances that may be found in sources of drinking water for endocrine disrupting potential. In response, the USEPA developed the Endocrine Disruptor Screening Program (EDSP), a two-tiered system to screen and test for endocrine disrupting compounds. Tier 1 was designed as a screen to identify chemicals having the potential to interact with the endocrine system. This tier includes five in vitro assays, four mammalian in vivo assays, and two ecotoxicology in vivo assays. There is some redundancy among assays to minimize false negative results and aid in mode of action (MoA) determinations. These Tier 1 screening assays, examined in conjunction with existing information in a weight of evidence evaluation, are used to determine whether the test chemical has potential activity in select endocrine pathways (i.e., estrogen, androgen, thyroid). If the collective data indicate the need for additional information, EDSP Tier 2 testing may be required. While Tier 1 assays are used as screens to determine the potential for endocrine activity, Tier 2 assays are longer term, often multigenerational studies in bird, invertebrate, amphibian, fish, and mammalian species that are meant to establish endocrine-related effects caused by each chemical to obtain information about effects at various doses.

In April 2009, the USEPA released an initial list of chemicals, which included pesticide active and pesticide inert chemistries, scheduled for Tier 1 EDSP screening. Official test orders for initiating Tier 1 screens of chemicals on this initial list were released in late 2009 and early 2010.

Two of the 11 assays in Tier 1 are ecotoxicological: the amphibian metamorphosis assay (AMA) and the fish short-term reproduction assay (FSTRA). Our laboratory conducted eight AMA assays and seven FSTRA assays during 2010 and 2011 using chemicals on the initial April 2009 list. In addition, we have conducted several positive and negative validation assays for both the AMA and

*Correspondence to: Katherine Coady, The Dow Chemical Company, 1803 Building, Washington Street, Midland, MI 48674. E-mail: kcoady@dow.com

Received 13 August 2013; Accepted 13 September 2013

Published online in Wiley Online Library (wileyonlinelibrary.com/journal/bdrb) DOI: 10.1002/bdrb.21081

FSTRA. Throughout the preparation, conduct, and interpretation of these Tier 1 screening assays, technical challenges were encountered. The purpose of this article is to discuss these challenges and, when possible, strategies to minimize or overcome them. The intent is that this information would be useful, not only to governmental agencies evaluating the outcomes of the Tier 1 testing, but also to laboratories conducting or preparing to conduct Tier 1 EDSP screening.

AMA Design

The phenomenon of amphibian metamorphosis is tightly coupled with thyroid function (Shi, 2000; Denver et al., 2002). Thus, amphibians are a developmental model with which to detect substances having the potential to disrupt the hypothalamic–pituitary–thyroid (HPT) axis. The AMA is the only Tier 1 assay using animals undergoing early morphological development (two other assays, the male and female pubertal rat assays, examine thyroid function only during sexual development). The HPT axis is fairly well-conserved among vertebrate taxa. In fact, results of assays focused on the thyroid pathway are generally very similar among the common mammalian model (the rat) and the amphibian (Pickford, 2010). Both the USEPA and the Organisation for Economic Co-operation and Development (OECD) have guidance documents outlining the conduct of the AMA (OECD, 2009a; USEPA, 2009a).

The AMA is designed to expose developing tadpoles to at least three waterborne concentrations of a test chemical plus control water for 21 days. The study begins with African clawed-frog tadpoles (*Xenopus laevis*) at Nieuwkoop and Faber (1994) stage 51, which is approximately 2 weeks posthatching. At this stage, also known as premetamorphosis, the tadpole thyroid is not yet functional. On study day 7, which is approximately when the thyroid gland begins functioning, five individual tadpoles are removed from each replicate tank (four tanks/exposure concentration). Endpoints (i.e., hindlimb length, snout-vent length, developmental stage, and wet weight) are collected on these individuals. Data from day 7 of the AMA are collected primarily to look for thyroid hormone agonists. Before day 7, the tadpole thyroid is still essentially nonfunctional so advanced development relative to controls can be attributable to compounds with this MoA, which are few in number (OECD, 2007a). Following the day 7 sampling, the remaining tadpoles are maintained in the test system for another 14 days. On day 21, the study is terminated and data are collected. Endpoint data collected on day 21 are the same as those collected on day 7, with the addition of thyroid histopathology that is performed on five tadpoles per replicate tank. By day 21, tadpoles have fully functional thyroid glands capable of secreting thyroid hormone, thus, HPT antagonists are clearly detectable at the day 21 time point. Hindlimb length (normalized by snout-vent length), developmental stage, and histopathology of the thyroid gland are more specific endpoints that are intended to give direct information on potential thyroid perturbations in tadpole metamorphic rate, while survival, wet weight, and snout-vent length serve as more generalized endpoints of tadpole growth and health in the assay.

AMA Study Initiation and Conduct

A high number of test organisms are used in the AMA. The test is initiated using 20 tadpoles in each of four replicate tanks per treatment level. Both the OECD and USEPA guidelines for the AMA indicate that tadpoles from the same spawning event should be used to stock a given AMA study (OECD, 2009a; USEPA, 2009a). Therefore, 320 tadpoles from a single spawning event are needed for a standard AMA, while more are required if a fourth dose or a solvent is used.

Most times, many more animals are generated for a study than are used in testing to ensure that sufficient tadpoles are available for test initiation. Logistic challenges arise when coordinating an adequate supply of tadpoles from a single spawn at the appropriate developmental stage (51) and age ($\leq$17 days) with the initiation of the study, which is often dependent on analytic confirmation that the test solution concentrations are accurate in the flow-through diluter test system. Test solution concentrations within the desired range of nominal must be established in the diluter system before adding the tadpoles to the test system and initiating the AMA. In our laboratory, it has been to our advantage to generate several cohorts of tadpoles about a week apart in age, then, if problems arise with the test system and delays are expected, the latter (younger) cohort of tadpoles can be used to initiate a study a week later without more extensive delays. Space limitations can be a factor when housing multiple tadpole cohorts; however, tight timelines and study logistics generally demand several tadpole cohorts be available for planned testing.

Before the initiation of the study, tadpoles are handled to assess tadpole stage, which is determined primarily by examining the hindlimb morphology. To stage a tadpole, it must be removed from culture, examined briefly under a dissecting scope, and then placed into a pool of acceptable tadpoles that are then randomly placed into exposure tanks. Mortality during the test is frequently observed within the first 24 hr after initiation, and is likely attributable to the handling stress of staging and placement in the study. Mortality >10% in the controls leads to an invalid test, so minimizing handling stress during initiation is vital. In our laboratory, anesthesia is not used during the tadpole staging process and care is taken to quickly examine tadpoles and transfer the tadpoles from one vessel to another with the use of large diameter transfer pipettes or small cups, when possible. The use of nets is minimized in the handling process as netting can cause damage to very small tadpoles.

Tadpole growth and developmental rates are greatly influenced by tadpole diet, so, feeding each tadpole a specific volume of food is important. Thus, the amount of food may vary from tank to tank depending on tadpole mortality across replicates. Both the OECD and USEPA AMA guidelines are specific with respect to the type of food (Sera Micron, Sera North America, Montgomeryville, PA) and the feeding rates. Our experience has shown that tadpoles grow and develop as expected when fed Sera Micron at the suggested rate in the OECD and USEPA AMA guidelines. Because thyroid function, and therefore, tadpole development, is dependent on iodide, ensuring tadpoles receive iodide is important. The

iodide content in Sera Micron was quantified in our labs and found to be 50 μg/g. However, iodide content in the laboratory dilution water obtained from Lake Huron was below the method detection limit of 10 μg/l. Since tadpoles were developing along the expected timeline, the measured iodide content in the Sera Micron feed was considered sufficient for promoting tadpole growth and development, even though iodide levels in the water supply were below the method detection limit.

AMA Endpoints

In our experience, the day 7 endpoints of hindlimb length, snout-vent length, and wet weight were more variable than those same endpoints on day 21, likely due to increased variability at earlier stages of growth (Table 1). Examining the endpoint results from control groups across multiple AMA studies ($n = 10$), we found the average coefficients of variations (CVs) for day 7 hindlimb length, snout-vent length, and wet weight were 12, 9.5, and 26%, respectively (Table 1). On day 21, the average CVs for hindlimb length, snout-vent length, and wet weight were 6.2, 2.7, and 7.5%, respectively. The CVs for length and weight measurements from control tadpoles in our laboratory are in line with the CVs reported in the integrated summary report for the AMA which were <15% for snout-vent measurements and <30% for wet weight measurements (OECD, 2007a).

If measuring hindlimb length and snout-vent length from a digital photograph of the tadpole, which is recommended by both the OECD and USEPA AMA guidelines, the accuracy of these measurements relies on tadpole/froglet position and placement relative to the dissecting scope platform and camera lens. We conducted trials in our laboratory in which we repeatedly placed and photographed the same tadpole on the dissection microscope; snout-vent and hindlimb measurements from these replicate pictures were then recorded. We found measuring to 0.1 mm for snout-vent length to be reasonable based on the precision of the measure due to variability in placement on the microscope for imaging. Because the vent can be difficult to see clearly at times, snout-vent length may be challenging to measure, therefore, measuring to the base of the abdomen may be an easier endpoint to capture. Using the base of the abdomen as a length measure has been approved by USEPA (http://www.epa.gov/endo/pubs/toresources/faqs.htm). Whichever method is selected for measuring snout-vent length, this method should be consistent within in a study and it is preferable to remain consistent between studies to establish a robust historical control data set for this parameter.

Developmental stage. While analysis of control responses from the AMAs in our laboratory indicated that developmental stage on both days 7 and 21 tended to be the least variable endpoint (mean CVs of 1.0 and 0.47%, respectively; Table 1), some cohorts of tadpoles had higher variability in developmental stage at day 21 than others. This is important, because one of the performance criteria described in the guidelines for the AMA indicates that the 10th and 90th percentile of the developmental stage distribution in the controls should not differ by more than four stages. However, 20% of the AMAs conducted in our laboratory (2 of 10 at the time of writing) had control treatments with a span of five developmental stages between the 10th and 90th percentiles. The reason for the slightly greater developmental spread in the controls in some studies but not others is unknown. Because there are slight variations among tadpole limb size and shape at stage 51, a way to minimize developmental spread is to carefully screen Nieuwkoop and Faber (NF) stage 51 tadpoles so there is close agreement among tadpoles used to initiate the study. From NF stage 51 to NF stage 58, the time to progress from one developmental stage to the next ranges from 2 to 6 days; however, the time it takes to progress from one developmental stage to the next in NF stages beyond 58 generally ranges from 1 to 2 days (Nieuwkoop and Faber, 1994). Given that beyond NF stage 58 the number of days between stages decreases, initiating studies with early stage 51 tadpoles may ensure that most tadpoles will be at relatively earlier stages of development, thereby minimizing the developmental stage span. It should also be noted that there is subjectivity involved with developmental staging, making it important that the same individual is available to stage the tadpoles at both day 7 and at study termination.

Thyroid histopathology. The test guidelines state that if advanced or asynchronous development is observed, the substance is considered thyroid active and histopathological examination of the thyroid tissues is not necessarily warranted. However, because histopathology is a very specific endpoint for determining HPT activity, it is often valuable to examine tadpole thyroids histologically even if not directly dictated by AMA guidance. Tadpole thyroid histopathology should follow specified guidance (OECD, 2007b; Grim et al., 2009) and be conducted by certified and trained personnel. Furthermore, stage-matching (i.e., examining tadpoles of a similar developmental stage) and knowledge of how thyroid histopathology alters with each developmental stage are both important when histologically examining thyroid glands because of changes in follicular cell height that correlate with development (Grim et al., 2009). It is also important to understand background levels of follicular cell hypertrophy and mild hyperplasia that may occur. In 11 of the 12 studies conducted in our laboratory, mild (grade 1) follicular cell hypertrophy was observed in thyroid glands of control tadpoles. Mild follicular cell hyperplasia was also observed in thyroid glands from control tadpoles; however, generally this observation was nonremarkable (<20% of the tissue affected). Although thyroid histopathology is a useful endpoint, these data are not typically statistically analyzed, and interpretation relies on expert judgment that can be subjective, making it important to have the same individual read all the histopathology slides within a single study so they are scored similarly. Our laboratory practices a peer-review process, in which the slides are initially read by a board certified pathologist, and then checked for accuracy and consistency in a semiblind fashion by another board certified pathologist.

Other Considerations

An issue with the use of *X. laevis* tadpoles is the occurrence of bent tails (scoliosis), which can occur at

Table 1
Control Performance in the Amphibian Metamorphosis Assay

Endpoint	Number of studies	Range of means (Overall mean)	Range of coefficients of variation (Overall mean CV)
Wet weight day 7 (g)	12	0.32–0.63 (0.45)	17–44% (26%)
Wet weight day 21 (g)	12	1.1–2.2 (1.8)	1.7–15% (7.5%)
Snout-vent length day 7 (mm)	10[a]	15–25 (17)	6.6–16% (9.5%)
Snout-vent length day 21 (mm)	10[a]	22–30 (28)	1.0–5.7% (2.7%)
Hindlimb length[b] day 7	10[a]	0.11–0.15 (0.13)	9.3–19% (12%)
Hindlimb length[b] day 21	10[a]	0.43–0.75 (0.59)	1.9–12% (6.2%)
Developmental stage day 7	12	53–54 (54)	0–2.3% (1.0%)
Developmental stage day 21	12	57–59 (58)	0.15–0.91% (0.47%)

[a]In two of the 12 assays, snout-vent length was measured to the termination of the abdomen not to the vent, thus, these length measurements were not included in this data set.
[b]Hindlimb length is normalized by snout-vent length.

a rate of up to 10 to 30% across an entire spawn of tadpoles for unknown reasons. In our experience, the occurrence of bent tails was not related to chemical exposure since the phenomenon occurred in controls and various treatment levels at the same incidence level. The bent tail phenomenon should be kept in mind when observing tadpoles at necropsy and interpreting study results.

Collection of tissues at the time of tadpole necropsy may be beneficial in cases where results of the AMA are ambiguous or potential follow-up investigations are warranted. Tadpole tissues that are particularly responsive to thyroid hormone action, including brains, limbs, and tails, can be collected at the time of test termination, flash frozen in liquid nitrogen, stored at −80°C, and subsequently investigated for genetic biomarkers of thyroid activity (Das et al., 2006; Zhang et al., 2006; Buchholz et al., 2007; Helbing et al., 2007a, 2007b). Preservation of additional tadpole tissues (e.g., liver and gonads) for potential histopathological examination may be useful for assessing toxicity to tadpoles via other pathways or modes of action.

FSTRA Design

The FSTRA was designed to identify substances which may interfere with the hypothalamus–pituitary–gonadal (HPG) axis. In this assay, reproductively mature fish are exposed to a test chemical for 21 days at which time data are collected and the study is terminated. Both OECD and USEPA provide guidelines for the conduct of this study (OECD, 2009b; USEPA, 2009b). Guidelines from these two agencies are generally concordant with respect to the conduct of the study; however, there are several significant differences between them with respect to the FSTRA endpoints. Both assays measure survival, behavior, body length and weight, fecundity, vitellogenin (VTG), and secondary sex characteristics (i.e., body coloration, presence of a fatpad, tubercle morphology). However, only the USEPA guideline advises recording fertilization success, gonadosomatic index (GSI), and gonad histopathology. USEPA also suggests measuring sex steroid concentrations (i.e., estradiol and testosterone) as an optional endpoint. Another difference between these two guidelines is in the proposed species used. The USEPA guideline specifies the use of fathead minnows (*Pimephales promelas*), while the OECD guideline allows the use of either fathead minnows, Japanese medaka (*Oryzias latipes*) or zebrafish (*Danio rerio*). Another difference between the USEPA and OECD FSTRA guidelines is in the stringency of the preexposure criteria. While the USEPA test guideline dictates that spawning occur at least two times in the preceding 7 days to exposure initiation and a minimum average of 15 eggs per female per tank per day for that replicate to be included in the study, the OECD guideline refrains from giving specific guidance on frequency of spawning or egg counts, and instead indicates that fish in the tank should be spawning and cites that 10 eggs per female per day is common. Other differences between the USEPA and OECD FSTRA guidelines include differences in guidance relative to setting the high test concentration. The USEPA guideline indicates that 100 mg/l of the test substance, the limit of water solubility, or the maximum tolerated concentration be used to set the high concentration level in the study, while the OECD guideline indicates that 10 mg/l, the limit of water solubility, or the maximum tolerated concentration be used to set the high concentration level in the study. Differences in these two guidelines make it important to determine how data generated from these studies will be used so the appropriate endpoints are collected.

FSTRA Study Initiation and Conduct

The FSTRA begins with sexually mature fathead minnows between $4^1/_2$ and 6 months old that are placed into a flow through test system that is not test substance dosed.

Table 2
Control Fish Performance in the Fish Short-Term Reproduction Assay

Endpoint	Number of studies	Range of means (Overall mean)	Range of coefficients of variation (CV) (Overall mean CV)
Fecundity (eggs/female/day) Breeding platforms with trays	8	14–37 (28)	8.9–52% (35%)
Fecundity (eggs/female/day) Breeding platforms without trays	3	9.9–19 (13)	36–52% (43%)
Fertility (%)	11	94–99 (98)	0.05–5.0% (1.5%)
Male Tubercle Score	11	16–35 (27)	4.4–28% (14%)
Male GSI[a] (%)	10	1.0–1.4 (1.2)	7.5–28% (16%)
Female GSI[a] (%)	10	11–16 (13)	4.8–32% (18%)
Male VTG[b] (mg/ml)	10	0.0004–0.0053 (0.0020)	28–130% (84%)
Female VTG[b] (mg/ml)	10	12–1.5 × 10^1 (64)	16–110% (47%)

[a]GSI, Gonadosomatic index.
[b]VTG, vitellogenin.

Four female and two male fish are placed into 10 l of water in tanks equipped with three breeding substrates per tank. Fish are monitored daily for mortality and signs of abnormal behavior. Daily egg counts are conducted to determine fecundity and fertility rates for replicate tanks. We have found that daily egg counts can be enhanced by adding a tray under the spawning substrate to capture falling eggs. A stainless steel mesh placed over this tray keeps fish from consuming eggs and can further increase egg counts. In our experience, breeding platforms equipped with mesh-covered trays approximately doubled egg counts (Table 2). It is to be expected that several groups of fish in the preexposure will fail to meet the criteria for use in a definitive FSTRA, either due to fish mortality or lower fecundity values; therefore, extra tanks are set up during the pre-exposure period. For example, a standard study design (water control with three test concentrations) requires 16 tanks of fish; therefore, at least 24 tanks should be set up initially to ensure enough spawning groups will meet this criterion. Even more additional replicate tanks may be required to ensure that enough tanks in the pre-exposure phase of the study meet the 15 eggs/female/day fecundity criterion. For example, at times our laboratory has used up to 30 tanks in the pre-exposure phase of the study to ensure that 16 tanks would meet the criteria for the exposure phase of the assay. This not only increases the number of fish used, but also the amount of time spent monitoring egg production daily. An additional consideration is that many diluter systems cannot hold these additional tanks, so multiple diluters may be required during this pre-exposure phase.

It is often difficult to minimize animal usage during a FSTRA. For instance, fish used in the test system should be within 20% of the arithmetic mean weight of all same sex fish. Because not all fish are within this range, extra fish are required to meet the minimum number required to stock the test pre-exposure tanks. Fish outside this range are ultimately euthanized, as are fish that are used in the pre-exposure, but do not go on to the definitive exposure.

Diseased or parasitized fish can compromise the results of a FSTRA. Parasitic microorganisms (e.g., mycobacteria, microsporidia, helminthes) can be present and potentially go unnoticed in a colony of fathead minnows and other species (Hoffman and Nagel, 1977; Francis-Floyd, 2011). Following low reproductive performance of seemingly healthy fathead minnows in our laboratory, histopathological investigations revealed the presence of both acid fast staining bacteria (mycobacteria) and microsporidial organisms in several fathead minnows received from outside aquaculture suppliers. Although no overt signs of ill health were observed (e.g., mortality, external lesions, abnormal behavior) in these fish, their fecundity was markedly decreased relative to historical controls. For this reason, in addition to daily inspections of fish health in the husbandry population, it is advisable to perform histopathological investigations on a subsample of fish from a given lot before their use in the pre-exposure phase to ensure that there are relatively low to no signs of infection. This is especially important when fish are procured from an outside supplier and the detailed history of health issues for that lot of fish is unknown. It should be noted that it is unlikely that a batch of fish will be entirely lacking in signs of parasitic infection, because various parasitic organisms, such as microsporidia and monogenean flatworms, can exist in a commensal relationship with their fish hosts, without causing any measurable adverse effects at low prevalence or intensity (Paperna, 1991; Sitja-Bobadilla, 2008). Thus, the level of infection and the associated tissue damage, if any, should be assessed by appropriately qualified personnel, preferably by a pathologist with experience in fish diseases. This assessment should help in arriving at a

reasonable prediction of whether the observed levels of parasitic infection occurring in a batch of fish are likely to cause adverse effects on fish health or reproduction. A considerable amount of time and unnecessary expense could be potentially avoided if precautionary steps like histopathological examination are undertaken to ensure the health status of a fish lot before their use in the FSTRA.

FSTRA Endpoints

Mortality. Mortality is an endpoint that signals overt toxicity; and concentrations resulting in overt toxicity are not intended in the FSTRA design. However, test conditions in the FSTRA can be stressful as fish may be expending considerable energies in spawning activities. In addition, disturbance to the tanks is a daily occurrence with removal and inspection of spawning tiles and, at times, daily tank cleaning activities. Low levels of infection have also been commonly reported among cyprinid fish that could contribute to mortality (Hoffman and Nagel, 1977; Francis-Floyd, 2011). These factors, either alone or in combination with low levels of a test compound, can increase mortality. Mortality in the control tanks must not exceed 10%, thus, if more than two fish among all control treatments die during the 21-day exposure period, the test may be considered invalid.

Fecundity. One of the performance criteria in the FSTRA is related to fecundity. In the USEPA FSTRA guideline, this criterion states that before initiating chemical exposures and in the control treatment during the exposure period, spawning should occur every 4 days or fecundity values should average 15 eggs/female/day/replicate (USEPA, 2009b). In addition to having to use increased animal numbers to meet the fecundity criterion at the end of the pre-exposure phase of the assay, another concern associated with the USEPA guideline performance criterion for fecundity is related to the fecundity values in the control tanks at the termination of the exposure phase of the assay. High fecundity values in replicate tanks in the pre-exposure do not necessarily mean that fecundity values will continue to be high in the controls for the subsequent 3 weeks of the exposure phase. In our experience, there was not consistently a good correlation between control fecundity in a replicate tank at the end of the pre-exposure phase versus control fecundity in that same replicate tank at the end of the exposure phase (Fig. 1). This indicates that the pre-exposure fecundity values can be limited in their ability to predict or ensure adequate fecundity values in the controls during the exposure phase.

The advantages of using fecundity as an endpoint is that it is clearly linked to population-level adverse effects and it has been shown to be a fairly sensitive endpoint in the assay (Ankley and Johnson, 2004; Dang et al., 2011). However, the fecundity endpoint in the assay is relatively variable and has low specificity for endocrine disruption (Table 2; USEPA, 2009b; Dang et al., 2011). Practically, counting eggs daily requires a lot of time and can be stressful to fish. It is also important to note that comparisons between laboratories may be impractical because differences in fecundity measures may be attributable to feeding regime, the use of covered trays under breeding tiles, or the methods used to enumerate eggs.

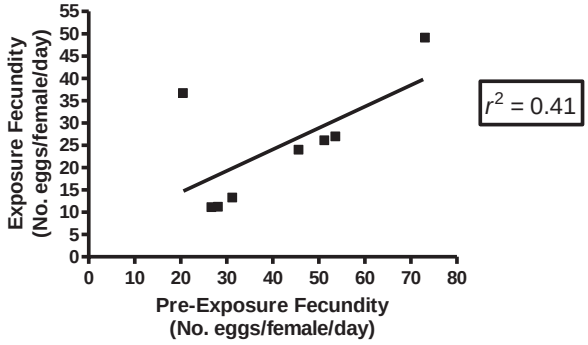

Fig. 1. Linear regression of the average fecundity values for the same control test vessels during the pre-exposure and exposure phases of the fish short-term reproduction assay. The goodness of fit of the linear regression (r^2) was 0.4083, and the slope was not significantly nonzero ($p = 0.0881$).

Fertility. Fertility among fish within a FSTRA does not vary much (Table 2). This endpoint has low specificity with respect to endocrine-mediated effects and, thus, is not a particularly indicative endpoint for determining potential endocrine activity in the HPG axis of fish. Furthermore, fertility tends to not be a particularly sensitive endpoint relative to fecundity (Dang et al., 2011).

Gonadosomatic index (GSI). GSI is the weight of the gonad normalized by the whole body weight of the fish. Overall, this endpoint has relatively low sensitivity for detecting potential endocrine activity (Dang et al., 2011), and changes in GSI can result from changes in either whole body wet weight and/or changes in gonad weight. For female fish, GSI can vary depending on spawning status, with smaller GSI values directly following spawning events (Jensen et al., 2001). Average GSI values for male and female control fish and the variability around this endpoint are summarized in Table 2.

Vitellogenin (VTG). VTG is a yolk-precursor protein normally expressed in female oviparous species and is a highly responsive biomarker for estrogen receptor agonists, especially in males who carry the VTG gene but do not ordinarily express it at appreciable levels (Sumpter and Jobling, 1995; Ankley et al, 2001; Arukwe and Goksoyr, 2003; Pawlowski et al., 2004; Dang et al., 2011). The VTG endpoint is relatively specific for detecting alterations in circulating levels of estrogen in fish; however, hepatotoxicity would also be expected to affect this endpoint since VTG in synthesized in the liver (Arukwe and Goksoyr, 2003). There are various methods by which to measure fish VTG levels (e.g., Enzyme Linked Immunosorbent Assay (ELISA), Liquid Chromatography (LC)/mass spectrometry; Zhang et al., 2004; Brodeur et al., 2006); however, the ELISA appears to be the most common assay and is specifically mentioned in the FSTRA guidelines (Jensen et al., 2001; Jensen and Ankley, 2006; Watanabe et al., 2007; USEPA 2009b; OECD 2009b). In our experience, VTG is a sensitive biomarker of potential endocrine activity; however, commercially available ELISA kits (Biosense Laboratories, Bergen, Norway) produced relatively variable results in both male and female fish (Table 2). VTG variability may be attributable to inherent variability among fish of the same sex, or issues associated

with immunoreactivity and protein stability (Arukwe and Goksoyr, 2003; Brodeur et al., 2006). For example, plasma samples may be frozen following collection and thawed when analyzed, which could result in a breakdown of the VTG protein, the degree to which could vary depending on how the samples were handled (e.g., how many times samples were submitted to a freeze-thaw cycle, whether or not samples were treated appropriately with aprotinin, and citrate buffer containing polyethylene glycol; Brodeur et al., 2006).

Gonad histopathology. Alterations in gonadal histopathology can be used to link molecular changes with specific organismal responses. Histopathological changes are often the result of the integration of a large number of interacting physiologic processes (van der Oost et al., 2003), and can be a useful tool in endocrine disruptor studies to determine a MoA that may be unexpected or counterintuitive (Leino et al., 2005). Histopathological alterations in the gonads of fathead minnows, medaka, and zebrafish were some of the most sensitive endpoints in the 21-day FSTRA (Dang et al., 2011). Background incidence of some histological changes in fathead minnow gonads is normal and should be noted. For instance, over the course of multiple studies in our laboratory, we observed the following in control fish: mild inflammation and mild to severe oocyte atresia in ovaries as well as mild mineralization of the efferent duct and mild granulomatous inflammation in the testes. As with the AMA, it is desirable to have a pathologist trained in examining fish gonads available to read and interpret gonadal histopathology in the FSTRA. Because of the degree of subjectivity involved when assessing the slides, this same individual should evaluate all of the samples from a study, and a peer review process for interpretation of the gonadal histopathology results is also desirable. Signs of infection should also be noted during assessment of the gonads, and the potential influence of noted infection(s) should be considered in the overall interpretation of the study results.

Sex steroids. Sex steroids are an optional endpoint in the FSTRA. Measuring sex steroid hormones can be accomplished by various methods, including Radioimmunoassay (RIA) and liquid chromatography/positive atmospheric pressure photoionization tandem mass spectrometry (Jensen et al., 2001; Zhang et al., 2009). Analytic issues can arise because of the sample sizes associated with this assay, which are small because of the low volumes of plasma available (i.e., generally $\leq$20 μl per individual fish). There are limitations to interpreting the meaning of increased or decreased sex steroid measures in fathead minnows as these endpoints are relatively variable (Watanabe et al., 2007; Jensen et al., 2001) and they represent concentrations at a single point in time (i.e., the time of necropsy).

Secondary sex characteristics. Modifications in secondary sex characteristics can be attributable to alterations in the endocrine system. For example, changes in tubercle number can indicate very specific modes of actions among chemicals, making it a useful endpoint. Androgen agonists can cause females to develop tubercles, whereas this is generally considered a male-only trait (Ankley et al., 2001, 2003). Estrogen receptor agonists, on the other hand, can decrease male tubercle prominence and number (Pawlowski et al., 2004; Filby et al., 2007; Saleirno and Kane, 2009). Although a method for scoring tubercles is well outlined in both FSTRA guidelines, this remains a subjective endpoint. Thus, it is important that a single individual scores each fish at the end of the study. In our experience, there is generally very little variability in tubercle scores (Table 2). Changes in secondary sex characteristics are reportedly less sensitive to endocrine activity in the FSTRA relative to gonad histopathology, fecundity, and VTG endpoints (Dang et al., 2011). Changes in other secondary sex characteristics, such as vertical banding coloration and fatpad development, are also recorded at test termination, but can be even more subjective and variable over time.

Interpretation of Results

Many endpoints may be collected during the course of a FSTRA, including survival, behavior, body length and weight, fecundity, VTG, and secondary sex characteristics, as well as fertilization success, GSI, gonadal histopathology, and sex steroid concentrations. The presence of so many endpoints in the assay increases the likelihood of false positive results. Thus, when making decisions regarding the endocrine disrupting potential of a compound, it is important to evaluate the cohesiveness of all the endpoint responses to a known mode of endocrine action in the HPG axis of fish.

There are limitations in the interpretation of FSTRA. Decisions on the potential for a test substance to exhibit endocrine activity are meant to be interpreted in a weight of evidence evaluation with the other 10 assays in the EDSP battery and other scientifically relevant information. For data with clear-cut responses, reporting No Observable Effect Concentration (NOEC), Lowest Observable Effect Concentration (LOEC) values based on fish survival, growth, and reproduction may be desired. However, this may not be appropriate as the standard design of the FSTRA (i.e., three dose levels and a control) was not designed to determine these values. If an NOEC and LOEC value is desired from an FSTRA, the test design should be modified to include more treatment levels and additional consideration should be given to setting the test concentrations (i.e., including environmentally relevant test concentrations) before conducting the FSTRA.

Caution should also be used when interpreting the individual endpoints in the FSTRA, especially in light of potential impacts of generalized stress or toxicity. For example, fish reproduction (i.e., fecundity) can be decreased in response to stress (i.e., via the hypothalamus–pituitary–interrenal axis and increased cortisol levels) rather than the endocrine activity directly targeting the HPG axis (Aluru and Vijayan, 2009; OECD, 2009b; USEPA, 2009b). Therefore, if a decrease in fecundity is observed in fish with no changes in other, more specific, endocrine-active endpoints (e.g., VTG), then fecundity has likely been reduced because of generalized toxicity or stress. This emphasizes the importance of considering all the data together when interpreting specific endpoints.

Other Considerations

Collection of tissues at the time of fish necropsy may be beneficial in cases where results of the FSTRA are

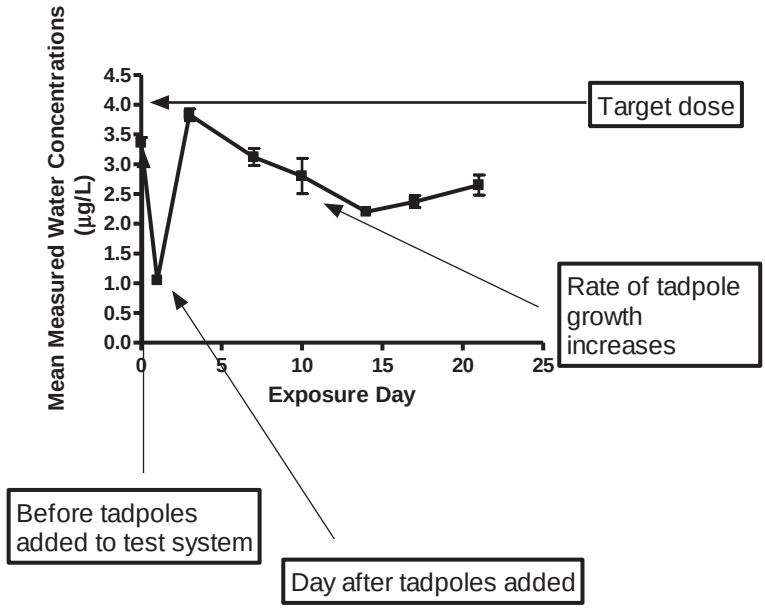

Fig. 2. Change in mean measured concentrations of a lipophilic and adsorptive test chemical during the 21-day amphibian metamorphosis assay. Loss of the test chemical from the water compartment was likely due to increased tadpole biomass and tadpole waste material.

ambiguous or potential follow-up investigations are warranted. Fish tissues that might be of particular interest include brain tissue for measuring inhibition of aromatase (Ankley et al., 2005) and liver tissue for assessing various toxicant responses. In addition, preservation of kidney and/or liver tissues for potential histopathological examination may be useful for assessing toxicity to fish via other pathways or modes of action.

Technical Challenges Associated with both the AMA and the FSTRA

Test concentration selection. The highest test concentration in both the AMA and the FSTRA is determined by either the water solubility limit, the maximum tolerated concentration (MTC), or 100 mg/l (or 10 mg/l in the OECD version of the FSTRA), whichever is lowest. The MTC is the highest concentration resulting in <10% acute mortality and, according to the OECD and USEPA guidelines, can generally be estimated by dividing the 96-hr LC50 (from a rangefinder or pre-existing study on another species) by three. There are limited acute data for *X. laevis*, and acute fish toxicity data may or may not be similar in sensitivity to amphibians (Bridges et al., 2002). Thus, a 96-hr acute toxicity test with *X. laevis* tadpoles may be necessary for estimating an appropriate high test concentration in the AMA. Unlike data for tadpoles in the AMA, fathead minnow acute data are far more common. However, there could be concerns with using acute toxicity data to predict toxicity in a 21-day test, particularly if the test substance has a structural alert associated with an increased potential for a large acute to chronic ratio (Ahlers et al., 2006). In these cases, prolonged range-finding tests (e.g., 14-day exposures) may be necessary to more accurately set the high test concentration, but this would add time onto an already lengthy screening test. Chronic data, if available, may be most useful for setting the dose concentrations in the AMA and FSTRA. When there is uncertainty in dose-setting, applying a 10-fold spacing factor between dose levels may help avoid generalized toxicity at multiple dose levels.

Maintenance of test concentrations. A performance criterion that can be difficult to meet in both the AMA and FSTRA is the maintenance of mean measured test concentrations within a specified range (i.e., coefficient of variation for mean measured concentrations ≤20%). This can be more problematic for some substances than others. For instance, some test substances are extremely lipophilic and/or adsorptive, and partition or adhere into biologic tissue or waste products rather than into the surrounding water. Thus, test concentrations may sharply decrease once biomass (tadpoles or fish) are added to the test system, and then increase as the system reaches a new equilibrium. Because in the AMA the tadpoles are continually growing and excreting increasing amounts of waste, the rate of loss from the water compartment may gradually increase over time. Figure 2 illustrates this scenario in an AMA with a known lipophilic and adsorptive test material. Testing biodegradable substances can also lead to declining test concentrations over time. The buildup of biofilm is common, particularly if using a solvent, and increases with time in the fish and tadpole exposure tanks, despite regular cleanings. While several steps can be taken to mitigate declines in test substance concentration (e.g., increasing diluter turnover rate, more frequently cleaning tanks, or replacing delivery tubes), these measures are often insufficient to maintain coefficients of variation below 20%. When there is foreknowledge or suspicion that test concentrations will be difficult to maintain in either the AMA or the FSTRA, an increasing frequency of analytical sampling is suggested to more closely follow the dynamics of the test material in the test system. A time weighted mean measured concentration should be calculated in cases

when the intervals among sampling time points are not uniform.

Difficult test substances. Many chemistries today present technical challenges when developing a way to deliver them to the test system. Some compounds with low water solubility or high adsorptive capacity are best delivered using a solvent. Using solvents may also be advantageous when testing substances at or near their limit of water solubility in a continuous flow system. For example, conducting an AMA without a solvent at the limit of water solubility with a 25 ml/min flow rate (the minimum recommended) in each of four replicate tanks in each treatment level would require a total flow of 100 ml/min, or 144 l/day, per treatment level. Creating and delivering such a high volume of stock is problematic for many testing facilities, and could be alleviated if concentrated stock solutions were made up in a carrier solvent. Using passive dosing methods (e.g., saturator columns) may alleviate the logistic problem of large stock volumes; however, additional preliminary work and the development of expertise would be necessary before the test.

While acetone, ethanol, methanol, triethylene glycol, and dimethylformamide (DMF) are solvents routinely used in aquatic toxicity testing, DMF is generally more preferred as it produces less biofilm on the test tanks (OECD, 2000). Biofilming can block delivery lines and can be particularly problematic when testing biodegradable compounds. Although a concentration of 0.1 ml solvent/l is acceptable, lowering solvent concentrations (e.g., 0.02 ml/l) may help mitigate biofilming (OECD, 2000; Hutchinson et al., 2006).

Using a solvent can present statistical challenges when there are differences between solvent and water-only controls. When statistically significant differences exist, the OECD AMA guidance document indicates that treatment groups should be compared back to water-only controls (OECD, 2009b). This is in direct contrast to earlier guidance provided by OECD in the guidance document entitled, "Current approaches in the statistical analysis of ecotoxicity data: A guidance to application," which indicates that in cases where there are statistical differences between control groups, the solvent control group is the appropriate control group for comparisons with treated groups (OECD No. 54, 2006). The USEPA guidelines for the AMA and FSTRA do not indicate how statistical evaluations ought to be performed in cases when solvent and water only controls significantly differ, but leaves the decision to best scientific judgment (USEPA, 2009a, 2009b). Taken together, there is no clear guidance on how statistical evaluations ought to be performed when solvent control data are statistically different from clean-water controls. Researchers should examine the data, assess the primary guidance document under which the study was conducted, and use best scientific judgment to decide how best to evaluate data in this scenario.

CONCLUSION

Both the AMA and the FSTRA, along with other Tier 1 battery assays and other scientifically relevant information, are useful assays for making decisions regarding potential endocrine activity of test substances. However, neither the conduct nor the interpretation of these assays is always straightforward; therefore, care in the conduct and interpretation of these studies is needed to ensure that the best decision making is made in regard to the potential endocrine activity of test substances.

REFERENCES

Ahlers J, Riedhammer C, Vogliano M, Ebert R-U, Kuhne R, Schuurmann G. 2006. Acute to chronic ratios in aquatic toxicity—Variation across trophic levels and relationship with chemical structure. Environ Toxicol Chem 25:2937–2945.

Aluru N, Vijayan MM. 2009. Stress transcriptomics in fish: A role for genomic cortisol signaling. Gen Comp Endocrinol 164:142–150.

Ankley GT, Johnson RD. 2004. Small fish models for identifying and assessing the effects of endocrine-disrupting chemicals. Inst Lab Anim Res J 45:469–483.

Ankley GT, Jensen KM, Kahl MD, Korte JJ, Makynen EA. 2001. Description and evaluation of a short-term reproduction test with the fathead minnow (*Pimephales promelas*). Environ Toxicol Chem 20: 1276–1290.

Ankley GT, Jensen KM, Makynen EA, Kahl MD, Korte JJ, Hornung MW, Henry TR, Denny JS, Leino RL, Wilson VS, Cardon MC, Hartig PC, Gray LE. 2003. Effects of the androgenic growth promoter 17-β-trenbolone on fecundity and reproductive endocrinology of the fathead minnow. Environ Toxicol Chem 22:1350–1360.

Ankley GT, Jensen KM, Durhan EJ, Makynen EA, Butterworth BC, Kahl MD, Villeneuve DL, Linnum A, Gray LE, Cardon M, Wilson VS. 2005. Effects of two fungicides with multiple modes of action on reproductive endocrine function in the fathead minnow (*Pimephales promelas)* Toxicol Sci 86:300–308.

Arukwe A, Goksoyr A. 2003. Eggshell and egg yolk proteins in fish: Hepatic proteins for the next generation—Oogenetic, population, and evolutionary implications of endocrine disruption. Comp Hepatol 2:4.

Bridges CM, Dwyer FJ, Hardesty DK, Whites DW. 2002. Comparative contaminant toxicity: Are amphibian larvae more sensitive than fish? Bull Environ Contam Toxicol 69:562–569.

Brodeur JC, Woodburn KB, Zhang F, Bartels MJ, Klecka GM. 2006. Plasma sampling and freezing procedures influence vitellogenin measurements by enzyme-linked immunoassay in the fathead minnow (*Pimephales promelas*). Environ Toxicol Chem 25:337–348.

Buchholz DR, Heimeier RA, Das H, Washington T, Shi YB. 2007. Pairing morphology with gene expression in thyroid hormone-induced intestinal remodeling and identification of a core set of TH-induced genes across tadpole tissues. Dev Biol 303:576–590.

Dang Z, Li K, Yin H, Hakkert B, Vermeire T. 2011. Endpoint sensitivity in fish endocrine disruption assays: Regulatory implications. Toxicol Lett 202(1):36–46.

Das B, Cai L, Carter MG, Piao Y-L, Sharov AA, Ko MSH, Brown DD. 2006. Gene expression changes at metamorphosis induced by thyroid hormone in *Xenopus laevis* tadpoles. Dev Biol 291:342–355.

Denver RJ, Glennmeier KA, Boorse GC. 2002. Endocrinology of complex life cycles: Amphibians. In: Pfaff D, Arnold A, Etgen A, Fahrbach S, Rubin R, editors. Hormones, Brain and Behavior. Vol. II, non-mammalian hormone-behavior systems. Elsevier Science, USA, p 469–513.

Filby AL, Thorpe KL, Maack G, Tyler CR. 2007. Gene expression profiles revealing the mechanisms of anti-androgen- and estrogen-induced feminization in fish. Aquat Toxicol 81:219–231.

Francis-Floyd R. 2011. Mycobacterial infections of fish. Southern Regional Aquaculture Center (SRAC) Publication No 4706, November 2011.

Grim KC, Wolfe M, Braunbeck T, Iguchi T, Ohta Y, Tooi O, Touart L, Wolfe DC, Tietge J. 2009. Thyroid histopathology assessments for the amphibian metamorphosis assay to detect thyroid-active substances. Toxicol Pathol 37:415–424.

Helbing CC, Bailey CM, Ji L, Gunderson MP, Zhang F, Veldhoen N, Skirrow RC, Mu R, Lesperance M, Holcombe GW, Kosian PA, Tietge J, Korte JJ, Degitz SJ. 2007a. Identification of gene expression indicators for thyroid axis disruption in a *Xenopus laevis* metamorphosis screening assay. Part 1. Effects on the brain. Aquat Toxicol 82: 227–241.

Helbing CC, Ji L, Bailey CM, Veldhoen N, Zhang F, Holcombe GW, Kosian PA, Tietge J, Korte JJ, Degitz SJ. 2007b. Identification of gene expression indicators for thyroid axis disruption in a *Xenopus laevis* metamorphosis screening assay Part 2. Effects on the tail and hindlimb. Aquat Toxicol 82:215–226.

Hoffman GL, Nagel ML. 1977. A new host for Pleistophora ovariae (Microsporidia). US Fish & Wildlife publications. Paper 113.

Hutchinson TH, Shillabeer N, Winter MJ, Pickford DB, 2006. Acute and chronic effects of carrier solvents in aquatic organisms: A critical review. Rev Aquat Toxicol 76:69–92.

Jensen KM, Ankley GT. 2006. Evaluation of a commercial kit for measuring vitellogenin in the fathead minnow (*Pimephales promelas*). Ecotoxicol Environ Saf 64:101–105.

Jensen KM, Korte JJ, Kahl MD, Pasha MS, Ankley GT. 2001. Aspects of basic reproductive biology and endocrinology in the fathead minnow (*Pimephales promelas*). Comp Biochem Physiol 128:127–141

Leino RL, Jensen KM, Ankley GT. 2005. Gonadal histology and characteristic histopathology associated with endocrine disruption in the adult fathead minnow *(Pimephales promelas*). Environ Toxicol Pharmacol 19:85–98.

Nieuwkoop PD, Faber J. 1994 Normal table of (*Xenopus laevis*). New York: Garland Publishing.

Organization of Economic Cooperation and Development (OECD). 2000. Guidance document on aquatic toxicity testing of difficult substances and mixtures. Environ Health Saf Publ Series on Testing and Assessment. No. 23, Paris, France.

Organization of Economic Cooperation and Development (OECD). 2006. Current approaches in the statistical analysis of ecotoxicity data: A guidance to application. OECD Series on Testing and Assessment. No. 54, Paris, France.

Organization of Economic Cooperation and Development (OECD). 2007a. Validation of the amphibian metamorphosis assay as a screen for thyroid-active chemicals: Integrated summary report. October 16, 2007.

Organization of Economic Cooperation and Development (OECD). 2007b. Guidance document on amphibian thyroid histology Part 2: Approach to reading studies, diagnostic criteria, severity grading, and atlas. Prepared by Christiana Grim, OSCP/EPA, USA, May 16, 2007.

Organization of Economic Cooperation and Development (OECD). 2009a. Amphibian Metamorphosis Assay. OECD Guideline for the testing of chemicals. No. 231, Paris, France.

Organization of Economic Cooperation and Development (OECD). 2009b. Fish Short Term Reproduction Assay. OECD Guideline for the testing of chemicals. No 229, Paris, France.

Paperna I. 1991. Diseases cause by parasites in the aquaculture of warm water fish. Annual Rev Fish Dis 155–194.

Pawlowski S, van Aerle R, Tyler CR, Braunbeck T. 2004. Effects of 17α-ethinylestradiol in a fathead minnow (*Pimephales promelas)* gonadal recrudescence assay. Ecotoxicol Environ Saf 57:330–345.

Pickford DB. 2010. Screening chemicals for thyroid-disrupting activity: A critical comparison of mammalian and amphibian models. Crit Rev Toxicol 40:845–892.

Salierno JD, Kane AS. 2009. 17α-estradiol alters reproductive behaviors, circulating hormones, and sexual morphology in male fathead minnows *(Pimephales promelas*). Environ Toxicol Chem 28:953–961.

Shi Y-B, 2000. Amphibian metamorphosis from morphology to molecular biology. New York, Wiley-Liss Press.

Sitja-Bobadilla A. 2008. Living off a fish: A trade-off between parasites and the immune system. Fish Shellfish Immunol 25:358–372.

Sumpter JP, Jobling S. 1995. Vitellogenesis as a biomarker of estrogenic contamination in the aquatic environment. Environ Health Perspect 103:173–178

United States Environmental Protection Agency (USEPA). 2009a. Endocrine disruptor screening program test guidelines OPPTS 890.1100: Amphibian metamorphosis assay (Frog). EPA 740-C-09–002.

United States Environmental Protection Agency (USEPA). 2009b. Endocrine disruptor screening program test guidelines OPPTS 890.1350: Fish short-term reproduction assay. EPA 740-C-09–007.

van der Oost R, Beyer J, Vermeulen NPE. 2003. Fish bioaccumulation and biomarkers in environmental risk assessment: A review. Environ Toxicol Pharmacol 13:57–149.

Watanabe KH, Jensen KM, Orlando EF, Ankley GT. 2007. What is normal? A characterization of the values and variability in reproductive endpoints in the fathead minnow. Comp Biochem Physiol 146:348–356.

Zhang F, Bartels MJ, Brodeur JC, Woodburn KB. 2004. Quantitative measurement of fathead minnow vitellogenin by liquid chromatography combined with tandem mass spectrometry using a signature peptide of vitellogenin. Environ Toxicol Chem 23: 1408–1415.

Zhang F, Degitz SJ, Holcombe GW, Kosian PA, Tietge J, Veldhoen N, Helbing CC. 2006. Evaluation of gene expression endpoints in the context of a *Xenopus laevis* metamorphosis-based bioassay to detect thyroid hormone disruptors. Aquat Toxicol 76:24–36.

Zhang F, Bartels MJ, Geter DR, Carr MS, McClymont LE, Marino TA, Klecka GM. 2009. Simultaneous quantitation of testosterone, estradiol, ethinyl estradiol, and 11-ketotestosterone in fathead minnow fish plasma by liquid chromatography/positive atmospheric pressure photoionization tandem mass spectrometry. Rapid Commun Mass Spectrom 23:3637–3646.